Towards Automatic Thermography-Based Leak Detection in District Heating Systems

Zur Erlangung des akademischen Grades eines

Doktors der Ingenieurwissenschaften (Dr.-Ing.)

von der KIT-Fakultät für Wirtschaftswissenschaften
des Karlsruher Instituts für Technologie (KIT)

genehmigte

Dissertation

von

M.Sc. Elena Maiken Vollmer

Tag der mündlichen Prüfung:	12. Februar 2026
Referent:	Prof. Dr. Frank Schultmann
Korreferent:	Prof. Dr. Andreas Oberweis

Karlsruhe, 2025

Abstract

As one of the greatest contributors to anthropogenic climate change, the building sector and its associated operational emissions need to be limited. District heating systems (DHSs) can be part of the solution by supplying climate neutral heat to urban environments via networks of mostly subterranean pipelines. However, continuous operation under extreme conditions makes these systems prone to leaks, which cause persistent losses and potential safety hazards due to the lack and imprecision of existing monitoring methods. Thermography-based leak detection (TLD) has emerged as a non-invasive, pipeline-agnostic alternative, pinpointing underground leaks through their associated elevated surface temperatures. Combined with airborne platforms such as unmanned aircraft systems (UASs), TLD can enable large-scale monitoring of DHSs. However, its wider adoption is limited, in particular, by the effort required to manually evaluate the large amount of resulting thermal infrared images. An automatic analysis is required that is robust, usable, generally applicable, and economically viable to bridge the gap between scientific research and real-world implementation.

The present dissertation addresses these four objectives and central research question through five studies. Reliability is advanced by developing and comparing various algorithms for thermal image processing, thermal anomaly detection, and false alarm removal (Studies A-D). Two novel deep learning (DL) models prove particularly effective for these core tasks (Studies C and D), while data quality is revealed to be a key determinant of performance (Studies A, B, and D). Usability is achieved by fully automating the analysis, including photogrammetric preprocessing, and designing easily interpretable outputs (Study A). Representativity is addressed through the creation and publication of novel UAS-based datasets, thereby establishing new benchmarks for TLD (Studies A and B) as well as DL-based semantic segmentation of spectral imagery (Studies C and D). Finally, a first economic break-even analysis for TLD demonstrates its viability as a cost-effective leak detection strategy, particularly when implemented with the described automation (Study E).

Together, these contributions establish an effective analysis and methodological foundation for automatic TLD, paving the way for its adoption for large-scale DHS monitoring and integration into future smart city applications.

Acknowledgments

This dissertation is the culmination of four years of work that would not have been possible without the support of numerous people. I would like to take this opportunity to say thank you to all who helped me in this endeavour.

First, I am sincerely grateful to my "doctor father" Prof. Frank Schultmann for his supervision and fostering of such an encouraging working environment at the Institute for Industrial Production (IIP). I owe a special thanks to Prof. Rebekka Volk, my former team leader, for our numerous one-on-one meetings and guidance throughout this journey of professional and personal growth. I would also like to extend my deep gratitude to Prof. Andreas Oberweis for being my second reviewer, Prof. Alexander Mädche in his role as examiner, and Prof. Fabian Krüger for completing my examination committee.

This PhD would not have been possible without funding. For three years, I had the pleasure of being part of the incredibly collaborative EU project "AI4EOSC - Artificial Intelligence for the European Open Science Cloud". From the Scientific Computing Centre (SCC) at KIT, I would like thank Borja, Fahimeh, Leo, Lisana, and Valentin for their dedicated support and interactive problem-solving - as well as being great travel buddies. Beyond the KIT, my special thanks goes to Ignacio for welcoming my many questions with such good-natured enthusiasm (I will cherish your User-of-the-Year "award" forever!).

During the scientific endeavours of my dissertation - namely my publications - I collaborated with several talented people I want to give my sincerest thanks to. For the majority of my PhD, I benefited immensely from the consulting of the Helmholtz AI Energy team at SCC. In particular, my sincere thanks go out to James and Mishal for countless discussions, hackathons, and feedback throughout the publication processes. I am profoundly grateful to Markus for going above and beyond in supporting and guiding me and for this incredible new opportunity at the SCC. I also want to say thank you to my former master students, Julian and Leon, for your exceptional enthusiasm, fruitful discussions, and showing me the joys of supervision.

Of course, none of my studies would have been possible without an extensive data foundation, which was provided by Marinus from Air Bavarian GmbH. I am deeply grateful for the enthusiastic support, countless hours spent drone-piloting, and opening invaluable doors for my economic assessment. I also want to take this opportunity to say thank you to all district heating operators and experts who, through their contributions, made my final paper possible. In addition, I give thanks to the Stadtwerke Karlsruhe, for answering my technical queries and providing me with hands-on insights and perspective.

Throughout my PhD, I had the pleasure of working alongside some lovely people at the IIP. Thank you to my whole "Resource Management in the Built Environment" team for the lively meetings, regular afternoon walks, and welcoming environment. I am especially grateful to Simon and Niklas for your friendship from day one and willingness to always lend an ear. My time at the IIP was brightened by many colleagues and a range of welcome distractions, from meandering tea-time discussions to karaoke nights. A special thanks goes to Thorsten for persuading me to leave my office more often and becoming my tea-and-climbing buddy.

Outside of the academic realm, I had the great fortune of being surrounded by a wonderful group of people. Kaleb, thanks for being so relentlessly encouraging on the final stretch, and now, an awesome new colleague and team leader. To my flatmates, bandmates, all HFK friends, and "Radlerlovers" crew: thank you for helping me forget my stress and filling these last four years with adventure and laughter. Bine, thank you for being my best friend, confidant, and ride-or-die sailing and skiing buddy. And of course, Frido - you have been a steadfast rock in the ocean of emotional turmoil that was this PhD journey; without you, I surely would not be here writing this.

Last, but without a doubt not least, I want to thank my family. Mum, dad, Annika (and Max!), Omi, Steffi, and Holger: Your endless support and belief in me means everything.

Karlsruhe, March 2026

Elena Vollmer

Contents

Abbreviations and Symbols

Abbreviations

AI artificial intelligence

BEA break-even analysis

BEP break-even point

CNN convolutional neural network

CV computer vision

DHS district heating system

DL deep learning

DTM digital terrain model

FE feature engineering

FN false negative

FP false positive

FPA focal plane array

GCP ground control point

GE general enhancement

GIS geographic information system

GNSS global navigation satellite system

GPS Global Positioning System

Grad-CAM Gradient-weighted Class Activation Mapping

GSD ground sampling distance

IMU inertial measurement unit

IR infrared

LD leak detection

LoD1 level of detail 1

LT local thresholding

ML machine learning

MLP multilayer perceptron

ODM OpenDroneMap

OSM OpenStreetMap

PIP pre-insulated pipe

PIRP pre-insulated rigid pipe

PS platform-specific

RF random forest

RGB red, green, blue

RGBT red, green, blue, thermal

RJPEG Radiometric Joint Photographic Experts Group

RS remote sensing

SD standard deviation

SM saliency mapping

TASeg Thermal Anomaly Segmenter

THT triangle-histogram-thresholding

TIR thermal infrared

TLD thermography-based leak detection

TN true negative

TP true positive

TUFSeg Thermal Urban Feature Segmenter

UA unmanned aircraft

UAS unmanned aircraft system

UM unsharp masking

VC vignetting correction

XAI explainable artificial intelligence

Symbols

A accuracy

DR detection rate of anomalies

DR_{30} detection rate of anomalies with more than 30 pixels

F_1 F_β-score with $\beta = 1$

F_2 F_β-score with $\beta = 2$ (favours recall R over precision P)

IoU intersection over union

P precision

R recall

ΔT temperature difference [°C]
(expressed in °C instead of K for consistency with existing literature in the field)

$\mathbf{T}_f$ full temperature array [°C]

$\mathbf{T}_m$ masked temperature array [°C]

$\mathbf{Y}$ ground truth array

$\hat{\mathbf{Y}}$ prediction array

List of Figures

List of Tables

Part I

Framework, Foundations, and Implications

1 Introduction

The building sector is widely recognised as one of the major drivers of climate change, generating over a third of global energy demand and associated carbon dioxide emissions [102]. Space and water heating alone account for 45 % of this demand and 80 % of related emissions [102]. Addressing such operational requirements is therefore central to achieving international climate goals, including the Paris Agreement [100] and the UN Sustainable Development Goals [101]. The issue has become even more crucial in light of the ongoing global energy crisis [42].

European governments have begun to tackle this challenge by introducing appropriate legal frameworks. In Germany, for example, recently passed legislation requires municipalities to develop comprehensive, climate-neutral heat supply plans for their cities and communes [11]. Comparable laws long established in Scandinavian countries have led to the adoption of centralised systems, specifically district heating systems (DHSs) [2]. These mainly subterranean pipeline networks distribute heat from central generation facilities, thereby offering a more technologically and economically viable alternative to individual solutions in urban areas [67]. Given the possibility of integrating renewable energy sources and leveraging excess heat and cogeneration, these systems have the potential for full climate neutrality [67]. Strong real-world examples of this already exist – for instance, in Denmark, where DHSs supplied around two-thirds of all households with 89 % climate-neutral heat in 2022 [2].

With more than 600,000 km of pipelines installed worldwide [74], DHSs are a promising technology to addressing the outlined issue. However, the systems face persistent challenges owing to their continuous use for decades under severe operating conditions [61]. In addition to heat dissipation, DHSs must contend with pipeline leaks [56, 74, 117], the most common fault to occur in these networks [60]. When the heated medium – commonly hot water – escapes into the surroundings, it not only greatly reduces the efficiency of the system, but may even cause soil collapse and public safety hazards [22, 60]. A timely localisation and repair is therefore crucial to efficient and safe DHS operation [117].

In practice, however, leaks can often persist given the lack or imprecision of monitoring systems in the underground pipes [22]. A promising solution has emerged to this end: thermography-based leak detection (TLD). Given that leaks mean the release of heated water into the surrounding soil, infrared sensors can be used to identify and locate them as thermal anomalies in so-called heat images [53]. This approach is both non-invasive and pipeline-agnostic and, when combined with aerial acquisition, can be leveraged as a large-scale monitoring approach. However, evaluating the tens of thousands of thermal infrared (TIR) images resulting from such survey expeditions poses a critical obstacle and

bottleneck to TLD adoption [22]. Automating this analysis is therefore crucial to enabling its effective and efficient use as a DHS leak monitoring method. While the past decade has seen progress in this field, a significant gap remains between current research and practical implementation. This dissertation is therefore centred around a key question: How can we bridge the gap between scientific research and real-world implementation to enable TLD as a viable leak monitoring method to DHS operators? To address it, four main research objectives are defined:

Objective 1: *Reliability*
While existing publications propose algorithms for automating the TIR image analysis, their varied data foundations make identifying the most suitable methods impossible. To establish trust in the approach, it is vital to develop the most performant algorithms for a robust and accurate analysis.

Objective 2: *Usability*
Although they design central parts of the analysis for automatic implementation, published works commonly retain several manual steps in their implementations – especially for data preprocessing. To truly enable method adoption, the entire analysis needs to be easy to use and automated in its entirety.

Objective 3: *Representativity*
To lay the basis for broader application, method development and comparisons should be founded on datasets that depict diverse scenery. While previous work often describe their TIR imagery as such, none share them publicly, meaning that no benchmark exists in the field. This is important to allow for future work and developments.

Objective 4: *Economic viability*
Existing publications focus solely on the analysis methodology, leaving the general question of economic viability of the TLD approach unanswered. Given the role of DHSs as heat suppliers, economic feasibility must be quantified to allow operators to make informed decisions on TLD adoption.

To address these, the present thesis is divided into two parts with five scientific studies at its core. This first, Part I, supplies both framework and context – from motivation and theoretical foundations to study overviews, overarching results, and implications. The second, Part II, allows for methodological deep-dives by providing all five associated publications in their full form.

Part I is structured as follows: After this introduction in Section 1, Section 2 provides background and motivation – from the historical growth of DHSs and need for leak detection to thermography and its potential. Section 3 delves into the important theoretical, conceptual, and methodological foundations related to the TLD approach. Section 4 then addresses the overarching objective and motivates the thesis aims through a comprehensive literature review. Section 5 summarises the five studies at the core of this dissertation. With the outlined goals and contributions in mind, Section 6 discusses broader findings, implications, and limitations, while situating these results within prior literature. Section 7 concludes this dissertation with a summary and outlook.

2 Background

To lay the groundwork for methodological discussions, this chapter provides the motivation and background for the focus of this thesis. After delving into DHSs and the need for their maintenance, various existing and novel leak detection (LD) approaches are introduced. A comparison of these is performed, from which thermography emerges as a technology with the potential for comprehensive, large-scale LD.

2.1 District heating systems

District heating, as indicated in Chapter 1, describes the approach of supplying numerous end users with heat through centralised production and an intricate network of pipelines [50, 118]. This manner of distribution has evolved over the years, allowing for more efficient, sustainable, and cost-effective solutions to rival decentralised heating options.

Historically, the development of DHSs can be categorised into several generations, as visualised in Figure 2.1 (s. next page). Emerging in the late 19th century, the first generation utilised steam as a heat carrier. However, this proved inefficient and dangerous due to the high temperatures involved of more than 200 °C. The second generation saw improvements with the adoption of pressurised hot water around the 1930s. Following the two oil crises of the 1970s, the third generation focused on increasing energy efficiency by lowering temperatures to below 100 °C and moving towards alternative means of heat generation, for instance combined heat and power plants, and fuel sources, such as coal, biomass and waste. A transition to the fourth generation is currently planned, characterised by the integration of renewable energy sources and reduction of grid losses to balance energy supply and conservation. Generally, all these developments are marked by lower distribution temperatures, material improvements, and pre-fabrication for simplified construction. [58]

DHSs consist of several key components. At the heart of the system lies the source itself, commonly one of the five so-called "major base heat supplies" [71]: combined heat and power, excess heat from industrial processes, waste-to-energy plants, geothermal heat, and biomass fuels [71]. These are used to heat the transportation medium, which is physically and chemically treated[1] to prevent it from causing internal corrosion. Pumps

[1] This includes removing particulate matter, hardness minerals, salts and oxygen from the water as well as adding chemicals to raise the pH, stabilise hardness, and prevent particle coalescence [33, 118].

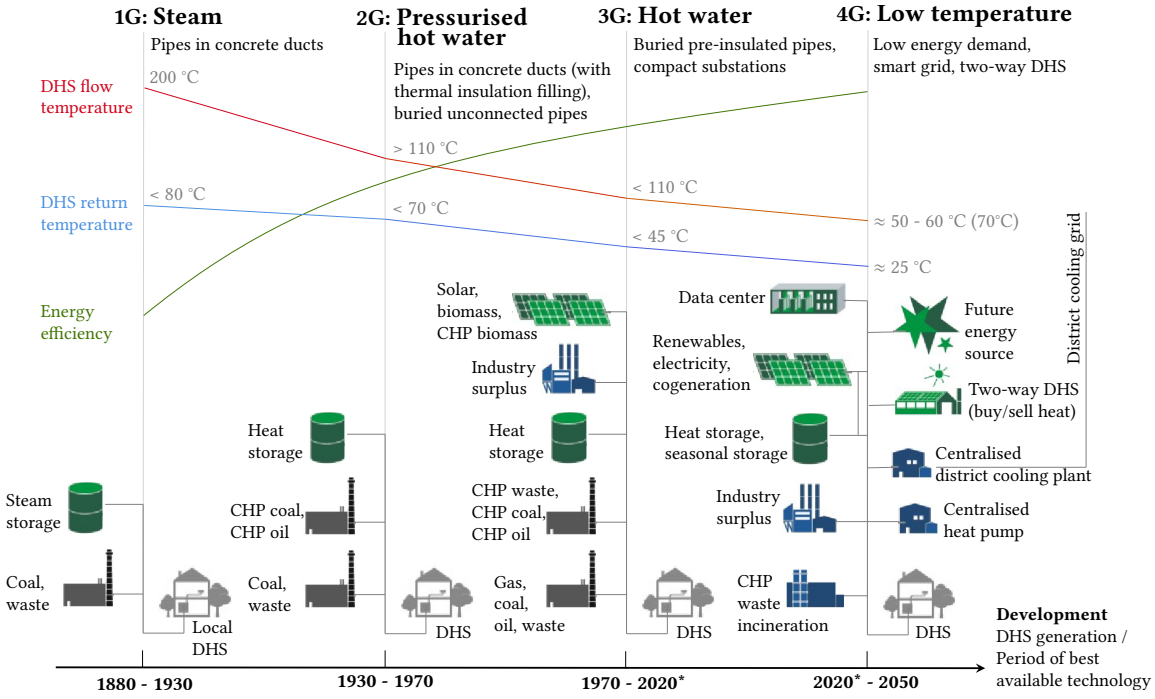


Figure 2.1: DHS development progression (reproduced from Heating [33] and Lund et al. [58]).[2]
Legend: CHP = combined heat and power, DHS = district heating system, * = approximate year

help distribute the water through a network of pipelines to end-energy users in residential and non-residential dwellings [118]. The heating systems of these buildings are either directly connected to the DHS or indirectly linked via interface equipment, such as heat exchangers [118]. After fulfilling its intended purpose, the cooled medium is commonly returned through a parallel set of pipes [118].

Although all components are relevant to system functionality, the pipelines themselves are arguably central to a successful and efficient distribution. Worldwide, they amount to over 600,000 km [74], of which a third are located in the European Union [41]. As depicted in Figure 2.1, their design developed alongside each DHS generation, motivated by a reduction of investment costs, space requirements, installation time, and operating and maintenance costs [4]. For this reason, the original steel service pipes in concrete ducts have since given way to pre-insulated pipes (PIPs) buried directly in the ground [118]. Today, the most widely installed system is the pre-insulated rigid pipe (PIRP), commonly as a composite of steel carrier and plastic sheath [33]. This means the majority of worldwide DHSs belongs to the third generation and is in need of maintenance and renovation [74].

[2] No absolute values for energy efficiency were given by the sources, only the visualised relative trend.

2.2 The need for leak detection

In their role as a provider of essential heating and societal well-being, DHSs can be classified as a vital part of critical infrastructure [61]. As such, their failure can have serious consequences, including discomfort for residents, disruptions in industrial processes, traffic, and broader socioeconomic impacts [61]. Given their placement and function within urban settings, they can compromise safety by interrupting much needed heating services[3] and can even cause deadly environmental hazards, as exhibited by the fatal pipeline burst of Zhengzhou, China, in 2021 [61].

Irrespective of their installation and pipeline type, DHSs are predisposed to degradation due to their operation under extreme conditions, namely high temperatures, pressure, and humidity [61]. Of the many components involved, the pipes themselves are the most susceptible to damage [36, 56, 60]. For instance, data accumulated over four years across the Heilongjiang province in China show that 56 % of faults stem from pipelines, which is more than 2.3 and 5.5 times the amount attributed to valves and compensators, respectively [85]. As highlighted by Table 2.1, faults and defects can be traced back to both system-related or external causes, in consequence of all manner of physical, operational, or environmental factors [74]. While these can be numerous – from different forms of corrosion to defective production, installation, or maintenance [36, 105] –, various studies find exterior corrosion to be the most significant contributor [36, 64, 130]. As an example, Murtazin et al. [64] name prolonged exposure to ground or surface water as the reason for up to 80 % of pipe ruptures in their study of Yekaterinburg, Russia.

Table 2.1: Factors influencing pipeline degradation (reproduced from Rafati and Shaker [74] and Tereshchenko and Nord [95]).[4]

Physical Factors	Environmental Factors	Operational Factors
Age and installation period	Seasonal variation	Previous failures
Corrosion	Soil conditions	Nearby excavation
Diameter		Pressure
Pipe length		Land use
Pipe and insulation material		Temperature levels
Dissimilar metals		Welding

However, the process of reaching an actual rupture is often a gradual one and precipitated by micro-damage [117]. As such, leaks are among the most common operational disruptors of DHSs [105]. Wojdyga and Chorzelski [117] found them to be the leading cause of water loss in Polish systems and problematic for several reasons. Indicated by an increase in the required water input [105], even minor leaks can cause considerable financial losses as a result of the continuous escape of chemically processed and thermally conditioned water [117]. This means that in addition to preceding more serious failures, they contribute

[3] Examples include the 2006 failure in Telšiai, Lithuania, that left thousands without heat for almost a week [130] or the 2020 pipe breakage in Harbin, China, that cut off over 10,000 people from the DHS [56].

[4] Only those listed by Tereshchenko and Nord [95] as the most frequently presumed factors for DHSs are included here. Numerous others exist – f.e. environmental factors such as ground water, climate, or seismic activity – which are not detailed, likely due to their challenging quantification.

to a constant reduction in operational efficiency and increase of costs associated with additional water treatment and heating [117, 129]. In their guide for DHS operators, the Association for District Heating in Switzerland [105] therefore state that, "depending on the size of the DHS and water treatment system, a few cubic metres of make-up water per day can already indicate a serious problem that must be addressed immediately".

All the afore-mentioned circumstances highlight the importance of robust monitoring and maintenance of DHSs, not just to safeguard urban life. Detection and repairs should occur as soon as possible to avoid the consistent losses incurred by leaks of all kinds [117]. The wear of the network is uneven by nature [64] and pipeline types may be affected differently[5] [117]. At minimum an annual, data-informed assessment of DHSs is essential for prioritising repairs and rehabilitation [64]. From this, it naturally follows that robust tools for LD are required to enable safe, reliable, and sustainable network operations.

2.3 Existing methods

Several methods exist for LD in DHS pipelines, varying in terms of technology, methodology, readiness level, and practical applicability. Generally, these systems are either dynamic, using mobile devices or materials, or static, based on fixed sensors [51]. Their operational requirements, possible application areas, and subsequent diagnostic capabilities can vary greatly. This section gives an overview of some of the most relevant approaches in existence, compiled from a wide range of sources [3, 30, 33, 50, 51, 118, 126, 129]. In particular, the German Working Committee on District Heating (AGFW) [3]'s report is used as a point of reference, given their focus on methods with true potential for practical implementation.

Operational changes Initially, DHS operators typically detect the presence of leaks through changes recorded by system sensors, such as pressure drops or increasing water losses [3, 30, 118]. While these are useful indicators, they enable at best the identification of a generally affected network section, not precise localisation [3]. As such, they serve as a preliminary assessment before deploying more accurate LD techniques [3].

Visual inspections Visual inspections by experts help assess pipeline conditions and detect early-stage defects or degradation, especially in older or complex areas like above-ground, duct-based, or basement pipes [3]. Depending on the pipeline type, different forms of direct or indirect inspection methods are possible. Accessible pipes, such as above-ground, duct-based, or basement, can be inspected by experts in person, while environmental cues – such as snowmelts [30] or steam emerging from nearby manholes – can be used as indirect indicators for leaks in PIPs [G. Roehl, personal communication,

[5] Statistical analyses of Polish DHSs by Wojdyga and Chorzelski [117] show that pipe segments with smaller diameters (DN40 to DN80) are disproportionately more affected, both in frequency and severity.

May 8, 2025]. Inaccessible duct-based pipes can be examined remotely using inspection crawlers (e.g., Crawler-Eye). Though such approaches are generally useful for pinpointing faults, the required personnel effort means they need to be focused on specific areas of interest.

Tracer gases Tracer techniques involve the addition of detectable substances (e.g., helium or Uranin) to the heat transport medium to pinpoint leak locations through their presence. Helium diffuses to the surface and can be detected by specialised sensors [3, 30, 126]. Uranin fluoresces and can reveal leaks through visual confirmation. This is especially useful when a risk of contamination with potable water exists, such as at construction sites [3, 30]. These methods do not disrupt operation, but require careful preparation, approval, and post-process flushing [30].

Integrated systems Some types of LD methods have been developed for particular pipeline systems. Modern PIRPs, for instance, are equipped with integrated leak detection systems in the form of wire circuitry within the insulation layer. Moisture ingress can thereby be identified and leak locations pinpointed [3, 33, 118]. These systems are based on one of two main measurement principles. The first is resistance comparison, whereby moisture is detected through the resistance changes it causes [33, 49]. While effective, this method can only locate one fault at a time [49]. Impulse runtime, on the other hand, sends high-frequency pulses and detects reflections, enabling precise localisation of multiple faults, albeit requiring higher moisture levels [33, 49]. All systems support central, decentral, or manual monitoring, ranging from centralised data collection to local inspections with portable equipment [33]. While potentially very effective for LD, their usability is highly dependent on a precise setup during pipeline installation [G. Roehl, personal communication, May 8, 2025].

Ground-penetrating radar Ground-penetrating radar uses electromagnetic radiation to create an image of the subsurface, including buried pipe systems. A depth of up to 5 m can be examined through reflected radar frequencies, which work independently of different pipe materials and other parameters. This allows for reliable and accurate leak detection. However, in addition to the need for experienced operators, the high cost of equipment and the required direct access to areas above the pipelines – including accompanying road closures – are considered disadvantages. [30, 31]

Thermography Another pipeline-independent, yet non-invasive approach for LD is TLD [3]. The diffusion of leaked heat through the ground causes increased temperatures on the surface that can be captured by TIR sensors. As anomalous hot-spots in thermal imagery, they are locatable given the image's position information. TLD can be performed from the ground – via handheld or vehicle-mounted cameras – or through remote sensing (RS) – using airplanes, helicopters, or unmanned aircraft systems (UASs)[3, 30, 129]. Aerial UAS-based surveys are particularly effective due to low cost, high resolution, and

flexibility [38, 128]. Challenges include high data volume and environmental sensitivity, necessitating specific acquisition conditions and expert interpretation [51, 74, 107, 111].

Acoustic and vibration sensors Acoustic methods detect high-frequency signals generated by pressurised fluid escaping through a leak. Among the most widely used techniques is correlation analysis, which uses synchronised sensors to estimate leak locations via signal delays [3, 51]. While this enables a sub-metre leak localisation during active pipeline operation, a clearly defined search area must be known in advance to avoid significant effort and at least two trained experts to take measurements [3]. Effectiveness may also be limited by pipe material and background noise [3, 51].

In-pipe methods In contrast to the visual inspection methods that are applied externally, inspection robots can be used to assess the walls of buried DHS pipes from the inside. Equipped with multiple ultrasonic sensors, they emit sound waves into the pipe wall and record reflections caused by material boundaries or defects such as corrosion. While real-time detection of severe wall thinning is possible during the scan, the main evaluation is conducted afterward by specialist personnel based on recorded data. The method offers high precision, but cannot be performed during operation as the pipe sections need to be dewatered and cleaned beforehand. [3]

2.4 Method comparison

To understand how these heterogeneous LD techniques compare to one another, Table 2.2 (s. next page) contrasts them across five operationally relevant aspects: pipeline applicability, coverage, diagnostic capability, application mode, and implementation effort.

- Pipeline applicability indicates the universality of the methods – meaning their usability across different pipeline types. Although most are agnostic, some techniques are only applicable to specific pipeline types, such as integrated systems to PIRPs [33, 49] or in-pipe robots to rigid pipes [3].
- Coverage distinguishes methods in terms of the scalability of their implementation. Where some require pre-defined search areas, others can be deployed network-wide.
- Diagnostic capability refers to the different types of tasks the methods are able to perform – from general DHS condition assessment[6] to specific leak-related methods. In terms of LD, a distinction is made between the general ability to detect the presence of leaks and general areas of interest and the more nuanced true leak localisation.
- The types of application modes highlight logistical constraints: fixed systems can run continuously once installed, while mobile variants require active planning and execution.

[6] While this is included for the sake of completeness, the true focus hereby lies on LD.

- Lastly, effort provides a general indication of the labour requirements for method use, as each has its own associated workload. For instance, mobile systems by nature demand a higher operational effort given the common requirement for on-site personnel and / or added equipment.

Table 2.2: Comparison of LD methods according to pipeline applicability, area coverage, capability, application mode, and effort (based on descriptions from German Working Committee on District Heating (AGFW) [3], Gurklienė et al. [30], Heating [33], Konstantin and Konstantin [50], Latif et al. [51], Woods [118], and El-Zahab and Zayed [126]).
Legend: DC = direct contact (at/in pipe), IC = indirect contact (close proximity), RS = remote sensing.

Method	Pipeline applicability	Coverage		Diagnostic capability			Application mode				Effort
		Network-wide	Chosen section	Condition overview	Leak presence	Leak pin-pointing	Fixed	Mobile DC	Mobile IC	Mobile RS	
Operational change	All	✓		✓	✓		✓				Low
Visual inspections	All exc. PIPs[7]		✓	✓	✓	✓		✓	✓		Mid
Tracer gas / dye	All	✓[8]	✓		✓	✓		✓			High
Integrated systems	PIRPs only	✓	✓	✓	✓	✓	✓		(✓)[9]		Low-mid
Radar	Buried only		✓	✓	✓	✓			✓		High
Thermography	All	✓	✓	✓	✓	✓			✓	✓	High
Vibro-acoustic	All		✓		✓	✓			✓		High
In-pipe robots	All exc. flexible		✓	✓	✓	✓		✓			High

Against this backdrop, TLD emerges as an interesting choice for three key reasons.

1. Breadth of applicability: The thermographic approach is pipeline-agnostic and – when implemented via RS – entirely non-contact. This enables the comprehensive surveillance of mixed networks, including PIPs, duct-based, pour-in-place, and others, and without physical access to the pipe [3].
2. Scalable coverage: In comparison to most other techniques, TLD does not require a pre-defined search area but can survey large areas and entire networks through airborne RS acquisition methods [30].
3. Diverse capabilities: Subsequent TIR evaluation not only allows the detection of problem areas, but also a precise leak localisation. As an additional benefit, the acquired data can be used to gain a general overview of the network and assess pipeline conditions [3, 30].

These advantages of thermography as a uniquely flexible, pipeline-agnostic, and useful LD tool may be the reason for its increasing adoption across DHSs [3, 51, 74], with initial field studies like Tuikka [99] showing how it has been able to greatly reduce reliance on public reporting. However, as described in Section 2.3, the approach is subject to its own set of disadvantages. One such critical downside is the associated effort – in particular the

[7] While some simple leak indicators exist for PIPs, such as snowmelts or steam from nearby manholes, these are not universally applicable as they depend on pipe locations. PIPs are therefore excluded.
[8] This only works for the entire DHS if there are no hydraulically separated or blocked areas [3].
[9] Manual systems require additional indirect contact access to the pipes in order to perform readings with portable measurement devices [33].

viewing and analysis of the acquired images, of which there may result tens of thousands given a RS-based coverage of large DHSs [22]. Given the mobile nature of the method, the effort must be repeatedly invested with each periodic acquisition flight [30].

For this reason, the past decade has seen a development of approaches for the computational analysis of TIR imagery to automate the leak localisation process and identify points of interest for network operators [9, 22, 38, 39, 86, 88, 122, 128]. However, before we can turn to these publications in detail, it is crucial to lay a methodological foundation to understand the considerations and constraints that govern thermography and TLD.

3 Theory, Concepts, and Methods

After exposing the need for DHS monitoring and potential of TLD, this chapter examines the chosen approach in more detail. Key concepts and methods are discussed to provide an understanding of the existing requirements and later serve as a basis for addressing the contributions this thesis is able to make. To this end, the chapter is organised to follow the procedural steps of the TLD approach, which encompasses two main parts:

- *Acquisition*: This aspect revolves around hardware, meaning the involved specialised sensors and acquisition platforms. Given the nature of thermography as a technology, specific conditions must be met during acquisition to ensure high-quality data is collected [54]. Only in this manner can its usability for the task be guaranteed.
- *Analysis*: Covering an entire DHS incurs a large number of images that require analysis [22]. Another requirement must therefore be the automatic evaluation of the collected data. Effective and performant methods are vital to deduce a useful list of leak candidates and make the approach viable to network operators [22]. This part therefore focusses on software, specifically an introduction into computer vision (CV) and relevant algorithms.

3.1 Data acquisition

Given that sufficient image quality is crucial for successful LD, it is important to understand the physical principles of thermal imaging and which conditions may impair its functionality. This section therefore first covers the basics of all involved technologies before discussing the requirements that must be met to achieve high-quality results. Two devices are vital for image acquisition: A camera that can capture TIRs and an acquisition platform from which data recording is performed.

3.1.1 Thermal images and thermographic sensors

Digital images are composed of pixels, each representing a portion of the captured scene's information. The number of pixels in an image defines its resolution, commonly expressed as image width by height. Each pixel may store a single value representing intensity (greyscale) or three values corresponding to colour channels. Greyscale images thus have one band, while colour images – also known as red, green, blues (RGBs) – have three. [48]

Imaging sensors, including standard visual cameras, operate by detecting specific regions of the electromagnetic spectrum. As shown in Figure 3.1, this spectrum comprises all forms of electromagnetic radiation, sorted by wavelengths and frequencies. Thermographic cameras make use of the infrared (IR) band, which lies just beyond visible light and is commonly subdivided into near, short, mid, long, and far wavelengths [24].

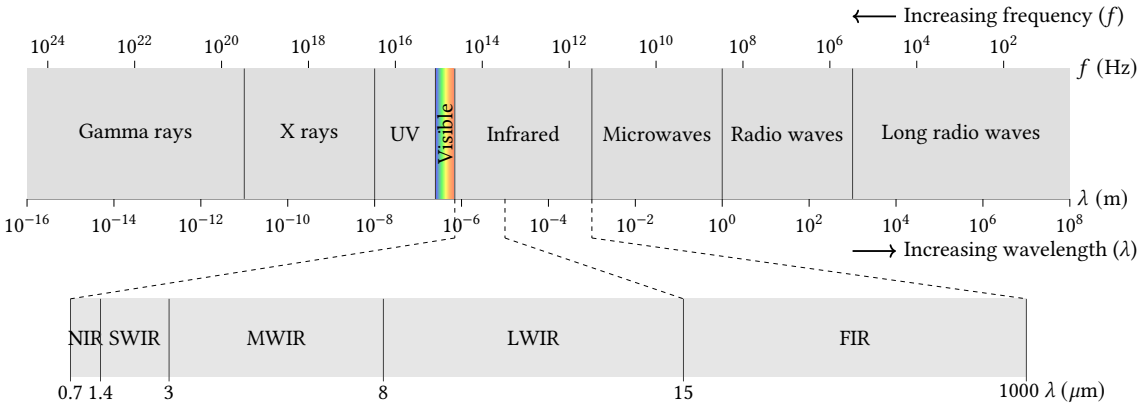


Figure 3.1: Electromagnetic spectrum including subdivisions of the IR spectrum (reproduced from [24]).
Legend: NIR = near, SWIR = short-wavelength, MWIR = mid-wavelength, LWIR = long-wavelength, FIR = far IR

Objects can both absorb and emit electromagnetic radiation. Those with temperatures between 190 K to 1000 K emit within the mid- and long-wavelength infrared ranges and their specific temperature values directly define wavelength and intensity of said emission. While theoretically ideal blackbodies can absorb all radiation and emit solely as a function of temperature, real-world objects suffer from emission losses due to their material properties and surface texture. Emissivity thereby describes how effectively a surface emits radiation compared to a black body of the same temperature. Most non-metallic materials perform well in the long-wave range, with high and relatively constant emissivity regardless of surface characteristics. [24, 34]

A thermal imager is a passive sensor that detects the IR radiation emitted by objects within its field of view. Its lens is made from materials that can transmit IR, such as Germanium. In place of single-point or row measurement approaches, modern systems employ an infrared focal plane array (FPA) – a two-dimensional arrangement of detectors to capture an entire scene at once. The two main technologies used for absorbing incoming IR radiation are photon and thermal detectors. While photon detectors directly convert radiation into electronic energy distributions, thermal detectors transform it into thermal energy, altering the physical properties of heat-sensitive materials and generating an electrical response. Photon detectors operate in the mid-wavelength IR range and offer high sensitivity to marginal radiation differences but require extensive cooling[1] to minimise thermal noise. Given the resulting high acquisition and maintenance costs, slightly less sensitive thermal detectors are often more practical. [24, 124]

[1] Typically below 77 K [24]

Uncooled thermal imaging systems typically capture long-wavelength IR radiation through an FPA consisting of ferroelectric detectors or microbolometers [15]. The two technologies differ mainly in terms of their materials. Ferroelectric detectors employ a dielectric component with a distinct phase transition, causing large changes in electrical polarisation with small temperature variations. Microbolometers, in contrast, are resistive elements whose electrical resistance changes with temperature. Today, microbolometers are more prevalent due to advantages such as higher sensitivity, greater sensor resolution, and finer temperature discrimination. [24, 125]

Despite this, TIR image resolution is still considerably smaller than that of standard visible-light RGBs. Uncooled microbolometers require larger, diffraction-limited pixels to capture long infrared wavelengths and more complex sensor fabrication [77, 124], resulting in resolutions of only 160×120 to 1280×1024 pixels [24, 32] versus 4000×3000 of common optical imagery [87]. Additionally, the field of view, focal distance, and pixel size of such thermal imaging systems are generally much more limited [87].[2]

In TIR imagery, each pixel corresponds to a temperature value derived from the IR energy directed at it. The extraction of raw temperature data varies by file format and manufacturer, a common example of which is the company Teledyne FLIR's Radiometric Joint Photographic Experts Group (RJPEG) [123]. It is important to note that accurate absolute temperature measurements are inherently challenging due to the numerous factors influencing IR radiation transmission from a scene to the sensor [24, 113]. While the measured signal is primarily determined by the scene's emissivity, it is also affected by ambient temperature, atmospheric moisture and composition, and reflected background radiation [24, 113]. Where possible, relative values should be used for a meaningful assessment. In terms of TLD, leaks can manifest with as little as 5 °C temperature difference ΔT to their surroundings [88].

3.1.2 Acquisition platform

IR imaging can be deployed in various different ways that can be distinguished into two key categories: ground-based devices or airborne aircraft.

Ground-based cameras, such as hand-held or vehicle-mounted, offer the highest spatial resolution or so-called ground sampling distance (GSD)[3]. However, these require full ground access, are extremely time-consuming for entire network coverage, and present automation challenges given their constantly changing ground perspectives. In contrast, so-called RS-based methods – such as airplanes or UASs – have several advantages. Fixed camera rigs produce more stable imagery, acquisition is contact free and non-invasive, and large-scale implementation becomes feasible. For the purpose of LD for DHSs, this thesis therefore places exclusive focus on airborne forms of TLD.

[2] For example, the sensor specifications listed by Sledz et al. [87] for each of the mentioned parameters are 45°, 13 mm, and 17 µm for their thermal versus 94°, 3.6 mm, and 1.56 µm for their visible.

[3] The GSD of an image defines the ground surface area represented by each pixel [34].

While airplanes were once preferred, their use has declined due to high operational costs, complex logistics, and lower GSD resulting from the high altitudes [88]. Technological advances have made UASs a natural successor – offering greater flexibility, significantly reduced costs, and typically achieving up to five times higher GSD than manned aircraft [88]. Broadly, the term unmanned aircraft (UA) covers all vehicles capable of flight without an onboard pilot and functionally encompasses sub-classes such as remotely piloted vehicles and drones.[4] Modern UA technology typically requires minimal piloting skills and includes auto-stabilisation, with operator input limited to altitude, speed, and heading angle. Integrated navigation systems enable pre-programmed flight paths and maintenance of constant altitude, attitude, and speed. To this end, positional information is customarily acquired by leveraging global navigation satellite systems (GNSSs) such as Global Positioning System (GPS) and attitude through inertial measurement units (IMUs). Measurement accuracy depends on system and sensors, with GNSS typically achieving submeter accuracy and additional corrective measures such as real-time kinematics reducing the positional error to only a few centimetres [63]. [21, 116]

In practice, the UA forms part of a complete UAS, comprising the aircraft, a ground control station, a communication data link, and the payload. The aircraft – whether fixed-wing, rotary-wing, or ducted fan – includes the airframe, propulsion, power system, controls, and its portion of the data link. The ground control station manages operations and handles the other end of the data link, enabling two-way communication for both control commands and telemetry. More advanced systems can also transmit video or radar data. For this dissertation's focus on TLD, a basic UAS configuration with the payload of a thermal imager and, optionally, additional RGB camera is all that is required. [21, 116]

Images resulting from the afore-described forms of RS-based capturing contain more than just pixel data; they also store metadata that describe the circumstances of acquisition. Technical metadata includes information such as image dimensions, compression type, and camera settings. Content metadata may record descriptive scene details and temporal data such as date and time. Positional metadata can include GNSS coordinates and altitude. Thermal cameras designed for RS applications are even able log platform attitude parameters – yaw, pitch, and roll angles – for both the aircraft and its gimbal [62, 69].

3.1.3 Acquisition requirements

While thermography can be an effective tool for DHS inspection, the nature of the technology (s. Section 3.1.1) means its usefulness is strongly dependent upon acquisition conditions. As hot-spots of interest may manifest with only a few degrees more than their surroundings, high-quality TIRs are essential for comprehensive TLD. One of the most important aspects of acquisition planning therefore involves the close monitoring of weather conditions [54].

[4] While these terms are used interchangeably in colloquial contexts, the former refers specifically to actively piloted vehicles, while the latter denotes a lack of sophistication and explicit use for monotonous, repetitive tasks only. [21]

- Image acquisition should take place at night, ideally a few hours before dawn [22, 34].
- Air temperatures should not exceed 10 °C [34].
- Snow, rain, fog, and high wind speeds[5] should be avoided [3, 22, 34].
- The ground surface should be dry and free of foliage, ice, or snow [3, 22, 34].
- Acquisition is generally limited from autumn to spring [22, 34]. In the northern hemisphere, the time window is defined as October to April.

Night-time acquisition ensures that most irrelevant heat sources are eliminated. Starting at dawn, atmospheric scattering of sunlight begins to warm the earth's surface. This effect intensifies under direct sunlight throughout the day, as buildings and other urban objects reradiate the absorbed heat. After dark, surface temperatures drop and cool down to more homogeneous background levels, while human activity decreases. Acquiring TIRs at night therefore significantly reduces the number of heat signatures caused by humans and their means of transport while enabling a greater temperature difference between DHS pipes and urban environment. Limiting the acquisition to colder months amplifies the latter effect, as DHSs run at higher temperatures then. [22, 34]

Where possible, the thermal camera should have an unobstructed view of the ground above the DHS. Pre-acquisition checks are essential to ensure that foliage or snow cover, which insulate the surface and obscure subsurface anomalies, are absent. Prolonged rainfall and high winds cool the ground, falsifying the recorded thermal signatures. [3, 22, 34]

Generally, precipitation such as snow, rain, or fog attenuates heat radiation, as water particles absorb IR energy before it can reach the camera sensor. This results in grainy TIRs with reduced detail and contrast. [3, 34]

Additional constraints relate to the physical characteristics of the DHS. Installation depth should not exceed 1 m for TLD and remain consistent throughout the network where possible. While the method is compatible with various pipe casings – including fibre cement, plastic, protective, steel, and flexible pipes – regardless of placement above or below ground, uncommon materials should be verified in advance for suitability. Network characteristics such as dimensions, age, installation depth, and medium temperature are best confirmed with the operator or energy supplier. [34]

Finally, legal considerations may also apply. Local regulations concerning RS operation, especially UAS, over private or company property may necessitate special permits or restrict acquisition in certain areas. While privacy concerns are minimal given that individuals cannot be identified in TIRs, data protection rules should still be observed. [34]

[5] Ljungberg and Rosengren [54] recommend a maximum of 3 m/s at ground level, while Heipke and Tödter [34] specify 2 m/s, though they note that they often exceeded said limit.

3.2 Data analysis

The automatic analysis of TIRs encompasses all algorithms required to programmatically extract a list of potential leak candidates. Since the data are image-based, this work falls within the domain of CV – an interdisciplinary field concerned with replicating or augmenting human vision through computers. CV enables machines to capture, analyse, understand, and interpret visual information from images [93]. This section introduces key concepts and techniques relevant to TLD and the previously described analysis task. Core methods range from image processing to pattern recognition – by means of both traditional and artificial intelligence (AI)-based approaches.

3.2.1 Data preprocessing

Before coming to the main task of leak candidate extraction, the raw TIR imagery should be prepared for what is to come – a step also known as data preprocessing [27]. These kinds of algorithms helps create a consistent, accurate, and potentially richer dataset in which patterns are more easily distinguishable [10]. In the case of RS-based TIRs, preprocessing revolves around three key aspects: 1. extracting temperature values from raw sensor outputs, 2. assessing thermographic quality and handling distortions, and 3. applying photogrammetric processing to enhance the data with spatial information.

3.2.1.1 Temperature extraction

The methods for extracting temperature values from acquired TIRs depend on the given format. Images of common proprietary formats, for instance RJPEGs, contain both visualised and raw thermal data packaged in one [123]. This means that standard processing software read the files as though they were conventional images, extracting the visualised data as intensity values. Each pixel thereby assumes only integer values between 0 (corresponding to black) and 255 (corresponding to white), giving the images an intensity resolution of 256 [48]. An example of this is shown on the left side of Figure 3.2.

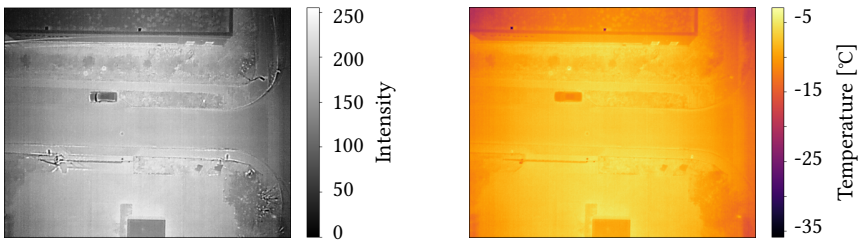


Figure 3.2: Example of the intensity and absolute temperature array extractable from a RJPEG TIR image (own visualisation).

To access the radiometric data stored in the images' metadata, specialised libraries such as the FlirImageExtractor can be leveraged. The thermal sensor values are extracted and

converted to decimal temperatures in degrees Celsius. The full temperature array $\mathbf{T}_f$ corresponding to the afore-mentioned example is visualised on the right side of Figure 3.2.

Figure 3.2 also exemplifies an issue mentioned in Section 3.1.1, namely that of incorrect temperature measurements common to long-wavelength IR thermal imagers [113]. In this case, values up to −35 °C were falsely registered, presumably on account of the metal surfaces on the roof [24]. As a countermeasure, it can be useful to computationally derive temperature boundaries with which to limit the TIR data to a realistic range.

3.2.1.2 Image quality

In contrast to standard RGB images, the use of TIRs comes with its own set of challenges. These are elicited by two aspects: The imaging technology itself (s. Section 3.1.1) and the RS-based acquisition (s. Section 3.1.2). The following discussion is centred around uncooled microbolometer FPAs as the most popular type of thermal imagers for these applications given their low weight and energy requirements [114, 125].

General issues First and foremost, it is important to keep in mind the significant differences between familiar visible-light RGB and the here utilised TIR imagery. A primary limitation is the comparatively low image resolution resulting from the microbolometer FPAs, which yields fewer pixels and thus coarser scene detail. This can reduce the ability to resolve fine structures (also known as the mixed-pixel problem in RS) [45] and complicate co-registration with higher-resolution RGB data [87].

General image quality can be quantified and common issues mitigated through the application of filters. Linear filtering modifies images by applying a kernel, a small matrix that is convolved with the input to generate the adapted output [27]. Common examples include the Gaussian filter for smoothing (shown in Figure 3.3) or the Sobel filter for edge detection [27, 93, 96]. Methods based on these, such as unsharp masking (UM), have been found to help counteract issues of blurring and unwanted noise [12].

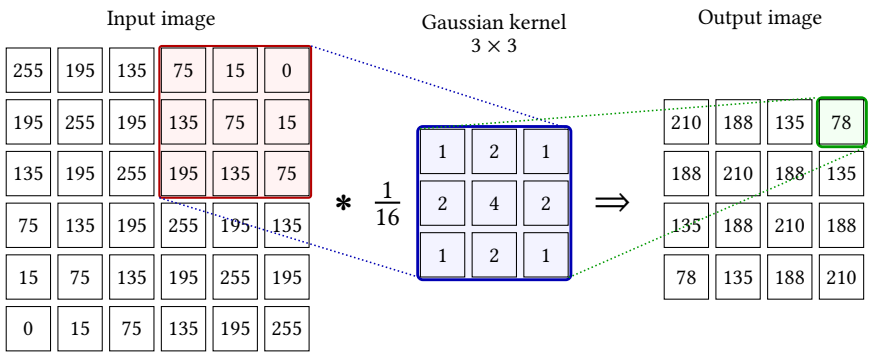


Figure 3.3: Kernel convolution exemplified with a Gaussian filter (based on descriptions from Gonzalez and Woods [27]). Colour highlights illustrate the operation in action: An input image patch (red) is weighted with the kernel (blue) via element-wise multiplication and summed to yield an output value (green). In this example, it is calculated as $\frac{1}{16}\,(1 \cdot 75 + 2 \cdot 15 + 1 \cdot 0 + 2 \cdot 135 + 4 \cdot 75 + 2 \cdot 15 + 1 \cdot 195 + 2 \cdot 135 + 1 \cdot 75) = 78$.

Thermal drift and vignetting A key issue is the susceptibility of uncooled microbolometer FPAs to various types of artefacts and distortions [125]. Regardless of calibration, such imagers are subject to thermal drift – a phenomenon where changing ambient and operational conditions cause deviations in measured temperatures over time [98, 114]. During UAS flights, influencing factors include ambient temperature, wind speed, light, internal electronics, and UA components [114, 125]. This has several important consequences for TIR quality and analysis.

The dynamic nature of these impacting factors means that TIRs can vary systematically even within a single flight [98, 125], as exemplified in Figure 3.4. Treier et al. [97], for instance, highlight significant temporal shifts within all their measurements, commonly more than 10 °C over the course of one flight line. At the same time, Yuan and Hua [125] emphasise the importance of warmup periods as they find sensor instabilities within the first hour of acclimatisation to be one of the major causes for measurement errors, especially given lower ambient temperatures. These temporal and inter-dataset aspects can complicate the central analysis task of thermal hot-spot detection. From a methodological perspective, selected algorithms should avoid assuming uniform dataset statistics as this can bias the results towards falsely registered areas of higher temperature.

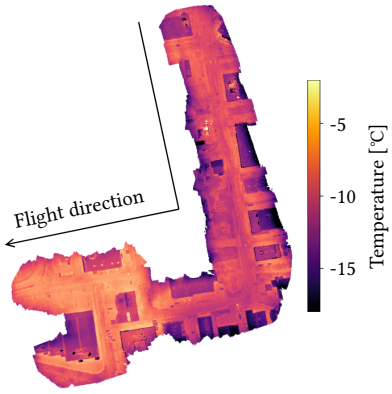


Figure 3.4: Inter-dataset thermal drift exemplified in a mosaic (reproduced from own image originally published in Vollmer et al. [111]). Despite consistent conditions, registered values increase over the course of the flight.

Within individual images, another challenge arises from the setup of thermal imagers as FPAs consisting of numerous microbolometers. Variations in sensitivity can produce inhomogeneity within the same output, also known as fixed pattern noise [114, 125]. To mitigate this, modern thermal imagers employ so-called non-uniformity correction to adjust offsets between the many detectors, such as shutter-based[6] [66, 114, 125]. However, its effectiveness is often limited due to fluctuations in environmental and operational factors – in other words thermal drift [114]. The resulting distortions commonly manifest as so-called vignetting [114], exemplified in Figure 3.5. Image edges appear systematically colder than the centre as a result of the opposing cool propeller slipstream and heated gimbal motor. This effect can persist despite the periodic and automatic implementation of corrective measures. In their study, Yuan and Hua [125] find that both the severity and pattern of vignetting to vary as a result of thermal drift. As lower ambient temperatures and higher wind speeds strengthen how it manifests, this issue is particularly relevant to the task at

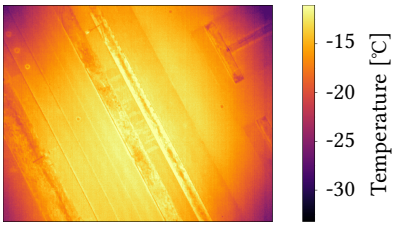


Figure 3.5: Inter-image vignetting effect exemplified in a single TIR image (own visualisation originally published in Vollmer et al. [109]).

[6] An opaque or semi-transparent, temperature-homogeneous shutter is periodically closed in front of the FPA for recalibration [66, 125].

hand. Post-acquisition corrections may be necessary to compensate for image falsification, which commonly leverage calibration imagery. [125]

3.2.1.3 Photogrammetric processing

While RS enables the efficient coverage of large areas, the acquisition form poses an additional challenge for the central aim of precise geographic leak localisation in TLD. Raw TIR imagery consists of pixel arrays with only coarse positional metadata (s. Section 3.1.2). Within the broader field of geomatics – the discipline of managing spatial data [16] – photogrammetric processing offers a means to bridge this divide. It centres around deriving and analysing spatial information from digital images [16, 57], whereby this dissertation focuses specifically on the methods of georeferencing and orthomosaicking. Before coming to these, some general concepts have to be introduced.

In spatial analysis, data can be represented in two main formats: vector and raster. Vector data are geometric objects (points, lines, or polygons) with explicit coordinate locations, while raster data are composed of grids of cells (two- or three-dimensional arrays) to represent space [16, 44]. Images such as the discussed TIRs therefore fall into the latter category, where each pixel can be tied to a location in geographic space using a spatial reference system. These consist of a coordinate system, a datum, and – if desired – a projection to define the exact location to which a given coordinate refers. Datums mathematically represent the Earth's curvature in 3D, while projections convert these into two-dimensional maps. In a geographic coordinate system, positions are expressed as angular coordinates (latitude, longitude) using a datum such as the global World Geodetic System 1984 or the more local European Terrestrial Reference System 1989. In a projected coordinate system, e.g. Universal Transverse Mercator, the curved Earth is mapped onto a plane, and positions are given in linear units such as metres. Spatial data can be displayed and analysed using a geographic information system (GIS). [25]

Georeferencing Georeferencing refers to the process of assigning image pixels to geographic coordinates within a chosen spatial reference system – in other words "mapping" an image [16, 25]. This can be achieved by calculating a transformation between the image space (rows and columns) and ground space (x, y, z), for instance via affine, polynomial, or spline geotransform [25, 44]. Methods are commonly based on known reference locations, so-called ground control points (GCPs) [25, 119]. As a result, each pixel in the given TIRs can be associated with a location on the Earth's surface, enabling true leak localisation and the potential exploitation of other geographic information in vector or raster format. For RS-based images, a significant challenge stems from internal and external geometric distortions caused by sensor perspective, motion, and terrain characteristics [116, 119]. The

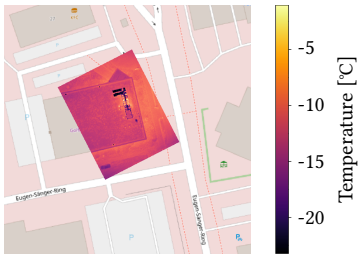


Figure 3.6: Example of a georeferenced TIR visualised in a GIS (own image originally published in Vollmer et al. [111]).

compensation of these, sometimes termed georectification, can require additional complex processing [119]. An example of a georeferenced image can be seen in Figure 3.6.

Orthomosaicking Orthomosaicking extends the aforedescribed mapping principle from individual image-based algorithms to datasets. Given the generally high amount of overlap in RS – and especially UAS – acquisition [7], images can be seamlessly stitched together to form mosaics [93]. When georectification is included, the outputs are called orthomosaics [46]. Various photogrammetry software exist to help automate the entire processing workflow, such as Pix4D and OpenDroneMap (ODM) [7]. Underlying algorithms are commonly based on automatic extraction and matching of key image features [93]. Some specialised software are even capable of handling non-visible data, such as the lower quality TIRs central to this thesis [7]. As shown in Figure 3.7, these output rasters of larger dataset areas, consisting of distortion-free averaged temperature values and geographic coordinates.

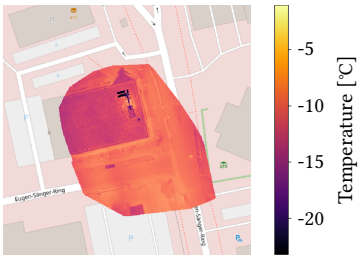


Figure 3.7: Example of a Pix4D orthomosaic visualised in a GIS (own image originally published in Vollmer et al. [111]).

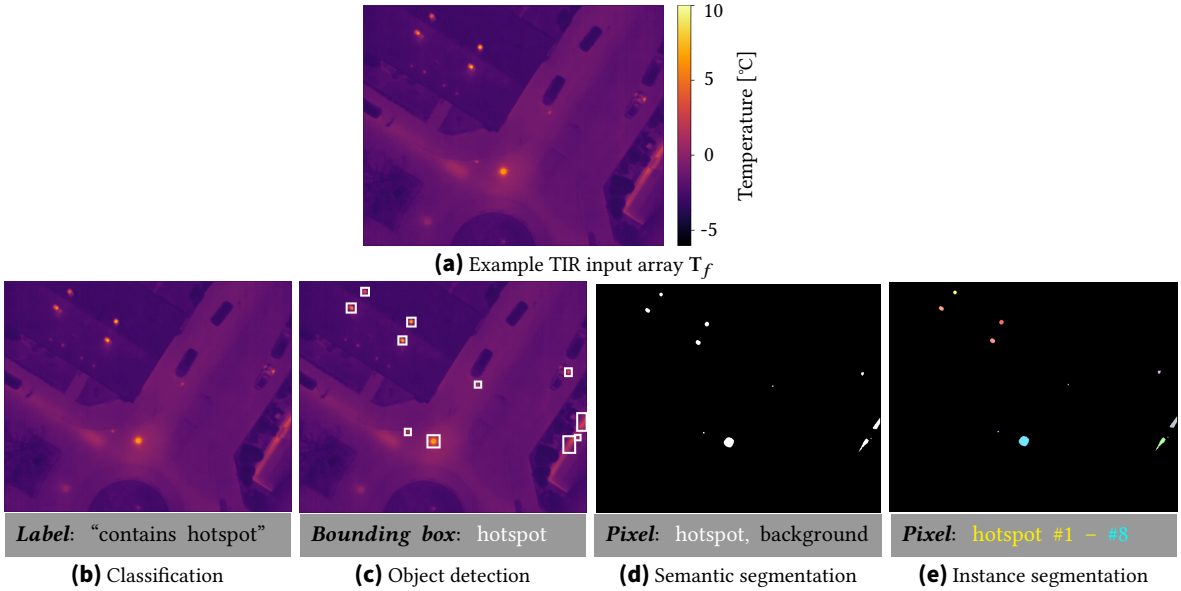
3.2.2 Pattern recognition

Where the previous subsection discussed the preprocessing of data (meaning both input and output are images), the following chapter is dedicated to methods which can produce attributes and characteristics as outputs from these processed inputs [27]. Pattern recognition is a central task in CV, which refers to the automated detection and classification of meaningful structures in data [10]. In the context of TIR imagery and this application, the task comprises finding anomalous hot-spot regions and, where possible, identifying whether they truly pertain to DHS pipeline network leaks.

As a foundation, some taxonomical clarifications should be made. A wide range of image-related CV tasks can be categorised into one of a few canonical problem types. They allow for different levels of localisation detail and differ in generated output [93, 96]. The types are visualised in Figure 3.8 (s. next page) using the example $\mathbf{T}_f$ in 3.8a as an input.

- *Image classification:* Assigns one class label to the entire image. For the mentioned example, the output might be a single label such as “contains hotspot” or “no hotspot” (s. Figure 3.8b).
- *Object detection:* Predicts bounding boxes and class labels for each object instance. The output here might be box locations of identified leak candidates within the image and associated “hotspot” might be provided (s. Figure 3.8c).
- *Semantic segmentation:* Assigns a class label to each individual pixel. An output image of the same dimensions would be returned with a definition per pixel as belonging to “hotspot” class or not (s. Figure 3.8d).

- *Instance segmentation:* Assigns both a class label and an instance ID to each pixel, thereby distinguishing between separate samples of the same class. An output image of the same dimensions would be returned again, but with all pixels that are identified as “hotspot” additionally being associated with a specific instance – such as “hotspot 1”, “hotspot 2”, and so on (s. Figure 3.8e).



(a) Example TIR input array $\mathbf{T}_f$

(b) Classification **(c)** Object detection **(d)** Semantic segmentation **(e)** Instance segmentation

Figure 3.8: Key problem types for image analysis tasks illustrated using an example TIR image.

Each of these type definitions can be formulated as either a binary or a multi-class problem, depending on whether the aim lies in distinguishing only two categories (e.g. “anomaly” versus “no anomaly”) or assigning one of several possible classes (e.g. “leak”, “street lamp”, “building”, “background”, etc.). For finding leak candidates in TIRs, the main focus lies on segmentation – one of the most challenging tasks to automate in image processing [27].

CV approaches can broadly be divided into classical methods – based on explicit rules, statistical models, and hand-crafted features – and AI-based methods, which learn discriminative features directly from data. Classical methods include techniques such as thresholding, filtering, or region growing, commonly valued for their simplicity, interpretability, and computational speed [27]. AI-based methods – known more as black-box approaches [37] – range from traditional machine learning (ML) to deep learning (DL) architectures, the latter reaching from conventional convolutional neural networks (CNNs) to the newer transformer models in the field of image analysis [28, 96]. Both families of methods are applicable to TIR imagery and may be used for the tasks at hand. Those algorithms relevant to this thesis’ publications are discussed in the following subsections.

A final aspect of pattern recognition lies in method evaluation. This can be done by comparing the generated output, also known as the prediction $\hat{\mathbf{Y}}$, with the desired result, the ground truth $\mathbf{Y}$. Metrics are used to quantify differences based on the amount of true positives (TPs), false positives (FPs), true negatives (TNs), and false negatives (FNs) [93]. Accuracy A, defined as $(TP + TN)/(TP + FP + FN + TN)$, is the proportion of correctly

classified pixels to all pixels. Precision P, equal to $TP/(TP + FP)$, refers to the fraction of predicted positives that are correct, while recall R, defined $TP/(TP + FN)$, measures the percentage of true pixels in $\mathbf{Y}$ that were successfully found. To balance the two, the F_1 score quantifies their harmonic mean, while a variation of this – the F_2 score – emphasises R to penalise missed detections. Lastly, intersection over union IoU is equal to $TP/(TP+FP+FN)$ and measures the overlap between $\hat{\mathbf{Y}}$ and $\mathbf{Y}$ relative to their union. [89]

3.2.2.1 Classical methods

As indicated, various classical methods exist with which to extract regions of interest from TIR imagery. This begins with generating a binary segmentation output of relevant pixels and concludes with clustering the pixels to anomaly regions.

Among the simplest methods for generating a binary segmentation output $g(x, y)$ from an input image $f(x, y)$ is thresholding [27, 83]. As the name suggests, this operation works by applying a limit T to every pixel (x, y) to define

$$g(x,y) = \begin{cases} 1, & \text{if } f(x,y) > T \\ 0, & \text{if } f(x,y) \leq T \end{cases} \tag{3.1}$$

The choice of T naturally depends on the segmentation objective and given circumstances. However, to enable an automatic analysis, it must be derived in a computational manner. Various approaches exist to this end [83]. Generally speaking, T can be defined as a constant over an entire image (global thresholding) or vary across it (variable thresholding) [27]. Latter methods are commonly called local or adaptive thresholding when T depends on local pixel neighbourhoods or the value of a pixel itself, respectively [27].

A way in which global thresholds can be defined is by leveraging histograms [27, 83]. Histograms show the distribution and frequency of continuous data by grouping them into ranges. In TIRs, these can therefore display the breadth of temperature values, with particularly high values (those of interest for TLD) at the uppermost end. An example of variable thresholding, on the other hand, might be the exploitation of pixel neighbourhood statistics [83]. A simple strategy is the comparison of a pixel's value against the mean and standard deviation of its immediate surroundings [38], which would – in the case of TIRs – allow for locally warm regions to be identified. More advanced techniques include maximum entropy thresholding [47], where the split between fore- and background is determined by maximising the combined information content, expressed as entropy.

To further hone the segmentation, different steps can precede thresholding. A simple approach can be the use of kernel operators (s. Section 3.2.1.2). Generating filtered versions of the inputs can help highlight interesting image structures before segmentation [27]. The Sobel filter, for instance, can be used to emphasise high temperature gradients in TIRs [38]. More complex methods may include the transformation of the input image into so-called saliency maps [43]. Based on visual human perception, these maps model how strongly each pixel stands out in the image context via colour, intensity, and orientation

features [43, 122]. In terms of TLD, they can indicate conspicuously warm areas of interest before thresholding is used for segmentation [122].

Additionally, various segmentation outputs can be combined to create even more refined output. For instance, applying a logical AND operator ensures that only pixels identified by multiple criteria (e.g., both intensity- and gradient-based thresholds) are selected, thereby focusing on holistic regions of interest. [27, 38]

Once a binary mask has been generated to define every image pixel as either relevant or irrelevant, the pixels of interest can be grouped into regions through clustering. Classical techniques include connected component algorithms or region growing [27, 93]. This step transforms the pixel-wise segmentations into larger coherent regions to represent potential leak candidates in the TIRs.

3.2.2.2 AI-based approaches

Rapid developments in recent years have enabled comparable image processing to be achievable via AI-based methods [75, 96]. To provide the foundational knowledge relevant for this thesis, the following section focuses on models for image analysis and, in particular, segmentation. Given the breadth of this field, a brief introduction of the involved taxonomy is called for.

Where AI encompasses the broad field of creating intelligent machines, ML is a subset of algorithms capable of learning from data. Instead of explicitly programming a solution to a given problem, ML models are trained on data to discover patterns that enable them to predict on unseen information. Generally, these methods can be categorised according to the form of training data available and desired learning experience. In supervised learning, models are meant to learn defined patterns and to this end are provided with labelled examples. This means the inputs are paired with the desired outputs – for example TIRs with corresponding "ground truth" segmentation masks. In unsupervised learning, only the data themselves are supplied, and the aim lies in identifying general structures or trends. As the absence of actual labels prevents the training process from being steered in the desired direction, this thesis focuses on methods of supervised learning. [28, 75]

To this end, a variety of ML algorithms have found application in the field of CV. Methods such as logistic regression, support vector machines, and random forests (RFs) typically rely on engineered image features rather than learning directly from pixel data. These stand in contrast to neural networks, a particularly effective class of models inspired by biological neurons in the human brain. While numerous variations of this style of architecture exist, the basic structure can be illustrated with a multilayer perceptron (MLP). As visualised in Figure 3.9 (s. next page), an MLP consists of fully interconnected artificial neurons arranged into an input layer, one or more hidden layers, and an output layer. The output y of each neuron is calculated by summing its weighted inputs, adding a bias term b, and applying a non-linear activation function ϕ. As the overall output is obtained through the step-by-step calculation of each layer, these structures are referred to as feed-forward networks. During training, the model output is compared to the provided

label and resulting error quantified by means of a loss function. To refine the model, this loss is minimised by adapting the weights and bias terms. Gradients needed for these updates can be computed efficiently through the backpropagation algorithm. [28, 70, 75]

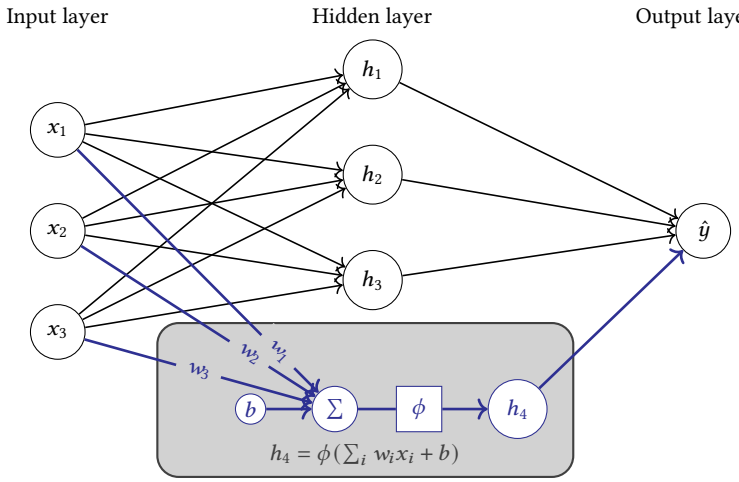


Figure 3.9: Example of an MLP with a call-out to illustrate the computation inside a single neuron (based on Pérez-Enciso et al. [70]).

The success of AI in CV is largely thanks to the development of these kinds of approaches, which in the field of ML are defined as belonging to the subspace of DL [96]. DL refers to models of significantly greater complexity and size, developed on larger datasets for solving more difficult tasks. The previously described concept of neural networks form the foundation of the advanced architectures implemented in this thesis.

In image analysis, a special kind of network has come to dominate the field: CNNs [75]. The idea originates from problems where the model inputs can be characterised as spatially-dependent, such as one-dimensional text or two-dimensional images [70]. The concept of convolution is used to automatically extract relevant features from the input data – in images, for instance, by applying convolutional kernels to identify edges, corners, and textures [28, 75]. This is combined with pooling, which merges multiple neighbouring data points into one representation to make the features more robust towards small input translations [28, 75]. As such, CNNs consist of at least one convolution-pooling block that precedes conventional fully connected layers. To enable automatic feature extraction, models learn both kernel parameters and weights during training.

In terms of task type, the specific problem of image segmentation was first addressed through fully convolutional networks [55], which enable a pixel-wise classification by replacing all remaining fully connected layers of a CNN with convolutional ones. Building on this idea, most modern semantic segmentation models follow one of two design principles: 1. Encoder–decoder structures, which progressively compress the image into a latent representation before reconstructing it, and 2. Pyramid pooling, which aggregates multi-scale contextual information [13]. These concepts are the foundation of many performant architectures, including the U-Net (an encoder–decoder with skip connections) [78],

PSPNet (a pyramid pooling variant) [127], and DeepLabV3+ (a hybrid design combining spatial pooling with a decoder) [14].

Another group of architectures, coined transformers, have recently been found to outperform CNN-based variants in different semantic segmentation tasks [52, 96]. Where CNNs focus on locality – the independent processing of small image regions – transformers are able to globalise and connect information across spatial distances [96]. They do so by leveraging the concept of attention in visual perception, similarly to the saliency maps mentioned in Section 3.2.2.1 [96]. As such, they are characterised by highly parallelised processing and self-attention mechanisms to dynamically assess how relevant image patches are to one another [18, 104]. Models typically consist of stacked encoders and decoders that transform inputs into latent vectors, which are in turn used to generate outputs. Although highly effective in numerous applications, these architectures require large datasets, significant computational resources, and specialised hardware. This has motivated the development of lightweight variants such as SegFormer [120], which achieves comparable performance to models such as the resource-intensive Swin-Transformer [52] despite its use of far fewer parameters.

Given the complexity and black-box nature of DL models, an expanding field of study revolves around understanding and explaining model behaviour, aptly coined explainable artificial intelligence (XAI) [37]. Approaches vary in terms of timing (ad-hoc – built into the model, or post-hoc – applied after training), range (model-specific or -agnostic), and requirements (using only prediction outputs or internal model layers) [26, 37]. While the use of such techniques has grown more common for classification, applications to semantic segmentation are still emerging [26]. Among the most popular methods for the former that are also applicable to the latter are gradient-based techniques [26]. One example is Gradient-weighted Class Activation Mapping (Grad-CAM), which produces a visual heatmap of regions the model finds important by tracing the gradients of a target class back to the activations of the final convolutional layer or latent representation [82].

4 Research Framework

After laying a motivational and methodological foundation, this chapter turns to the specific research objectives required to enable airborne thermography as an effective and viable tool for large-scale LD in DHSs. Given the narrow scope of this research field, these are best motivated by a review of related literature. Section 4.1 therefore begins by summarising existing publications implementing automatic TIR image analysis for LD. Section 4.2 condenses their approaches into a general procedural framework. Using this foundation, Section 4.3 compares the current state-of-the-art and identifies research gaps related to achieving the overarching objective. Finally, Section 4.4 outlines the focus areas of this thesis, defining the remaining requirements to bridge the gap between scientific research and real-world implementation to enable TLD for DHSs.

4.1 Related work

The fact that leaks in underground DHS pipes can be identified via thermography was initially discovered and described by Axelsson [5] and Ljungberg and Rosengren [53] in the late 1980's. They demonstrate that TIR imagery from mobile or airborne platforms can reveal leaks by the resulting surface hot-spots. Despite this initial revelation, the automation of image analysis for large-scale TLD has only become an area of interest in the past decade. To this end, four research groups have contributed to the field with various publications [9, 22, 38, 39, 86, 88, 122, 128].

Friman et al. [22] were the first to describe a method for the automatic detection of pipeline leaks in TIRs in 2014. Acquired via manned aircraft in Sweden at 800 m altitude (GSD of 24 cm), the authors first georeference the images using "semi-automatic commercial off-the-shelf software" [8] based on the recorded GPS and IMU information. The search space is then reduced by masking the TIRs with buffered GIS data of the pipeline routes, thus removing all image areas that are not above or in close proximity to the DHS. Subsequently, Friman et al. [22] define thermal anomaly as pixels that exceed a set intensity threshold of the probability distribution above the pipe network. Of the significant number of potential leaks identified in this manner, a considerable amount pertains to false alarms. To address these false positives – many of which are found to be caused by buildings – the authors apply building segmentation via watershed transform and later refine results using a classifier. However, Friman et al. [22]'s methodology also suffers from drawbacks, which are mainly the comparatively high number of false alarms and the misclassification of buildings. Berg et al. [9] therefore expand upon the approach by extracting additional

image features for binary ML-based classification of found detections as true or false, improving building segmentation using OpenStreetMap (OSM) data, and introducing temporal analysis to compare images across acquisition dates. [9, 22]

Research by Hossain et al. [38, 39] builds upon the work of Berg et al. [9] by extending the AI-based classification approach. Their studies' 243,082 TIR images are acquired in Denmark by means of UAS[1]. The authors begin by identifying thermal anomalies via region extraction algorithm based on a variety of filter kernels. Their selection is based on two assumptions regarding how leaks manifest in TIRs: that they present as local thermal maxima and have low edge gradients due to heat dispersion. To this end, the approach incorporates both a pixel intensity filter using arithmetic mean and standard deviation of radius-based surroundings and a Sobel filter for the exclusion of high gradients. The two resulting segmentation masks are combined by logical AND operator and smoothed with a majority voting kernel. Using these candidates and manually created labels, the authors proceed to compare several conventional ML classifiers – including Berg et al. [9]'s RF – with a more complex CNN. The DL model is found to outperform standard ML for this binary classification. [38, 39]

Xu et al. [122] propose an alternative leak detection approach based on saliency analyses. This method models the human visual system and attention-directing mechanisms under the principle that image regions of particular visual significance can thus be identified as salient. In their study, thermal imagery acquired by manned aircraft in Sweden (GSD of 20 cm) is georeferenced and masked with GIS pipeline data, manually buffered via ArcGIS. Saliency maps are generated using a simplified Itti et al. [43] model incorporating colour, orientation, and intensity features and anomalous pixels are identified via subsequent threshold segmentation. In their later work, Zhong et al. [128] combine local saliency maps – comparing regions to their immediate neighbourhoods – with global maps that calculate feature rarity across the image. The fused IR saliency maps are segmented using a maximum entropy method to detect anomalies. Zhong et al. [128]'s TIR datasets comprise one from a manned flight in Sweden and two UAS-based from Sweden and China at flight heights of 120 m to 150 m. As an alternative to using GIS pipeline data, they propose algorithms for segmenting roads from simultaneously acquired RGB imagery, based on the assumption that pipelines are frequently located beneath them.

Adopting a different strategy, Sledz et al. [88] leverage blob detection for the anomaly identification. Approximately 3,000 UAS-based TIRs acquired in Germany at 40 m flight height (GSD 5.2 cm) are semi-automatically processed into orthomosaics using Agisoft Metashape and masked with buffered GIS pipeline data. Elliptical hot-spots are detected via a Laplacian of Gaussian blob detector, with anomalies classified as leaks if the temperature difference ΔT between the warmest region and its surroundings exceeds a defined threshold. The resulting numerous false alarms are manually categorised to determine their primary causes. The authors suggest that a binary classifier combined with digital surface model-based filtering could automatically exclude above-ground sources.

[1] Hossain et al. [38, 39] do not provide further details on their acquisition conditions.

Given this recommendation, Sledz and Heipke [86] implement a RF classifier using manually created features based, in part, on a digital surface model. They also replace blob detection with a saliency-based anomaly detection method, though it differs from Xu et al. [122]'s and Zhong et al. [128]'s approaches. The change is motivated by limitations in said prior work, particularly a tendency to mark all visually unusual regions as salient (not just warm ones) and a normalisation step which can suppress less salient anomalies in the presence of more significant ones. To address these issues, Sledz and Heipke [86] modify the intensity feature map calculation of Itti et al. [43] by considering only positive intensity deviations via a max operation and constraining values to a percentile-defined interval before normalisation, ensuring weaker anomalies remain detectable. For segmentation, they fuse the resulting saliency map with information from simultaneously acquired RGB images using the Dempster–Shafer evidence theory [17, 84].

4.2 General analysis procedure

While each research group has contributed unique approaches and algorithms, their workflows follow similar procedures to perform the image analysis task. In general, all proposed steps can be assigned to one of the following three aspects:

- *Data processing*: Given the type of imagery and subsequent or prior tasks, processing may be required to prepare the data for (optimal) usage.
- *Thermal anomaly detection*: At the core of the LD task lies the identification of pixels of interest within the TIRs, which, when clustered, form the coveted thermal anomalies and initial list of potential leak candidates.
- *False alarm removal*: Given the urban placement of DHSs, many detected anomalies will be irrelevant. Another key task lies in refining the results to meaningful leak candidates. In this context, a conservative filtering strategy is essential: the cost of prematurely removing a genuine leak far outweighs that of including additional false positives (R over P). It is therefore preferable to produce a slightly longer list of candidates for manual inspection rather than risk excluding true leaks.

As the descriptions in Section 4.1 indicate, the sequence of these steps can vary. For example, false alarm removal is sometimes applied before thermal anomaly detection, typically by reducing the search space to areas around the pipelines [9, 22, 88, 122, 128]. This involves overlaying pipe blueprints on the TIRs and excluding non-intersecting areas, which in turn requires additional data preprocessing such as georeferencing or mosaicking [22]. In general, the chronological order of steps for automatic leak candidate identification can be defined as in Figure 4.1.

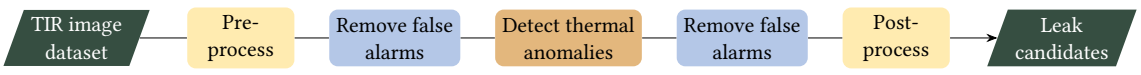


Figure 4.1: General image analysis procedure (based on descriptions from Berg et al. [9], Friman et al. [22], Hossain et al. [38, 39], Sledz and Heipke [86], Sledz et al. [88], Xu et al. [122], and Zhong et al. [128]).

4.3 Research gap

Building upon the provided procedural framework (s. Section 4.2) and literature review (s. Section 4.1), the existing publications can be categorised according to their data foundation and developed methodologies. To this end, Table 4.1 provides a summary and comparison of the named publications.

Table 4.1: Overview of existing literature's data foundation and implemented methods (based on descriptions from Berg et al. [9], Friman et al. [22], Hossain et al. [38, 39], Sledz and Heipke [86], Sledz et al. [88], Xu et al. [122], and Zhong et al. [128]).
Legend: ✓ = automatic implementation, ✓* = semi-automatic or manual implementation, (✓) = implemented but not used in final analysis.

Research groups		Data foundation			Analysis algorithms										
		Region	Aerial vehicle	GSD	Georeference	Orthomosaic	Mask with DHS	Intensity histogram	Saliency map	Local filters	Blob detector	Building removal	AI binary classifier	Categorisation	Temporal analysis
Sweden	Friman et al. (2014) [22]	Sweden & Norway	Airplane	24 cm	✓*	(✓*)	✓	✓				✓			
	Berg et al. (2016) [9]				✓*	(✓*)	✓	✓				✓	✓		✓
China	Xu et al. (2016) [122]	Sweden & China	Airplane & UAS	19 – 24 cm	✓*		✓		✓						
	Zhong et al. (2019) [128]				✓*	✓*	✓		✓						
Denmark	Hossain et al. (2019) [38]	Denmark	UAS	-						✓			✓		
	Hossain et al. (2020) [39]									✓			✓		
Germany	Sledz et al. (2020) [88]	Germany	UAS	5.2 cm		✓*	✓				✓	✓		✓*	
	Sledz et al. (2021) [86]					✓*			✓			✓	✓		

This overview highlights substantial contributions that have already been made beyond the general procedural framework. Diverse algorithms have been tested for the key task of anomaly detection – most notably thresholding, saliency mapping, and local filtering – with all research groups emphasising the effectiveness of their methods in identifying regions of interest. For false alarm reduction, search space minimisation via DHS pipeline masking has emerged as a central approach, complemented by building removal and binary classification. However, despite these advances, a significant gap remains before RS-based thermography can be implemented as a large-scale monitoring solution for LD in DHSs.

All authors highlight the proficiency of their implemented methodology, in particular the developed thermal anomaly detection algorithms. However, as Table 4.1 shows, the utilised data foundations and dataset sizes differ considerably from one another. While some few approaches can be contrasted[2], a comprehensive comparison of all algorithms is as such not possible. In fact, some authors like Hossain et al. [39] explicitly report poor performance when testing other methods, such as Zhong et al. [128]'s saliency mapping, which they ascribe to differences in dataset size, complexity, and GSD. The issue of methodology

[2] For instance, Hossain et al. [39] implement Berg et al. [9]'s RF in their test of various AI-based binary classifiers.

choice extends beyond anomaly detection. Table 4.1 also reveals contention over the most suitable type of photogrammetric data processing. While Berg and Ahlberg [8] emphasise the use of mosaics solely for explorative purposes for qualitative reasons, their usage is wide-spread among other research groups such as Sledz and Heipke [86] and Sledz et al. [88]. At present, it remains unclear which methodology is most effective for the tasks at hand, hindering the optimal exploitation of TIRs for LD.

A second issue that requires improvement revolves around the aspect of automation. As indicated by the starred ticks in Table 4.1, several of the listed methods are implemented manually or only semi-automated via third-party software. This is particularly evident in preprocessing steps required for search space minimisation, namely georeferencing and mosaicking. Additionally, the GIS blueprint of DHS pipelines used for masking is also described as having been manually buffered [122, 128], as are misalignment corrections between photogrammetrically processed TIRs and GIS ground truth [86, 88]. Much of the existing work focuses on algorithmic development, giving the impression that implementations start at the image analysis stage rather than with preprocessing. This, as well as a notable lack of post-processing methods, highlights a very research-centric perspective that does not properly take the real-world goal and requirements of the approach into account.

From this follows a final unexplored aspect in the field: the holistic assessment of RS-based TLD. Publications motivate the use of thermography for this application with the negative consequences of pipeline leaks and "costly and tedious" [22] alternatives, such as repeated manual excavation. With a sole focus on methodological development, no existing study quantifies the true impact of DHS leaks and examines the viability of TLD to network operators from an economic perspective.

4.4 Thesis objectives

With the various weaknesses of the current state-of-the-art revealed, several research objectives can be defined to enable the viable use of TLD for DHS leak detection. They are centred around four key aspects:

Objective 1: *Reliability*
The most effective algorithms should be identified for each task to ensure the developed software is both robust and accurate.

Objective 2: *Usability*
To achieve a high technology readiness level, all processing steps should be fully automated. Given the non-computer science background of many network operators, this is essential for the approach to be adopted.

Objective 3: *Representativity*
The data should depict diverse scenery to enable a valid methodological comparison as well as lay the basis for broader application. This data should be shared and made open source for future use and development.

Objective 4: *Economic viability*
The TLD approach should be quantified in terms of its economic feasibility for network operators, enabling informed decisions on its adoption.

Collectively, these objectives act as a common thread for the work presented in the following sections. By addressing methodological development, automation, and real-world applicability – both technical and economic – this thesis seeks to bridge the gap between the current research-oriented TLD and its practical deployment for large-scale DHS monitoring.

5 Summary of Studies and Results

With the research gap and key objectives outlined, this section presents the five companion studies that comprise this dissertation. As such, Studies A – D primarily address **Objectives 1 – 3** (*Reliability*, *Usability*, and *Representativity*), while Study E focuses on **Objective 4** (*Economic viability*). Each of the following sections relates to a publication, providing an overview of their contributions and how these relate to the key objectives, as well as details on study-specific motivation, methodology, findings, and conclusions. Full versions of the papers are included in Part II.

5.1 Study A: Automating the analysis of thermal images to detect leaks in district heating systems

This section summarises the paper "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems", authored by Elena Vollmer, Rebekka Volk, and Frank Schultmann. It was published in the *International Journal of Remote Sensing* in 2023 and is cited here as Vollmer et al. [111]. The full version can be viewed in Part II, Section A, on page 89.

5.1.1 Context and contributions

Within the context of the current literature and research gap presented in Section 4, this study places a particular emphasis on building a complete processing software to fully automate TIR image analysis for TLD. In doing so, it contributes to several of the key objectives and existing state-of-the-art in the following ways:

Objective 1: In terms of preprocessing, developed methods include a quality indicator and comparison of photogrammetric preprocessing approaches to identify the most suitable. Additional contributions comprise the novel implementation of a traditional CV anomaly detection algorithm (triangle-histogram-thresholding (THT)), new approaches for false alarm removal based on leak features and location, and a categorisation to convey urgency.

Objective 2: A complete processing pipeline is created by automating all steps, including all afore-mentioned. This places particular focus on photogrammetric preprocessing, achieved for both individual image georeferencing and orthomosaic generation via remote access of ODM and Pix4D.

Objective 3: A novel dataset of UAS-based TIR imagery from Munich, Germany, and surrounding areas is acquired and published alongside the developed code [112].

As the given dissertation's first contribution to this field of research, Study A is closely aligned with related work (s. Section 4.1) and builds upon existing methodological findings. Consequently, implemented algorithms and novel contributions are designed to follow the general procedural framework defined in Section 4.2. They can be attributed to each analysis step as shown in Figure 5.1.

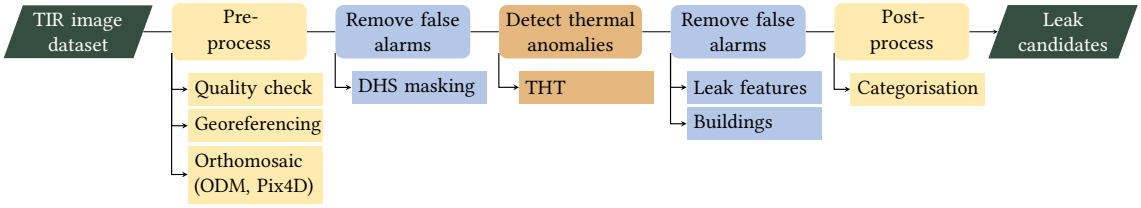


Figure 5.1: Contributions and implemented algorithms of Study A within the general analysis framework.

5.1.2 Methodology

Of the 49 UAS flights conducted in and around Munich, Germany, four sets of in total 3400 TIRs form the foundation of this study. These were selected for their urban and suburban diversity and the presence of critical and potential leaks. DJI's Matrice 600 Pro UA [91] equipped with FLIR's hybrid Zenmuse XT2 camera [92] was used to enable an almost entirely automatic acquisition at 60 m height. The latter consists of both a thermal and a visual sensor to simultaneously capture RJPEG TIRs and RGBs of 640×512 and 4000×3000 resolutions, respectively. Only TIRs are used in this study.

The fully automatic analysis software developed in this study is shown in Figure 5.2 (s. next page). Designed to process one set at a time, it can return a list of leak candidates and their geographic locations given some additional raster data inputs – specifically a DHS blueprint, a level of detail 1 (LoD1) model[1] of surrounding buildings, and a digital terrain model (DTM) of local ground elevation. To begin with, various preprocessing approaches are implemented to compensate for several of the challenges described in Section 3.2.1. After extracting the full image temperature arrays $\mathbf{T}_f$, a broad temperature range is defined based on the set's arithmetic mean and standard deviation. This limits the values to a realistic range and helps remove particularly low outliers. Next, a quality indicator is calculated per set to determine what subsequent photogrammetric processing is applicable. Photogrammetry software was found to be unable to generate viable orthomosaics from TIRs that lack distinctive features. The Canny algorithm – a standard for precise and robust edge localisation [27] – is therefore used to determine a mean edge pixel count, which allows the set to be categorised in terms of its usability for mosaicking.

[1] This simplifies buildings to basic geometric blocks with metadata such as identifiers, location, and height.

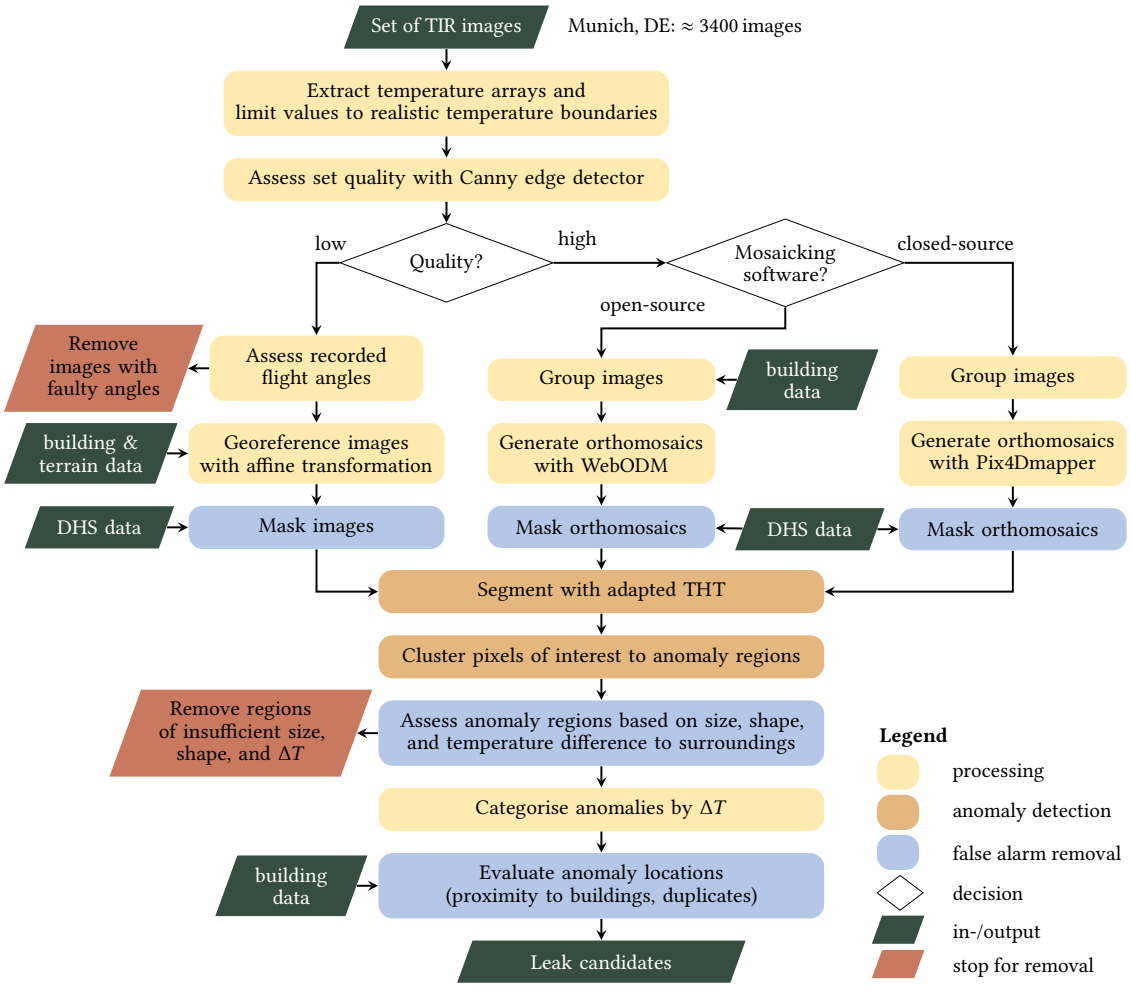


Figure 5.2: Flowchart of the analysis software developed in Study A (based on own figure originally published in Vollmer et al. [111]).

Sets that satisfy this requirement are processed by programmatically interfacing with two RS mapping software: the freely available WebODM [115] and commercial Pix4Dmapper [73]. Though orthomosaics can be generated from entire sets, inter-dataset drift (s. Section 3.2.1.2) and occasionally observed misalignment artifacts prevent their use for leak detection. These issues can be circumvented by processing in smaller batches. For WebODM, manually testing indicated that group sizes of 15 to 30 images are most suitable, with buildings used as origin points to prevent clipping of relevant fringe areas. For Pix4Dmapper, highest-quality mosaics are empirically achieved using around 30 images, without requiring the same consideration for buildings. If orthomosaic generation fails, built-in mechanisms retry their creation with progressively larger group sizes (within a defined limit) before continuing with the next batch.

Sets suffering from a lack of distinguishing features are photogrammetrically processed through automatic georeferencing. Relevant internal and external parameters are extracted from the image metadata and used to calculate GSD and image footprint. Based on these, geographic coordinates of all four image corners can be determined, which – together

with the recorded GNSS image coordinate – form a list of automatically generated GCPs. These are leveraged to estimate an affine geotransform and thereby define the image's geographical placement. Additional measures are implemented to account for undesirable aspects of UAS-based acquisition. First, the UA's flight attitude is not always recorded correctly, which leads to inaccurate rotation in georeferencing. A comparison of individually registered UA and sensor heading angles allows images with erroneous measurements to be removed automatically. Secondly, the comparatively small flight height causes high relief displacement (s. Section 3.2.1.3), which is mitigated by redefining the recorded flight height. To this end, the extracted image location is compared with the given LoD1 and DTM models to compensate for buildings or terrain changes in the immediate vicinity.

Given this photogrammetric processing on individual image or group level, the search space can be minimised to areas above and around the DHS pipelines by masking with the network blueprint [9, 22, 88, 122, 128], which is automatically buffered by 3.5 m on all sides. Masked temperature arrays $\mathbf{T}_m$ are thereby created.

Having undergone extensive preprocessing, the data are now ready for thermal anomaly detection. Working under Friman et al. [22]'s assumption that leak-related pixels reside in the upper tail of a histogram, segmentation is achieved with the THT algorithm. Instead of manually choosing a specific percentile per set, this approach automatically determines a threshold as shown in Figure 5.3 by connecting the histogram peak $P1$ with its upper tail-end $P2$ and selecting the value $P3$ with the maximum orthogonal distance d_{max}. Thresholds are calculated per image to enable drift-invariance. Adaptations tailor the algorithm to the anomaly detection task, specifically peak selection in case of multiple maxima and percentage-based threshold adjustment in case of excessive pixel selection. Pixels of interest in the resulting binary segmentation outputs are clustered to anomaly regions using a customised flood-fill algorithm.

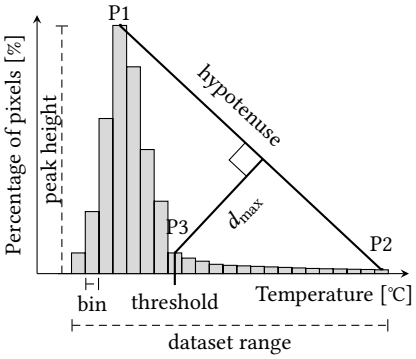


Figure 5.3: Visualisation of THT algorithm (own figure originally published in Vollmer et al. [111]).

For additional false alarm removal, novel feature-specific constraints are applied. These are based on anomaly size, ellipticity of shape, and temperature difference ΔT to surroundings. The latter value is used to perform a categorisation in terms of urgency: Candidates with a delta of 5 °C to 10 °C are defined as "potential", 10 °C to 15 °C as "definite", and more than 15 °C as "critical". An additional location analysis helps compare leak positions with one another and with surrounding buildings, identifying duplicates across images and building-related false alarms such as chimneys, vents, and entrances. The remaining anomalies form the list of leak candidates, stored in image, spreadsheet, and map format.

Method effectiveness is assessed through a comprehensive evaluation using the four case study sets. For comparative purposes, individual georeferencing is applied to all, not only sets of lower quality. Manual data screening determines relevant leak candidates and helps build a baseline for assessing overall performance. Additionally, the outputs of each algorithmic step post anomaly detection are registered for impact evaluation.

5.1.3 Key findings and discussion

This study developed an entirely automatic TIR analysis software for UAS-based TLD, focussing on general pre- and photogrammetric processing methods as well as implementation of new approaches for anomaly detection and false alarm removal.

Regarding photogrammetric approaches, the study reveals that automatic orthomosaic generation is not robust enough to be used in this capacity. Despite included quality and failure control mechanisms, both software fail to generate results at times, meaning relevant groups of images are prematurely removed from consideration. Individual image georeferencing is more reliable. Incorrect flight attitude measurements are discovered to result from fast changes in UA direction (e.g. sharp turns), which do not occur often and can be avoided in flight planning. However, no form of photogrammetric processing is precise enough to allow for location-based false alarm removal, keeping in mind that this requires very high accuracy to ensure true leaks are not unintentionally eliminated.

In terms of anomaly detection, the adapted THT reliably identifies leak candidates of various sizes, including the given critical ground truth leak (as shown in Figure 5.4). Masking minimises the search space to great effect, reducing all images by at least 40 % and 78 % on average. A comparison with manually identified leak candidates shows that this is able to proactively remove 15 % of conspicuous anomalies. The novel inclusion of sequentially applied feature constraints reduces the thousands of automatically identified regions in the masked images by at least 92 %. Remaining false alarms are found to be caused by common urban features in the urban environment, such as warm vehicles, buildings, or manholes.

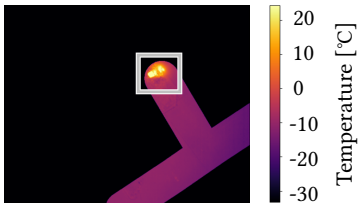


Figure 5.4: Example of the critical leak present in the study's dataset (own visualisation originally published in Vollmer et al. [111]).

While the summary highlights the proficiency of the developed analysis pipeline, several limitations remain. As the programme is tailored to the given sets, its use is limited to images acquired under similar conditions. Additionally, the data stem from a single geographical region, meaning its generalisability remains uncertain. Lastly, how well implemented algorithms – particularly for anomaly detection – perform compared to state-of-the-art methods in literature is as yet unknown.

5.2 Study B: Comparing traditional computer vision methods for anomaly detectionzur Bindung

This section summarises the paper "Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods", authored by Elena Vollmer, Julian Ruck, Rebekka Volk, and Frank Schultmann. It was published in the journal *Automation in Construction* in 2024 and is cited here as Vollmer et al. [109]. The full version can be viewed in Part II, Section B, on page 123.

5.2.1 Context and contributions

As highlighted in the discussion of related work (s. Section 4.3) and limitations of the previous study, a key gap in literature stems from the current impossibility of identifying best methodological approaches. For photogrammetric processing, this was addressed in Study A by comparing automated georeferencing and orthomosaicking. Study B focusses on the aspect of comparison for anomaly detection, the centre of the image analysis process. To this end, it contributes to the thesis aims in the following manner:

Objective 1: Regarding thermal anomaly detection, two of the most promising traditional CV methods from literature – namely local thresholding (LT) and saliency mapping (SM) – are compared to the adapted THT approach developed in Study A. Additionally, preprocessing is expanded to further improve TIR quality by including vignetting correction (VC), something that existing literature in the field does not do.

Objective 2: Closely aligned with Study A, all steps are implemented for automatic analysis.

Objective 3: Data from Munich is expanded to include new UAS-based TIR imagery from the city of Karlsruhe, Germany, which diversifies the foundation and enables generalisation. Datasets are published alongside the code [79].

In terms of the general procedural framework, this study builds upon the developments of Study A as illustrated in Figure 5.5.

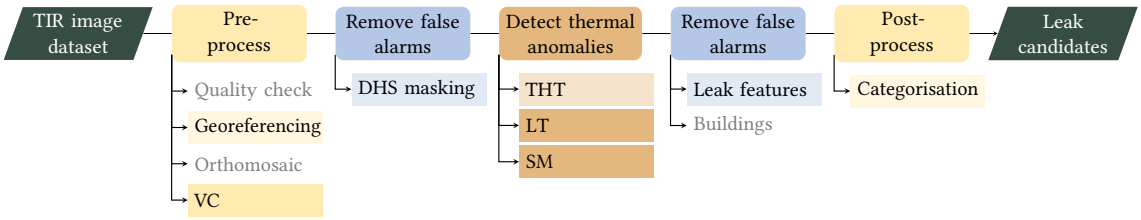


Figure 5.5: Contributions and implemented algorithms of Study B within the general analysis framework. Newly included methods are fully opaque, retained prior methods are semi-transparent, and unused prior methods are greyed out.

5.2.2 Methodology

The foundational data base for this study comprises five sets from Munich, selected for their high quality, varying levels of urbanisation, and presence of leaks and leak candidates. This is enhanced by two sets of around 1500 images from the city of Karlsruhe, Germany, to enable diversification and generalisable algorithm development. The methodology built in this study is depicted in Figure 5.6 (s. next page). To begin with, a novel preprocessing step is included to compensate for existing vignetting in all given TIR data (s. Section 3.2.1.2). Following observations from Yuan and Hua [125], a universally applicable approach is developed that does not require additional calibration imagery. Temperatures at each pixel

location across all TIRs of one flight are averaged to form a pixel-wise mean, from which a set-specific correction mask can be derived. This is applied to all extracted temperature arrays $\mathbf{T}_f$ before continuing processing according to Study A, namely limiting to set-specific temperature boundaries and georeferencing with image-specific affine geotransforms.

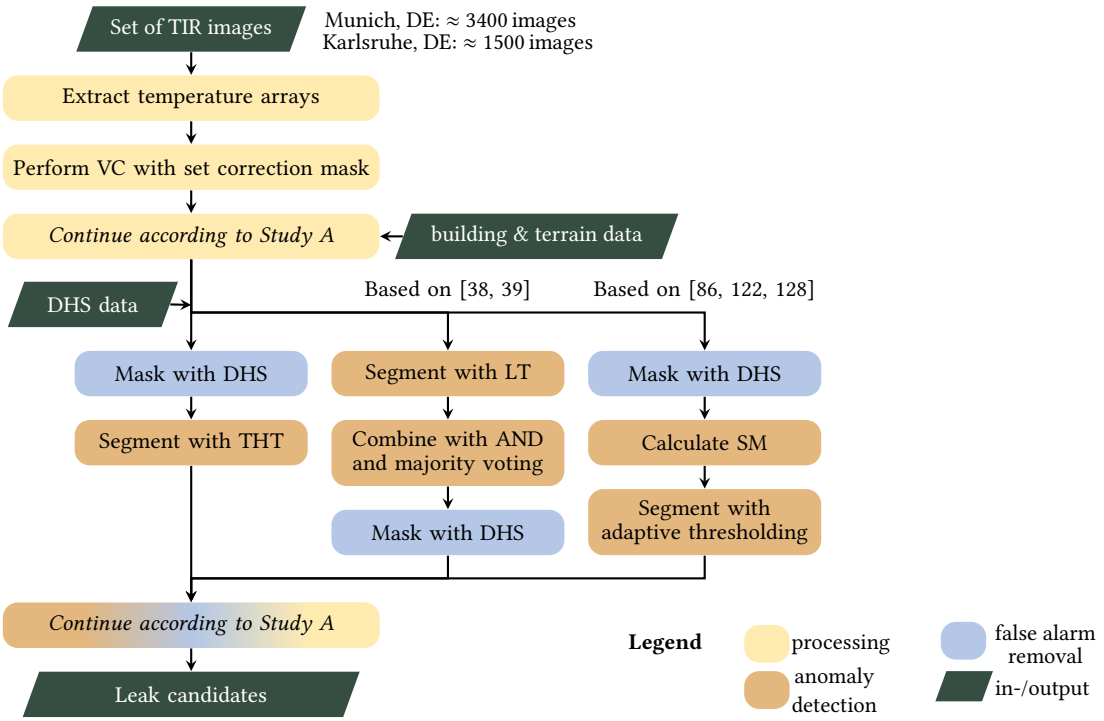


Figure 5.6: Flowchart of the analysis software including expansions from Study B (based on descriptions published in Vollmer et al. [109]).

For the task of anomaly detection, Study A's THT is compared to Hossain et al. [38, 39]'s LT algorithm of combined filter kernels and Zhong et al. [128] and Sledz and Heipke [86]'s SM approach[2]. These three present key branches of image-wise, traditional CV methods from literature. However, the implementations of the two alternate approaches require considerable adaptation to perform as intended.

While Hossain et al. [38, 39]'s LT combines a local maxima and gradient filter, an analysis reveals the latter to eliminate relevant leak candidates, such as the critical ground truth leak in the case study set. The approach is therefore adapted to use only the former temperature-based filter. Instead of simply defining the local neighbourhood with one radius, kernels are created for several radii to ensure selected pixels are relevant in smaller local as well as the entire global image context. Following Hossain et al. [38, 39], the segmentation outputs are combined by logical AND operator and smoothed with majority voting kernel, before masking with the DHS. Figure 5.7 (s. next page) exemplifies the adapted methodology.

[2] For details on method fundamentals, see Sections 3.2.2.1 and 4.1 as well as Chapter B.2.2 in Study B.

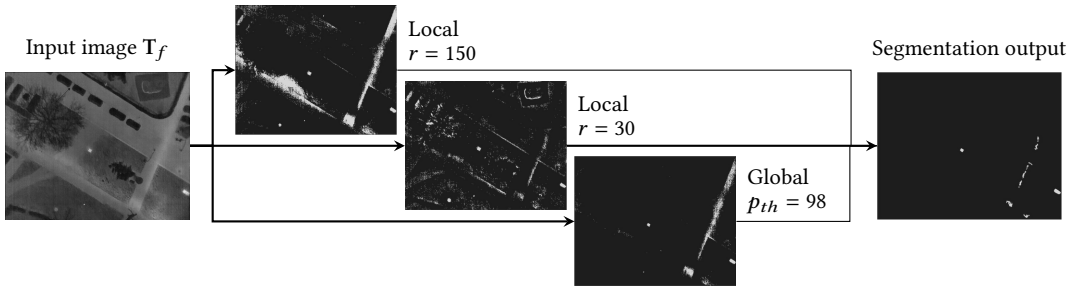


Figure 5.7: Example of the adapted LT method (based on own figure originally published in Vollmer et al. [109]). Local filters are defined by neighbourhood radii r, the global one by overall percentile p_{th}.

Despite the various existing implementations of SM, this study uncovers several methodological shortcomings. Similarly to Hossain et al. [38], experiments with Zhong et al. [128]'s combination of local and global SMs return significantly lower accuracy than described. Sledz and Heipke [86]'s customisation to suppress cold salient regions and a percentile-based normalisation to promote local maxima improves upon the general approach, however resulting maps are found to be highly sensitive to the choice of percentile and overestimate irrelevant regions. Several enhancements are therefore made. Systematic overestimation is prevented by including a reference square, a small area of increased temperature in the masked-out image area that simulates a thermal anomaly. Additionally, salient seams around cold objects are suppressed through clipping and SMs of the masked images $\mathbf{T}_m$ are calculated for initial search space minimisation. In alignment with literature, Zhong et al. [128]'s maximum entropy segmentation is used to generate binary segmentation outputs, as Sledz and Heipke [86]'s evidence theory approach requires RGBs acquired at daytime. Although initial tests find the former to be robust, the suggested threshold can cause irrelevant regions to be included or pixels of interest to be rejected. A custom adaptive thresholding approach is therefore developed that additionally leverages the saliency of the newly introduced reference square. The SM segmentation procedure is visualised in Figure 5.8.[3]

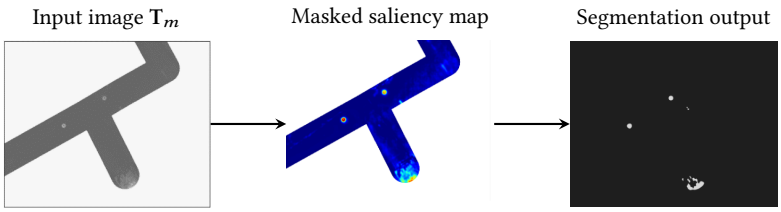


Figure 5.8: Example of the adapted SM method.

As in Study A, anomaly regions are identified through customised clustering. Using the same false alarm removal, these are again classified by relevance according to their local ΔT. Around 300 TIRs are selected from the given imagery and manually annotated to create a custom evaluation dataset. Partitioning the data into training, validation, and test splits is achieved via heuristic greedy algorithm to prevent scene overlap, while one

[3] For explicit visualisations of all of the described adaptations, see Study B's Chapter B.2.2.3.

set – containing the ground truth leak – is reserved entirely for the test data. Both of the newly adapted classical CV depend upon several parameters, such as the set of radii r in LT or the size of the reference square in SM. The training set is therefore leveraged to find the most performant variant of each, using a grid search to evaluate 448 and 20,000 combinations respectively.

In terms of metrics, R, P, IoU, and F_2 are used to assess segmentation quality, while custom detection rates DR and DR_{30}[4] indicate how many ground truth anomalies have been found. As can be expected, these run opposed: higher R or DR scores come at the cost of lower P and IoU. Four suitable parameter configurations are identified: MaxIoU (maximises IoU), MaxIoU@85/90 (maximises IoU with a R greater than 85 % or 90 %), and MaxF_2 (maximises F_2 to set a minimum R limit).

A comprehensive evaluation is conducted, comprising quantitative, qualitative, and holistic assessment. For the first, the validation and test sets are utilised, while the second leverages example imagery to highlight algorithm characteristics. The third examines suitability within the TLD context by comparing results from two representative software pipeline runs, using one set from each city.

5.2.3 Key findings and discussion

This study developed and compared three, traditional CV anomaly detection algorithms for TLD based on diverse TIR data from two German cities.

All methods are found to reliably detect major hot-spots, particularly those from critical leaks with a high local ΔT. However, they differ in their robustness for finding weaker anomalies and handling of complex scenery with strong temperature gradients. The quantitive evaluation finds THT with VC to achieve a good overall balance, with high DR (88.6 %/79.8 %) and IoU (59.8 %/47.8 %) on validation and test sets, respectively. A comparison with the variant without VC highlights the extreme importance of preprocessing

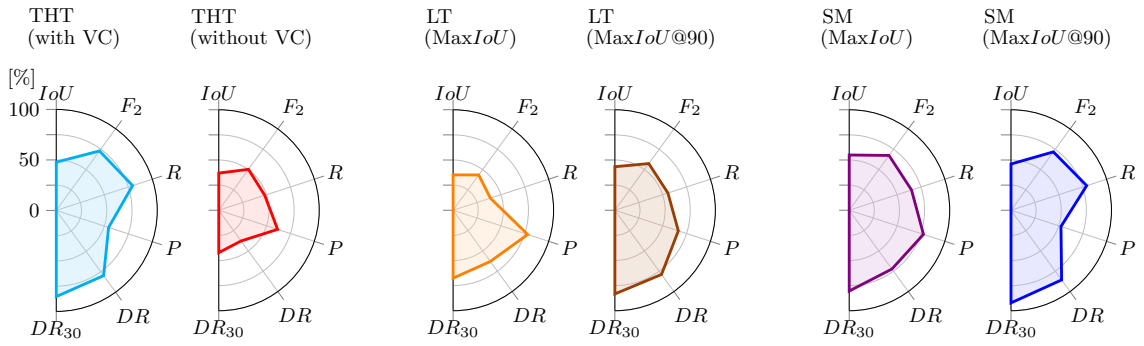


Figure 5.9: Quantitative evaluation of select algorithm variants on test set (based on data in Vollmer et al. [109]).

[4] DR is the share of anomalies in the output that have at minimum a 30 % overlap with the ground truth. DR_{30} is only considers anomalies larger than 30 pixels.

for reliable detection. In line with the *P*-*R* tradeoff, optimisations for higher *R* or *DR* generally come at the expense of lower *P* and *IoU*. Depending on the configuration, SM is able to reach a higher *IoU* (60.3 %/55.0 %) at the cost of *DR* (Max*IoU*) or the overall highest *DR* (85.7 %) while causing a significant false positive increase (Max*IoU*@90). LT performs worst overall, commonly achieving low scores and weak robustness across sets. Table 5.9 (s. previous page) visualises the results of the discussed configurations on the test set.

The qualitative evaluation reveals several interesting aspects of algorithmic behaviour. Excerpts from the most promising methods – THT with VC and SM at Max*IoU*@90 – are shown in Figure 5.10. In Example 5.10.(1), which shows a ground truth leak together with smaller hot-spots, THT identifies both the critical as well as both comparatively less significant anomalies. By contrast, SM seems to suppress smaller hot-spots in the presence of large ones. LT is found to classify only the warmest pixels within an anomaly, a behaviour that would explain the low *R* score. Example 5.10.(2), on the other hand, shows an inclination of SM to overclassify homogenous areas. This type of complex imagery also highlights how the choice of uniform threshold can have drawbacks, as THT fails to identify smaller hot-spots on the building roof. Despite this, THT is found to be the most consistently reliable method overall given the observed inconsistencies in SM and LT.

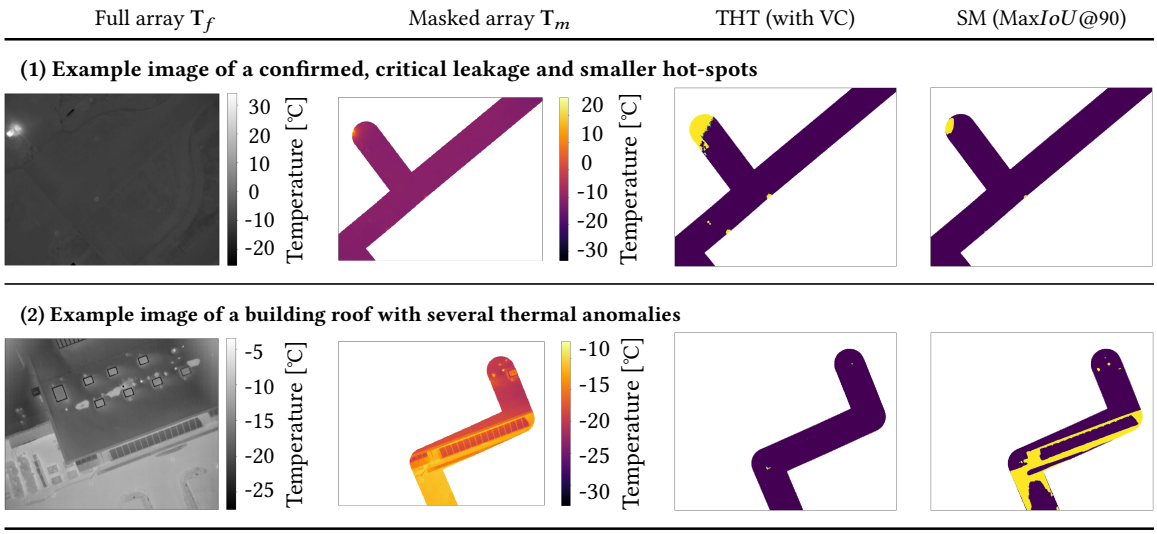


Figure 5.10: Qualitative evaluation of the most promising algorithmic variants for two example scenarios (own figure originally published in Vollmer et al. [109])). The outputs show anomalous pixels in yellow, background pixels in purple.

An evaluation of the three methods within the setting of the analysis pipeline finds all reliable at identifying the confirmed leak. The most common false alarms are caused by manholes and cars, which are identified in near equal measure across methods. While THT finds the fewest regions of interest in absolute terms, a classification by ΔT shows that those missed are irrelevant.

This study is contingent upon some limitations. Despite diverse data, there are few confirmed ground truth leaks, meaning the focus lies on anomaly detection instead of

incorporated false alarm removal for explicit and direct leak detection. While the algorithms from literature were implemented as best as possible, a lack of shared code and detailed explanations prevented their exact replication. Despite developing and using a custom labelling tool, the manually created evaluation dataset is still subject to human error. Having identified the most reliable traditional CV method as Study A's THT, the question of how AI approaches compare remains unanswered. This can be expanded to post-detection anomaly classification, to which thus far only simpler ML models have been applied.

5.3 Study C: Comparing deep learning with traditional computer vision for anomaly detection

This section summarises the paper "Leak detection using thermal imagery: Deep learning versus traditional computer vision state-of-the-art", authored by Elena Vollmer, Julian Ruck, Rebekka Volk, and Frank Schultmann. It was published in the *ISPRS Journal of Photogrammetry and Remote Sensing* in 2025 and is cited here as Vollmer et al. [110]. The full version can be viewed in Part II, Section C, on page 151.

5.3.1 Context and contributions

With Study A's THT identified as the most reliable traditional CV method among existing state-of-the-art, Study C delves into the implementation of AI, specifically DL, for the anomaly detection task and how such models compare to traditional CV methods described in Study B. The development of this DL model – coined Thermal Anomaly Segmenter (TASeg) – and associated dataset leads to several new contributions:

Objective 1: A custom pipeline for DL model training is developed for handling small, unconventional datasets, various state-of-the-art DL architectures are leveraged to identify the most suitable one, and a comprehensive evaluation in line with Study B enables a comparison to traditional CV.

Objective 2: Following Studies A and B, all steps are designed for automatic analysis.

Objective 3: A novel, binary semantic segmentation dataset for DL model training consisting of UAS-based TIR imagery from the cities of Munich and Karlsruhe is created and published [80].

Within the general framework, this publication expands upon Studies A and B as shown in Figure 5.11 (s. next page).

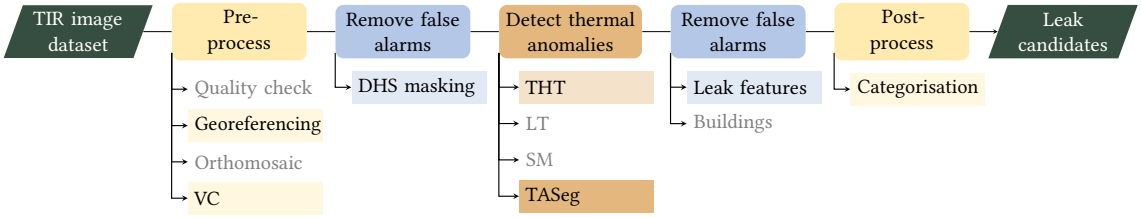


Figure 5.11: Contributions and implemented algorithms of Study C within the general analysis framework. Newly included methods are fully opaque, retained prior methods are semi-transparent, and unused prior methods are greyed out.

5.3.2 Methodology

As shown in Figure 5.12, this study adds a new methodological branch to compare a DL model – coined TASeg – to previously developed traditional CV algorithms. In terms of task specification, anomaly detection is defined here as a supervised binary semantic segmentation problem to identify hot-spots in TIRs on a pixel level. This enables both the mentioned comparison as well as further feature-based false alarm removal. Vignetting corrected TIRs from Study B form the basis of the TASeg dataset in both their full $\mathbf{T}_f$ and masked $\mathbf{T}_m$ form. For DL, the best results are found to be achieved by stacking the temperature arrays to three-channel inputs, with $\mathbf{T}_m$ doubled and $\mathbf{T}_f$ included once for context.

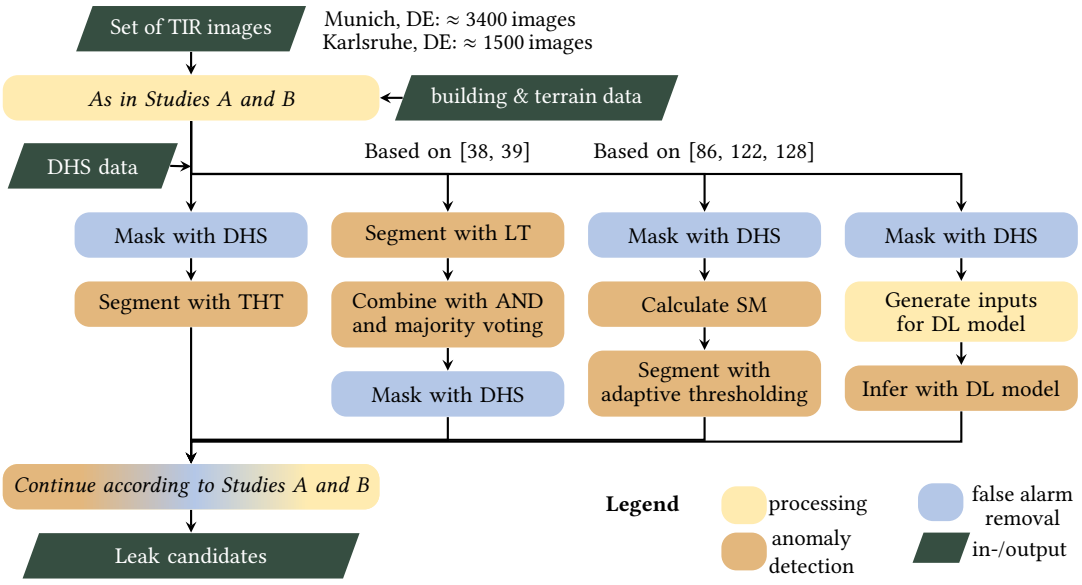


Figure 5.12: Flowchart of the analysis software including expansions from Study C (based on descriptions published in Vollmer et al. [110]).

Developing a performant TASeg model is challenging given the circumstances of this use case. Most prominently, it is characterised by an unconventional type of imagery (TIRs), a limited amount of labelled data (around 300 images), and a strong class imbalance (far more background pixels than anomalous ones). To address these challenges and leverage

all available data with minimal effort, a multi-stage, model-independent training procedure is developed. As visualised in Figure 5.13, this slowly shifts the focus from a commonly known domain (multi-class segmentation of ground-based RGBs) to the target one (binary segmentation of UAS-based TIRs). First, transfer learning is implemented to provide the model with a general direction, meaning the model is initialised with publicly available, RGB-based weights. Training is then structured into three phases to hone the model to the described target. To this end, the bipartite TASeg dataset is created. The manual part consists of a small number of laboriously annotated data – namely the evaluation set from Study B of around 300 images. To avoid exceptional labelling efforts, the remaining TIRs are used within the so-called generated set. The most performant traditional CV method identified in Study B – THT – is used to generate ground truth segmentation masks, thus enabling the creation of a large, labelled dataset. This is used in the first two phases: initially with the encoder frozen to preserve pre-trained feature extraction capabilities, then with all weights to adapt to the imbalanced, binary TIR problem. In the final and longest training phase, the TASeg model is fine-tuned on the manual set to grasp the nuances of the desired segmentation behaviour, and the variant with the highest IoU on the validation split is selected. Ablation studies find this three-phase training procedure, combining generated and manual datasets, to be highly effective and achieve the highest performance across common metrics (P, R, IoU, and F_2). Additional experiments help determine Dice and Tversky as the most promising loss functions, together with a PolynomialLR learning rate scheduler and a custom set of data augmentation transformations.

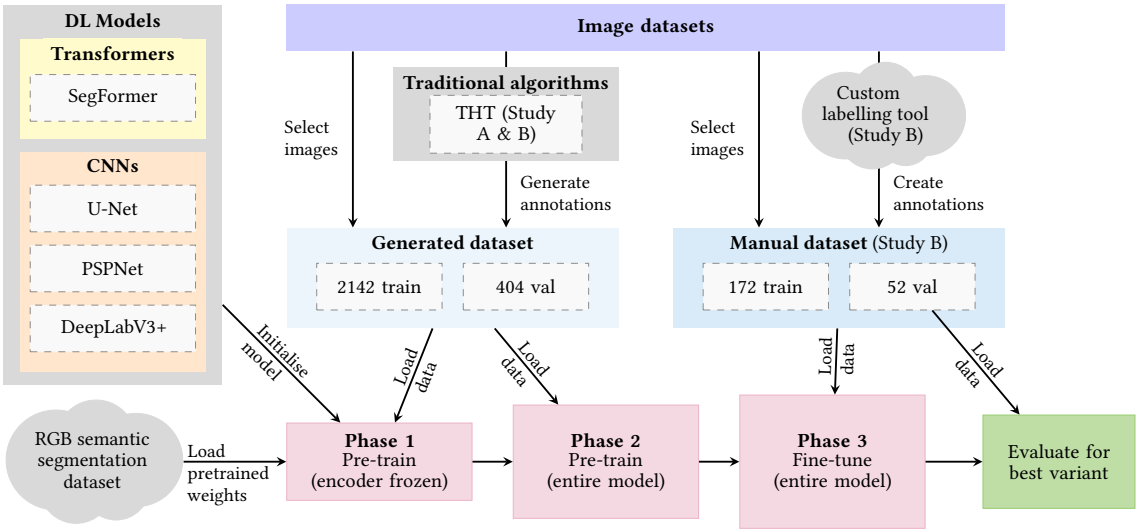


Figure 5.13: Visualisation of the developed multi-stage DL model training procedure (own figure originally published in Vollmer et al. [110]).

As Figure 5.13 indicates, the DL model itself is selected through a comparison of conventional CNN and transformer architectures. For this semantic segmentation task, these include three prominent CNNs – the U-Net [78], the PSPNet [127], and the DeepLabV3+ [14] – with a ResNet101 encoder as well as the lightweight SegFormer [120] with mix transformer encoders B0 to B4. Among CNNs, DeepLabV3+ and U-Net are most performant, with PSPNet scoring consistently lower. However, the overall best variant is identified as

the SegFormer-B2 with Tversky loss, which achieves the highest R and F_2 and second best IoU on the validation split.

This DL model is therefore chosen as the TASeg and compared to Study B's traditional state-of-the-art through an even more expansive assessment. To this end, model predictions are binarised with thresholds of 0.1 or 0.5 [1] before region clustering and feature-based false alarm removal. Given the "black box" nature of DL models, the evaluation is expanded with a visualisation-based XAI technique to provide further insights into model behaviour [37]. Specifically, the Seg-Grad-CAM variant [106] of the popular Grad-CAM approach [82] is modified to handle models based on continuous non-integer TIRs.

5.3.3 Key findings and discussion

By leveraging given data in novel ways, this study was able to train and compare various supervised semantic segmentation models for thermal anomaly detection in spite of the challenging image type, imbalance, and lack of labels.

The quantitative evaluation finds that the SegFormer excels at this segmentation task. Compared to traditional state-of-the-art from Study B, the TASeg achieves the highest IoU scores on both validation and test set, outperforming prior bests by 9.7 %pt and 6.5 %pt respectively. It is able to greatly increase P while maintaining high R and DR, indicating a substantial reduction of FPs. Overall, both SegFormer configurations commonly outperform traditional methods. Lower thresholds yield slightly better results, though this comes at the cost of considerably more detected regions and false alarms.

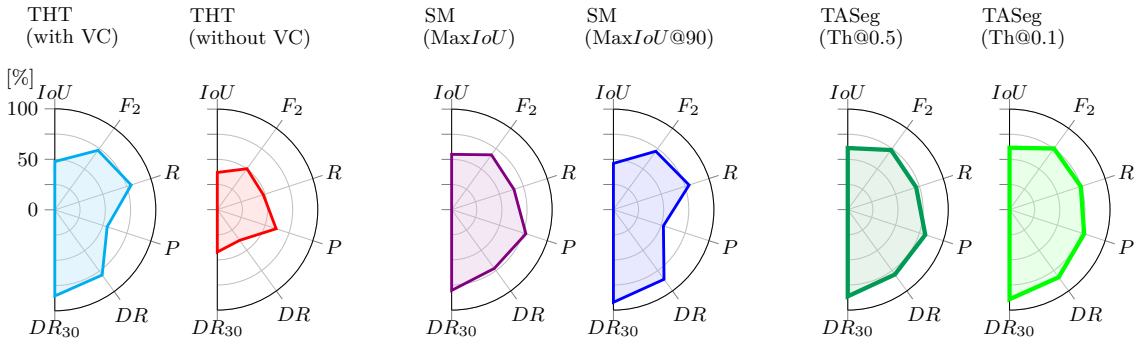


Figure 5.14: Quantitative evaluation of traditional CV and DL algorithms on the test set (based on data in Vollmer et al. [109] and Vollmer et al. [110]).

A qualitative evaluation of common leak detection scenarios solidifies the SegFormer as the most performant approach. As Figure 5.15 (s. next page) shows, the TASeg model is able to produce more robust results. Where hot-spots of different sizes are present (Example 5.15.(1)), it finds all of them, drawing precise contours without overestimating region sizes. In complex scenes that challenge THT (Example 5.15.(2)), the model successfully identifies all relevant anomalies.

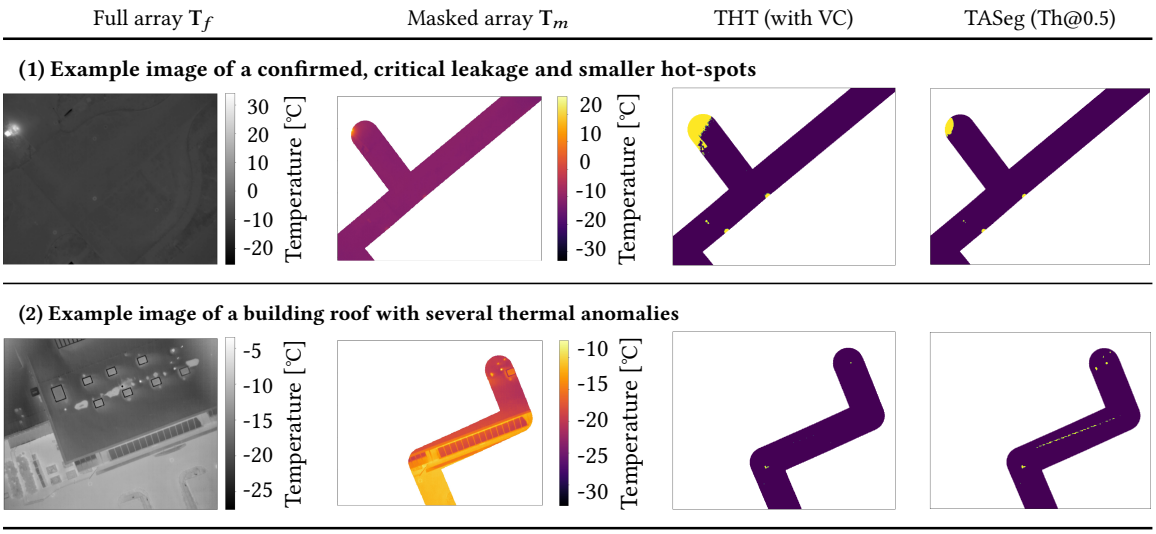
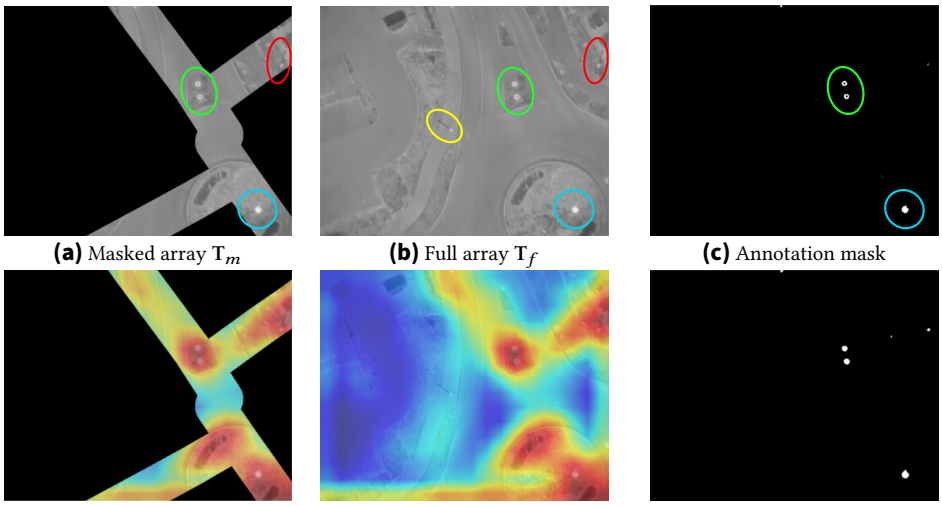


Figure 5.15: Qualitative comparison of the most promising traditional CV and DL algorithms for two example scenarios (own figure originally published in Vollmer et al. [110])). The outputs show anomalous pixels in yellow, background pixels in purple.

Additional characteristics of the TASeg are highlighted by the XAI analysis, an excerpt of which is shown in Figure 5.16 (s. next page). This confirms that the use of double $\mathbf{T}_m$ and single $\mathbf{T}_f$ array effectively focusses the model on areas above and around the DHS network – as illustrated by the disregarded street lamp on the centre left of 5.16b, circled in yellow. Where prediction (5.16f) and annotation (5.16c) differ, the Grad-CAM explanations show that this is because the model attributes similar importance to the centre (green) and bottom (cyan) manholes as the street lamps (red) in the upper right corner. The annotation is less nuanced and does not include the street lamp of lower temperature. In general, the size of the attention-rich areas indicates that the model considers anomaly surroundings for its decision, explaining its more subtle segmentation behaviour.

A final evaluation in the context of the leak detection pipeline finds the TASeg to discern more regions of interest in absolute numbers, demonstrating a more sensitive prediction approach. Detected anomalies are smaller on average, supporting previous findings of more fitted boundary definitions. The SegFormer identifies critical leaks equally reliably as the THT. Fewer manholes make the final leak candidate list due to the segmentation behaviour discerned in the XAI analysis, as the included colder centres cause some of the anomalies to fall short of the 5 °C threshold. By contrast, other common urban features are more commonly identified, such as those on building rooftops. An additional analysis of run times finds that, despite greater complexity, the TASeg model is nearly as fast as the THT. As the recorded times are still significantly longer than those listed in the SegFormer publication [120], a potential for improvement remains for real-time applications.

Some limitations apply to this study. The generalisability of the developed TASeg model to data from other geographic regions and acquisition setups is unclear. Improving the model through retraining with new TIRs requires additional computing resources. While the DL approach was found to outperform traditional state-of-the-art approaches, the



(a) Masked array $\mathbf{T}_m$ **(b)** Full array $\mathbf{T}_f$ **(c)** Annotation mask

(d) Explanation for $\mathbf{T}_m$ **(e)** Explanation for $\mathbf{T}_f$ **(f)** Prediction mask

Figure 5.16: Grad-CAM explanations for an example TIR input consisting of ($\mathbf{T}_m$, $\mathbf{T}_m$, $\mathbf{T}_f$) (based on own figure originally published in Vollmer et al. [110])). Coloured circles highlight key thermal regions. Explanations are shown as heat maps, where interest increases from blue to red.

model's more subtle segmentation behaviour emphasises the need for a more expansive and effective false alarm removal.

5.4 Study D: Implementing deep learning for false alarm removal of common thermal urban features

This section summarises the paper "Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering", authored by Elena Vollmer, Mishal Benz, James Kahn, Leon Klug, Rebekka Volk, Frank Schultmann, and Markus Götz. It was published in the *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* in 2025 and is cited here as Vollmer et al. [107]. The full version can be viewed in Part II, Section D, on page 183.

5.4.1 Context and contributions

Following Studies B and C, which focused on anomaly detection within the TIR analysis pipeline, this publication delves into the central challenge of false alarm removal. While Study A made initial strides in this direction through the automatic assessment of leak features, all studies find a considerable number of false alarms still remain that stem from common features in the urban environment. Existing publications (Section 4.1) tend to implement specific algorithms for building removal or AI – mainly standard ML – methods for binary classifications into "leak" and "non-leak" categories. In light of the demonstrated success of DL for anomaly detection (Study C), this paper instead combines

the two strategies from literature and key finding regarding the origins of common false alarms by approaching their removal as a multi-class semantic segmentation problem.

However, extending beyond model development, the main focus of this study lies on the data itself. As shown on numerous occasions in previous work, in particular Study B, data quality strongly influences traditional CV algorithm performance. Given the noise- and artefact-susceptibility of spectral TIR imagery, one might assume data preparation – known as feature engineering (FE) in AI – to be a central aspect of associated DL studies. However, a review of related literature shows the issue is often overlooked due to an application's experimental stage [32], use of benchmark datasets [65], or assumption that models themselves can compensate [12, 121]. This paper therefore examines the impact of data-related aspects on DL model performance, from the use of simultaneously acquired RGB imagery (s. Study A) to quality-enhancing filters like unsharp masking (UM) and Retinex. The development of this model – coined Thermal Urban Feature Segmenter (TUFSeg) – and associated dataset generates several new scientific contributions:

Objective 1: Central factors affecting UAS-based red, green, blue, thermal (RGBT) image quality and corrective algorithms are identified, while a comprehensive ablation study examines numerous data-related aspects by leveraging of the most widely used semantic segmentation architectures in RS. An extensive evaluation – including resource usage – helps draw best-practice rules for thermography-centred RS applications.

Objective 2: All FE and model-related algorithms are designed to enable automatic implementation.

Objective 3: A novel multispectral, multi-class semantic segmentation dataset for DL model training consisting of registered TIR and RGB images from the cities of Munich and Karlsruhe is created and published [108].

This study's contributions to the general analysis framework are depicted in Figure 5.17.

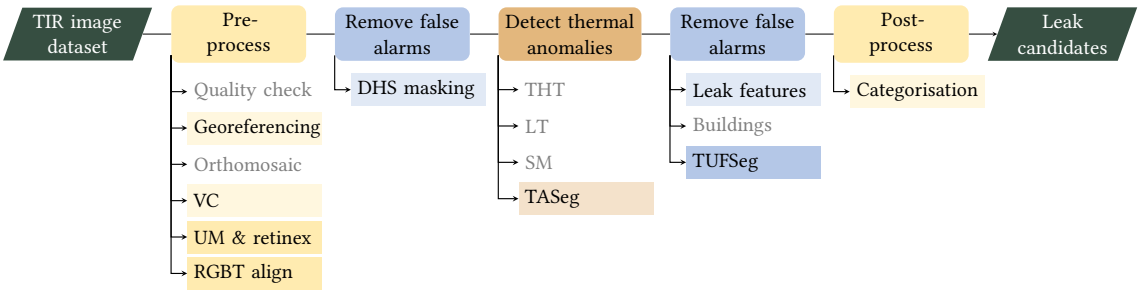


Figure 5.17: Contributions and implemented algorithms of Study D within the general analysis framework. Newly included methods are fully opaque, retained prior methods are semi-transparent, and unused prior methods are greyed out.

5.4.2 Methodology

This study is based on 12 datasets from Munich and 2 from Karlsruhe of simultaneously acquired TIRs and RGBs, chosen for their depiction of diverse and common urban scenery. For the TUFSeg dataset, these 8,452 images are narrowed down to 700 from Munich and 93 from Karlsruhe to ensure unique scene depiction without overlap. An overview of the study's developed data processing methodology and experimental setup is depicted in Figure 5.18. It consists of two main parts: Image registration to create the multispectral RGBT TUFSeg dataset and comprehensive ablation study to examine various FE aspects. While the first is centred around compensating differing sensor fields of view, aspect ratios, and data resolution, the second focusses on issues of quality and image artefacts.

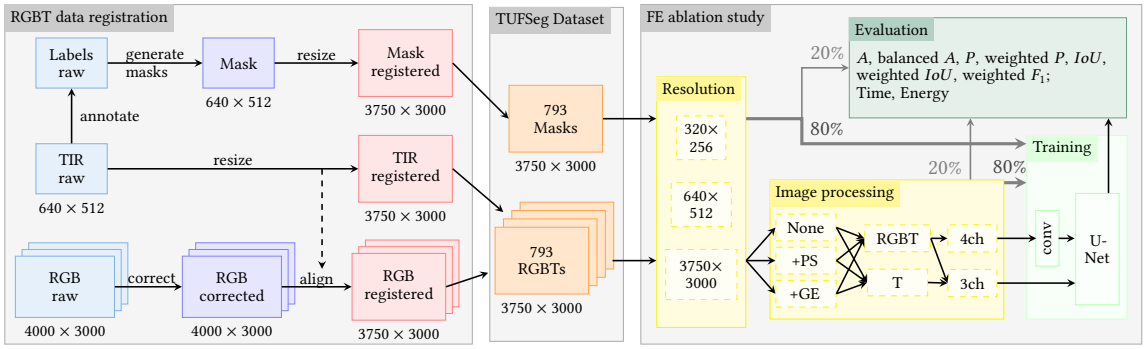


Figure 5.18: Overview of the data processing and experiment setup in Study D: From RGB and TIR image registration for RGBT dataset creation to comprehensive FE ablation study (based on own figure originally published in Vollmer et al. [107]).

As dedicated by the involved technology (Section 3.1.1), TIRs commonly have a much smaller resolution than RGBs, in this case 640 × 512 versus 4000 × 3000. The RGB sensor's greater field of view and aspect ratio mean a pixel-level alignment is necessary. However, corrective measures must be taken beforehand if distortions are present. In the case of the given data, the RGBs suffer from a radial distortion known as the fish-eye effect, which is compensated through experimentally determined intrinsic matrix and distortion coefficients. Subsequently, the offset between TIRs and RGBs is corrected by estimating a homography matrix from manually sampled tie points. When applied, this matrix transforms the data to the same resolution – in this case 3750 × 3000 to match the TIR aspect ratio and RGB resolution.

For the task of false alarm removal, nine classes of common urban features are identified: buildings, car (cold), car (warm), manhole (cold), manhole (warm), miscellaneous, person, street lamp (cold), and street lamp (warm). The TIRs are annotated with a total of 8010 polygons, whose distribution across the classes is highly uneven given the varying object occurrences. Analogous to Study C, pixel-wise predictions are required to enable the fusion of anomaly detection and false alarm removal outputs. The drawn polygons are therefore transformed into semantic segmentation masks, including an additional background class. Together with the registered four-channel RGBT imagery, these form the TUFSeg dataset depicted in the centre of Figure 5.18.

For FE, the central focus lies on addressing qualitative issues commonly affecting raw imagery. This study divides these into two aspects: platform-specific (PS), which covers unwanted effects caused by RS method (here UAS), and general enhancement (GE), which additionally encompasses sensor-related aspects. The former, as already indicated by Study B, is centred around the non-uniformity induced in TIRs due to UAS flight – namely vignetting. PS therefore consists of a per-image VC approach via radial polynomial function [6]. GE expands upon this by also focussing on the key issues of contrast and blurring. The Retinex algorithm [68], designed to handle varied lighting conditions, is found to effectively compensate for poor illumination and contrast, as exemplified in the nighttime-acquired RGBs in Figure 5.19a. For the latter, the UM algorithm helps rectify inherent blurring and prevent noise artifacts, as exemplified via the TIR in Figure 5.19b.

(a) Retinex correction exemplified with an RGB **(b)** UM correction exemplified with a TIR

Figure 5.19: Examples for GE methods (own figures originally published in Vollmer et al. [107])). Each pair shows the original (left) and the enhanced (right) image.

Applied in parallel for better results [35], these algorithms form the basis of three key image enhancement options: none, PS, and GE. Alongside these, various other aspects are investigated within the FE ablation study:

1. Three image sizes (3750×3000, 640×512, and 320×256) are tested to examine the impact of resolution,
2. The reduction of the four channel to three-channel inputs (by using RGBs converted to greyscale) helps assess the relevance of colour,
3. The impact of RGB data is investigated by comparing to the use of only thermal images as inputs.

Together with four seeds used for weight initialisation, this leads to a total of 108 model configurations. Specifically, the U-Net architecture [78] is chosen for maximum impact given its exceeding popularity for RS applications [59], prevalence in urban semantic segmentation [65], and demonstrated proficiency in multispectral data analysis [40]. Following related literature [65], a ResNet-based encoder and cross entropy loss function are selected – specifically, ResNet-152 and sigmoid focal cross entropy loss for highly imbalanced datasets. As in Study C, transfer learning is implemented for encoder weight initialisation. The model is customised with an added convolutional layer to accept four-channel inputs.

A comprehensive evaluation of both performance and resources follows the training of all model variants. The former is assessed using the most popular metrics in RGBT segmentation [90] and RS-based urban feature detection [65]: A, balanced A, P, weighted P,

weighted F_1, mean IoU, and weighted mean IoU. Balanced and weighted metrics combine scores per class through (weighted) averages. In terms of resources, time and energy consumption are measured. An additional qualitative evaluation helps characterise the behaviour of certain TUFSeg configurations.

5.4.3 Key findings and discussion

Through the developed data registration pipeline and exhaustive ablation study, this paper was able to create a first-of-its-kind RGBT dataset for thermal urban feature segmentation and analyse the impact of FE and information loss on DL model performance.

The extensive quantitative evaluation reveals several interesting findings. Overall, customised FE has an impact of between 7 %pt to 10 %pt per metric, thus disproving the notion that DL infers all engineered features on its own. The results highlight the strong class imbalance, as metrics that consider classes equally (balanced A, P, IoU) score significantly lower than those that weigh according to class size (A, weighted P, weighted IoU, weighted F_1). Specifically, these former, so-called "low" performing metrics achieve 40 % to 58 %, while the latter ("high") reach 85 % to 96 %. Generally, raw RGBTs score highest in the "low" metrics, while GE – followed by PS – excels in the "high" ones. This indicates that applied FE helps overall performance, although it benefits under-represented classes less. Channel comparisons reveal that loss of colour information has little impact, with four- and three-channel variants performing similarly. While RGBTs achieve their best results without FE, the overall strongest performance is achieved by GE using only TIRs. When relying solely on TIRs, FE can generate up to 5 %pt performance improvement. All this suggests that, contrary to assumptions, additional RGBs data are not required when custom FE is applied. An analysis of image sizes finds the highest resolutions to be most beneficial to improve "low" metrics, while mid sizes are usually sufficient for "high" ones.

A qualitative evaluation confirms these conclusions. As the excerpt in Figure 5.20 (s. next page) illustrates, the U-Net model generally excels at this form of semantic segmentation, although it struggles with less prevalent classes. Two representative models are compared: one trained on raw RGBT imagery, the other on only TIRs with full FE. Example 5.20.(1) shows the latter's finer discrimination of common objects such as warm and cold cars, actually surpassing the annotation mask in detail by associating only the hood pixels with the warm class. The RGBT-based model tends to over-classify (e.g. warm cars), which makes it somewhat more sensitive to under-represented classes such as warm manholes. Contrary to the expectation that including RGBs will improve prediction capabilities, Example 5.20.(2) shows how the raw RGBT model fails at identifying the given cold cars, while the TIR-based model correctly classifies the unobscured one. This supports the quantitative finding that select FE can outperform added RGB information.

In terms of resource usage, time and energy consumption depend primarily on image resolution, following a near-logarithmic trend with a saturation threshold between small and mid sizes. This means that high-resolution imagery not only achieves the best results

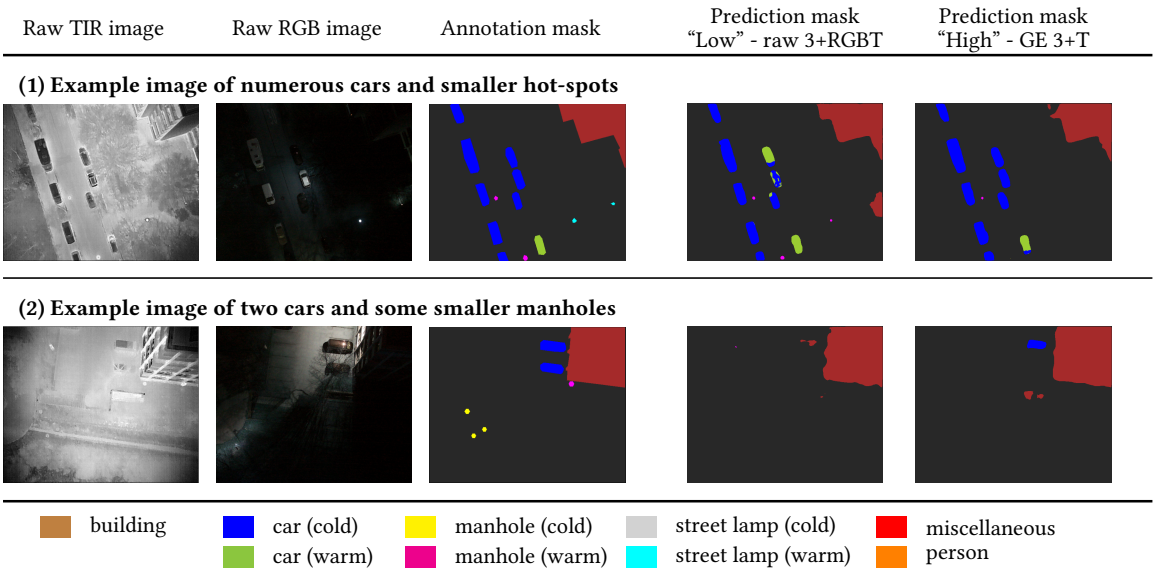


Figure 5.20: Qualitative comparison of "low" (raw 3-channel RGBT-based) versus "high" (GE 3-channel TIR-based) performing metric winners, exemplified via the model variant using resolution 640 × 512 and seed 42 (own figure originally published in Vollmer et al. [107]).

most commonly, but also requires surprisingly few additional resources. The effect of FE is minor, only increasing energy requirements and inference times slightly.

The limitations of this study include the use of only four random seeds and potential for human error in annotation and registration. Given the described findings, further experiments are necessary to isolate the contribution of individual filters and assess generalisability across models and datasets. Nonetheless, the analysis is already able to show the considerable impact of manual FE – particularly the holistic GE. The benefits of using only TIRs mean more cost-effective acquisitions using only TIR sensors can suffice in future and similar applications. While the full impact of the TUFSeg model within the leak detection pipeline is not quantified, its demonstrated proficiency at identifying common thermal urban features, in particular buildings and warm cars, can greatly improve the removal of remaining false alarms.

5.5 Study E: Assessing the economic viability of thermography-based leak detection

This section summarises the paper "Assessing the economic viability of thermography-based leak detection for district heating systems", authored by Elena Vollmer, Rebekka Volk, and Frank Schultmann. The article has been submitted for publication in a scientific journal. The full version can be viewed in Part II, Section E, on page 209.

5.5.1 Context and contributions

In contrast to the previous four publications that focused on the methodological development of the TIR analysis software for TLD, this last study delves into the economic viability of the approach in general. This is a hitherto overlooked aspect of leak detection in DHS pipelines, especially within the reviewed literature for thermography-based methods. A reason for this may be the novelty of the method and absence of sufficient data to evaluate the full economic implications, both of the approach itself and of leaks within DHSs. While new studies such as Tuikka [99] initiate the discussion by outlining the cost-effectiveness and reliability of LD used by Finnish DHS operators, such qualitative descriptions fall short of providing tangible assessments and lack an explicit focus on TLD. Against this backdrop, Study E is able to make the following contributions:

Objective 4: Newly collected empirical data from Europe's largest DHSs region are introduced and analysed, leak development and repair costs are modelled to assess lasting impacts, a first break-even analysis establishes economic thresholds for TLD viability, and results are used to generate recommendations for the adoption of LD methods.

5.5.2 Methodology

In light of the afore-mentioned lack of data, an empirical study is conducted to enable an estimation of the economic impact of both pipe leaks and TLD. The study's focus is centred on German-speaking regions given Germany's position as the largest DHS market in Europe, with both the highest pipeline share and growth [20]. Of the 35 approached DHS operators, eight German and one Swiss one contributed by providing answers to the developed survey. Given the characteristics of the interviewed operators' networks, these are categorised into four groups based on size: 1. small (L_{DHS} <30 km), 2. medium (30 km< L_{DHS} <100 km), 3. large (100 km< L_{DHS} <300 km), 4. very large (300 km< L_{DHS}). To shed a more general light on LD in practice, an additional interview was held with a representative of the German-centred integrated leak detection company LANCIER monitoring. Based on this data foundation, DHS leak costs can be estimated and contrasted with expenses for TLD to assess the method's economic viability.

In general, costs associated with leaks can be divided into ongoing – incurred during existence for the heating and treatment of make-up water – and one-time – required for the repair. Both are subject to a certain time-dependency, as the leak growth over time has an impact on how much water needs replacing as well as how extensive the required subsequent repairs will be [117]. While in practice each is contingent upon numerous DHS- and instance-specific factors, the survey responses and information from existing publications are combined to enable cost estimation.

For ongoing expenses, the growth of an exemplary, conservative leak is approximated. Given the lack of literature regarding DHS pipeline leaks, Guo et al. [29]'s study of water distribution networks is leveraged, which identifies the Richard's function to best

simulate said growth. Parameters are adapted according to the characteristics described by participants of the empirical survey and quantified observations in literature such as Fuchs and Frommhold [23]. Despite existing contracts with treatment companies, few DHS operators were able to provide definitive numbers on the cost of make-up water. The most conservative value of $5\,€/m^3$ is selected, resulting in the cost estimation for ongoing expenses in Equation 5.1 (s. next page).

$$C_{\text{ongoing}}(t) = C_{\text{make-up}} \cdot \sum_{t=0}^{T} f_{\text{growth}}(t) = 5\,\frac{€}{m^3} \cdot \sum_{t=0}^{T} \left(\frac{30\,\frac{m^3}{day}}{\left(1 + e^{(1-0.01\cdot t)}\right)^{\frac{1}{0.4}}} \right) \tag{5.1}$$

In terms of one-time costs, the interviewed operators were able to provide more information, detailing leak repairs that spanned 2000 € to over 1,000,000 €. This extremely broad range is attributed to situational complexity (urbanisation, access, time of year, and material and customer requirements) as well as the previously mentioned temporal factor. For a more exact and representative example, modelling is centred around PIRP pipelines as the most commonly installed and – as the empirical study finds – most prone to leaks. Based on operator statements, the time-dependent leak repair costs require a growth function to be delineated effectively. The established Logistic function is leveraged (Equation 5.2) to define a conservative, exemplary leak repair.

$$C_{\text{repair}}(t) = f_{\text{growth}}(t) = \frac{20{,}000\,€}{\left(1 + e^{(1.6-0.012\cdot t)}\right)} \tag{5.2}$$

These afore-mentioned costs are combined by summation to define the cumulative expenses of an exemplary, conservative DHS leak, as visualised in Figure 5.21.

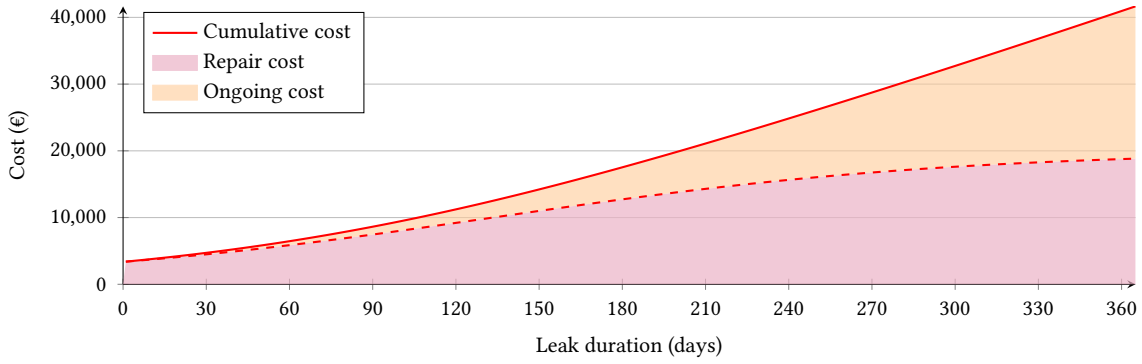


Figure 5.21: Combined cumulative ongoing and repair costs for an exemplary DHS leak (own figure from Study E).

The cost of implementing TLD can be estimated based on responses from operators who have already tested TLD as well as experience with UAS acquisition from the previous studies. An assessment of weather conditions throughout Germany finds the average number of viable days for RS-based TIR acquisition to allow for an annual coverage of all network sizes. As such, the cost of TLD as a service depends on the length of the pipelines

that need to be covered and consists of expenses for the flights themselves, operational requirements, and data analysis. This last aspect, attributed with between 25 % to 40 % for UAS flights, is currently performed entirely by hand. In light of the discussed research into automation, such as the approaches developed throughout Studies A to D, this analysis cost can be discounted. With the total service cost of UAS-based acquisition averaging at 450 €/km and the most conservative amount of savings assumed – namely 25 % – the general cost emerges as $C_{\text{TLD}} = L_{\text{DHS}} \cdot 337.5\,€/\text{km}$.

With these two costs estimated, a break-even analysis (BEA) is performed to assess the economic viability of TLD and identify its break-even point (BEP). Three scenarios are defined: 1. leaks are not actively sought and never found (Eq. 5.3), 2. leaks are not actively sought, but are identified through means that do not incur costs, i.e. a call from the public, and removed on day T (Eq. 5.4), 3. leaks are located via TLD and removed on day t_d (Eq. 5.5), where t_d depends upon the required time for TIR acquisition, analysis, and repair.

$$TC_{\text{inf}} = n_{\text{leaks}} \cdot \lim_{t \to \infty} C_{\text{ongoing}}(t) \tag{5.3}$$

$$TC_{\text{none}} = n_{\text{leaks}} \cdot \left(C_{\text{ongoing}}(T) + C_{\text{repair}}(T)\right) \tag{5.4}$$

$$TC_{\text{TLD}} = n_{\text{leaks}} \cdot \left(C_{\text{ongoing}}(t_d) + C_{\text{repair}}(t_d)\right) + C_{\text{TLD}}(L_{\text{DHS}}) \tag{5.5}$$

The number of leaks n_{leaks} assumed to occur varies for each of the four network categories depending on each DHS size L_{DHS} and leak rate. A representative network for each, with associated $t_{d_{\text{min}}}$[5], is defined for the BEA. The analysis considers a time frame of over a year, where T extends beyond that and the comparison includes $t_d \in \left\{t_{d_{\text{min}}},\ 30,\ 91,\ 183,\ 274,\ 365\right\}$ days. It is assumed that all leaks occur simultaneously and C_{TLD} costs are incurred at the start. A sensitivity analysis is included to assess the impact of automation on TLD and general economic viability.

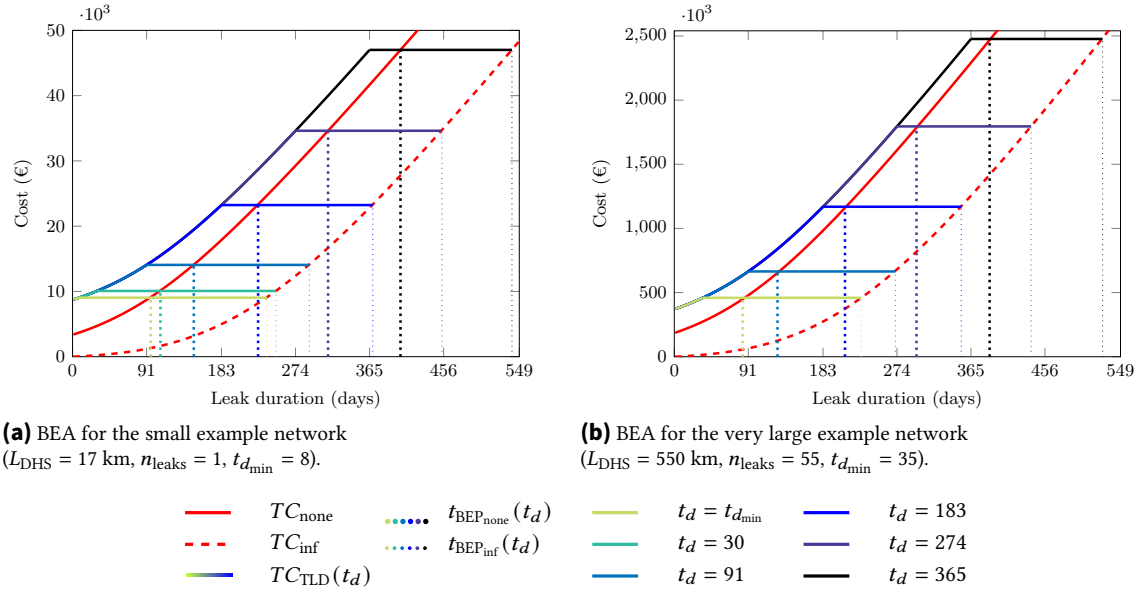
5.5.3 Key findings and discussion

This study revolved around the assessment of TLD in terms of economic viability, whereby the involved costs of both DHS leaks and TLD method were approximated using real-world data from a novel empirical study in order to perform a meaningful BEA.

Several interesting conclusions can be drawn from this analysis, an excerpt of which is shown in Figure 5.22 (s. next page). In general, the BEP is reached in a surprisingly short amount of time – less than three months after t_d compared to TC_{none} and less than a year for TC_{inf}. The comparison between Figures 5.22a and 5.22b shows how TLD pays off faster with increasing network size, indicated by the smaller horizontal distances between both red and all aqua-coloured lines. Looking at, for instance, the darkest blue (where detection

[5] The duration for TIR acquisition depends on the network size L_{DHS}, while analysis and repair are assumed as requiring one week at minimum.

and repairs occur in 365 days) shows $\Delta t_{BEP_{none}}$ is reached in 41 days in the small network, but achieved after just 25 days in the very large. This can be attributed to leak rates, which the empirical study found grow with DHS length. Regardless of network size, amortisation speed increases with t_d, as seen by the shortening horizontal distances between aqua-coloured and red lines. In the small network (Fig. 5.22a), for instance, a comparison of the green (t_d = 30 days) and darkest blue (t_d = 365 days) repair scenarios shows how $\Delta t_{BEP_{none}}$ is halved from 82 to 41 days. This stems from the non-linearity of ongoing and repair costs. TLD can therefore remain competitive despite a growing probability of localisation by other methods over time. This fact complements its methodological strength of being able to find more obscure leaks which are not easily located quickly by other means.



(a) BEA for the small example network (L_{DHS} = 17 km, n_{leaks} = 1, $t_{d_{\text{min}}}$ = 8).

(b) BEA for the very large example network (L_{DHS} = 550 km, n_{leaks} = 55, $t_{d_{\text{min}}}$ = 35).

Figure 5.22: BEAs for two different network sizes (own figure from Study E). Red lines depict the reference scenario costs, aqua-coloured lines the TLD costs for $t_d \in \{t_{d_{\text{min}}}, 30, 91, 183, 274, 365\}$ days, with the bends in the curves indicating when t_d has been reached. To show when TLD costs amortise, the BEPs are marked by dotted vertical lines – thick for $t_{\text{BEP}_{\text{none}}}$, thin for $t_{\text{BEP}_{\text{inf}}}$.

The accompanying sensitivity analysis finds that, aside from reducing TC_{TLD}, automating the TIR analysis shortens the amortisation period by a percentage that matches or even exceeds said reduction. Even a conservatively estimated automation can therefore generate exceedingly high savings, which can be expected to increase further given drops in other shares such as UAS flight costs. An aspect that the quantitative analysis omits yet DHS operators mention favourably is the additional advantage of acquiring TIRs, namely their use in providing insight into general network and pipeline conditions beyond only LD.

Several conclusions and recommendations can be derived from the combined empirical study and BEA. Key findings include:

1. Even conservative leaks cause considerable expenses, emphasising the need for systematic LD which – according to the interviewed operators – is currently not always implemented as required.

2. Operators should especially take ongoing costs into consideration (s. Fig. 5.21), which the empirical study shows are often underestimated.
3. The interviews reveal how traditional LD and existing integrated systems suffer from poor usability, incorrect or irregular usage, and malfunction issues, highlighting the need for complementary techniques.

In this, TLD is found to be a promising option both terms of its capabilities as well as economic viability as demonstrated by the BEA. TLD also offers additional benefits that are not easily quantifiable, such as its use for DHS condition assessment and temporal analyses to identify network degradation. When considering its adoption, several network-specific aspects must be taken into account, including installation depth and length, leak rate, and TLD service cost.

This study is limited in several ways. Given the focus on German-speaking countries, findings may not be generalisable to other regions, especially outside of Europe. The methodology is built upon a comparatively small empirical study, and derived assumptions and illustrative examples cannot depict the full breadth of possible occurrences. Instead of modelling only a single leak, different variants of growth and size could be included in future, as well as other time-dependent aspects such as seasonality and interim repairs. Quantifying the economic impact of other LD methods and expanding the study to other countries can create an even more comprehensive analysis and solidify findings.

6 Discussion

Having summarised and assessed the individual studies that comprise this dissertation in Section 5, this chapter takes a broader perspective to discuss the thesis as a whole. To this end, contributions and implications are first portrayed in relation to the four thesis objectives, before the universal limitations and benefits are presented.

6.1 Contributions and implications

As outlined in Section 4, the central aim of this thesis lay in advancing TLD as a method for large-scale DHS pipeline leak monitoring. Section 4.4 defined four key objectives to bridge the gap between current research-centred state-of-the-art and the requirements of real-world adoption. While each was addressed in the individual study summaries, they are now considered more universally. Alongside these descriptions, Table 6.1 on page 64 visualises the new state-of-the-art by placing the first four study contributions within the context of existing literature, as detailed in Section 4.1 and initially portrayed in Table 4.1.

Objective 1: *Reliability*

The first objective is centred around algorithm quality and the use of accurate and robust methods to ensure a reliable automatic analysis. This is necessary to create a foundation of trust essential to the adoption of any new technology. Studies A, B, C, and D all contribute to the objective by identifying performant approaches for the core tasks of automatic TIR image evaluation: data processing, anomaly detection, and false alarm removal. As Table 6.1 illustrates, this is achieved through a combination of method development and holistic comparisons.

A central part of this thesis is the emphasis placed on data preprocessing, cementing it as a fundamental and highly relevant part of the analysis pipeline. Various algorithms are implemented to improve data quality to great effect – as shown by VC in Study B and the comprehensive enhancements in Study D. This thesis is able to address the contention regarding photogrammetric processing in Study A by comparing approaches and finding individual image georeferencing to be more robust and reliable than orthomosaicking. For the central task of anomaly detection, Studies A, B, and C together are able to paint a comprehensive picture by developing and comparing a range of methods. While Study A's THT is found to be the most reliable amongst promising traditional CV algorithms,

Study C reveals that DL methods can outperform conventional approaches through the development of the TASeg model. For false alarm removal, Studies A and D both contribute by presenting effective approaches. Study A refines DHS pipeline masking and introduces feature-based sorting by size, shape, and temperature difference ΔT, while Study D combines ideas from literature and prior findings by creating the TUFSeg model to sort out false alarms prevalent in the urban environment. Method development is rounded off by Study A's introduction of a post-processing ΔT-based categorisation to attribute the resulting leak candidates with a measure of importance.

Together, the studies and their contributions not only further the state-of-the-art in this research field, but increase confidence in the approach by establishing a methodologically reliable TIR analysis.

Objective 2: *Usability*

The second objective focusses on the aspect of ease of use as an essential factor to the adoption of TLD. As Study E highlights, the analysis of TIRs by hand presents a challenge to operators in terms of resources and expertise. Automation can greatly alleviate the effort, but requires careful consideration. Existing literature, for instance, only addresses it for the steps of anomaly detection and false alarm removal, disregarding the consequent effort and obstacle to technology adoption this creates.

As Table 6.1 demonstrates, this dissertation ensures Objective 2 is considered throughout. While all the first four studies contribute by adhering to the overarching requirement, none enables it more than Study A through the automation of the previously manually implemented photogrammetric processing. By leveraging building and terrain data, the robust individual image georeferencing approach is enhanced to programmatically remove displacement distortions, thereby offering a strong approximation of manual corrections. Other preprocessing steps, such as quality assessment (Study A) and improvement algorithms (Studies B and D) are also designed with this aspect in mind. Beyond automation, the software outputs are considered as well. The implemented severity categorisation and mapping of leak candidates provides network operators with results in an easily interpretable way, thus enabling increased trust and faster decision making.

While additional inputs are necessary to enable similar performance to manual implementations, the full automation and interpretable outputs developed in this thesis marks a substantial step toward the adoption of TLD.

Objective 3: *Representativity*

The third objective is centred around a solid data foundation and key to ensuring a valid methodological formation and evaluation. This is even more important given the current lack of shared data by existing literature.

To this end, the thesis bases all its developments on newly created, real-world datasets acquired in Europe's leading DHS market [20]. Study A begins with a data foundation

of TIR images in and around the urban area of Munich – including the presence of a confirmed and critical leak. In Studies B, C, and D, this is expanded to the city of Karlsruhe. Combined, the data thereby exemplify varied levels of urbanisation, scene diversity, and – given fluctuating acquisition conditions – image quality. Each study uses several hundred to several thousand TIRs, which in D is expanded to include RGBs, thus forming a first-of-its-kind multispectral UAS-based RGBT dataset. Meanwhile, Study E encompasses an entirely new empirical study, providing insights into the current condition of DHSs, management strategies, and the potential for TLD adoption in practice. This foundation enables the methodological developments and comparisons in Objectives 1 and 2 that would have otherwise been impossible given the current lack of shared or existing data. By publishing the created datasets alongside the studies [79, 80, 108, 112], this dissertation ensures higher transparency and reproducibility.

While generalisability is still limited owing to the singular geographic focus and unvaried acquisition style (s. Table 6.1), the size, real-world character, and accessibility represent a substantial step forward and provide the first foundation for broader application and future TLD development.

Objective 4: *Economic viability*

The final objective is focused on assessing the TLD approach in terms of its economic viability – a crucial aspect to the technology's adoption given that DHSs operate as centralised heat suppliers and should act upon a sound economic basis. As Table 6.1 implicitly indicates, literature to date has disregarded this aspect in favour of methodological development.

Study E is therefore the first to do so by estimating the costs of both an exemplary leak and TLD as a UAS service based on the afore-mentioned novel empirical study. The analysis not only reveals the economic and practical need for such LD methods, but showcases how quickly the cost of TLD can amortise. Despite the regional focus and required assumptions, this thesis is able to confirm the approach as an economically viable option.

Table 6.1: Comparison of existing literature's data foundation and implemented methods with this dissertation's contributions.
Legend: ✓ = automatic implementation, ✓* = semi-automatic or manual implementation, (✓) = implemented but not used in final analysis.

Research groups		Data foundation			Analysis algorithms																
		Region	Aerial vehicle	GSD	Data quality (i.e. VC)	Georeference	Orthomosaic	Mask with DHS	Intensity histogram	Triangle histogram (THT)	Saliency mapping (SM)	Local filters (LT)	Blob detector	DL	Building removal	AI binary classifier	Categorisation	Leak features	Temporal analysis	Severity evaluation	Leak mapping
Sweden	Friman et al. (2014) [22]	Sweden & Norway	Airplane	20 cm		✓*	(✓*)	✓	✓						✓						
	Berg et al. (2016) [9]					✓*	(✓*)	✓	✓						✓	✓			✓		
China	Xu et al. (2016) [122]	Sweden & China	Airplane & UAS	19 – 24 cm		✓*		✓			✓										
	Zhong et al. (2019) [128]					✓*	✓*	✓			✓										
Denmark	Hossain et al. (2019) [38]	Denmark	UAS	-								✓				✓					
	Hossain et al. (2020) [39]											✓				✓					
Germany	Sledz et al. (2020) [88]	Germany	UAS	5.2 cm			✓*	✓					✓		✓		✓*				
	Sledz et al. (2021) [86]						✓*				✓				✓	✓					
Germany	Study A (2023) [111]	Germany	UAS	7.8 cm		✓	✓	✓		✓								✓		✓	✓
	Study B (2024) [109]				✓	✓		✓		✓	✓	✓						✓		✓	✓
	Study C (2025) [110]				✓	✓		✓						✓				✓		✓	✓
	Study D (2025) [107]				✓											✓					

6.2 Critical appraisal

While this dissertation makes various contributions to both existing state-of-the-art and the four central objectives, it is also subject to several limitations. The following discussion goes beyond the restrictions outlined in the individual study summaries in Section 5 to provide a broader critique. Additionally, this critical appraisal includes general benefits of the thesis transcending the afore-mentioned objective-centric ones.

6.2.1 Limitations

The overarching limitations of this thesis can be considered on different levels: those inherent to TLD itself, those arising from methodological choices and data availability, and those relating to the versatility of the developed analysis.

General As a leak monitoring method, TLD is generally limited by its reliance upon the repeated capture of TIR imagery, which is subject to specific acquisition requirements (s. Section 3.1.3). Given that the findings of this thesis frequently highlight the importance of data quality, the aspect becomes all the more critical in ensuring that the methods and developed analysis can be used reliably. This places constraints on the implementation of TLD as a large-scale monitoring approach, as it requires DHS operators to ensure acquisition conditions and frequencies are met.

Methodology Specifically addressing the developed automatic TIR analysis, several limitations need to be considered. As Table 6.1 and previous descriptions emphasise, chosen methodology was often motivated and grounded in existing literature. However, missing and incomplete descriptions as well as a lack of shared code in all prior work complicated genuine replication and method development. This was particularly challenging in Studies B and E, where numerous assumptions had to be made to compensate for insufficient or lacking information. These deductions, though carefully reasoned, could not always be tested for validity on a broader scale and may not be universally applicable. For the central analysis tasks, the two developed DL models are likewise limited in interpretability due to their black-box nature and dependence on manually annotated data, which are inherently subject to some degree of human error and ambiguity.

While the individual analysis steps were evaluated comprehensively throughout the studies, the entire pipeline with all integrated algorithms was not conclusively assessed in its final form. A true evaluation would require the application to new data from hitherto unexplored DHSs, so as to fully analyse effectiveness and generalisability. Despite the significant strides made in increasing the technology readiness level, the approach still requires comprehensive testing to this end.

Data These aspects highlight a recurring theme across all studies: the geographic, scale, and acquisition-related limitations of the given data. The restriction to only a few DHSs within one country and fixed platform and sensor parameters particularly affects generalisability – both of the methodology and underlying assumptions. The lack of existing or shared data, in particular by previous research groups (s. Section 4.1), made further diversification difficult, as it required the sole reliance upon self-acquired datasets. However, given its consistent focus on Germany as the largest DHS market in Europe, the thesis can be understood as an initial analysis guideline, based on a prominent use case. Investigating its applicability to other regions and varying acquisition parameters, as well as different RS platforms would be very interesting for future work to assess transferability and versatility.

Customisation A final limitation concerns the aspect of versatility. With a focus on robust and reliable method development and automation, various avenues for enhanced approaches and customisation remain unexplored in the implementation's current form – for example temporal analyses similar to Berg et al. [9]. In instances where little data is available to DHSs, the current dependency on additional input information – building and terrain information for georeferencing and DHS blueprints for masking – may be a hindrance. Further research could explore replacing these with less exact alternative approaches, such as extracting streets from images [128] and OSM data usage. Where DHS operators have access to more detailed information, false alarm removal could be enhanced to automatically include a comparison to network components such as maintenance shafts. For some steps, such as the ΔT leak categorisation, conservatism could be parametrised to depend on network and soil characteristics that impact heat propagation – thus taking into account different influences on how leaks manifest in TIR imagery.

6.2.2 Benefits

Next to the afore-mentioned limitations, the thesis includes some overarching benefits beyond the objective-specific contributions presented in Section 6.1. Grouped thematically, these encompass datasets, methodological aspects, and the potential for broader applicability.

Datasets While the data-related advantages and disadvantages have been discussed at length, the preparation, creation, and publication of real-world TIR imagery lays the foundation for the first community-wide benchmark for in-field evaluation. This extends beyond the use for TLD-based automatic analyses, as two DL datasets for UAS-based semantic segmentation were developed and published. Consisting of several hundred to thousand images and annotation masks, both the TASeg dataset of binary TIRs and multiclass, multispectral RGBT TUFSeg dataset can be leveraged for all manner of analysis tasks, such as thermal fault identification, urban land classification, and object counting. Given that very few such datasets are publicly available, especially ones including a distinction

by thermal state, this thesis contributes initial benchmarks to the field of thermography- and UAS-based semantic segmentation.

Data preprocessing Although the focus on data quality solidified stringent TIR acquisition as a necessary requirement and limitation of TLD, the studied effects of preprocessing have a wide-reaching impact. Simple corrective measures such as VC in Study B can easily be transferred to other RS-based thermography applications, enabling more effective data usage and improved performance. Assessing data quality can expand possibilities, as demonstrated by Study D, in which data enhancement was found to allow a cheaper acquisition. These aspects are relevant to all forms of TIR analysis, not only this dissertation's TLD application.

Deep learning The development of the DL models also allowed for some general contributions to less common applications in the field of AI. The multi-stage training procedure developed in Study C for tackling the problems of unconventional data, limited annotation, and class imbalance is both task- and model-independent, making it easily transferable to similarly challenging applications. The adaption of a common XAI method to the peculiarities of TIR data not only helped provide explanations for prediction behaviour, but also highlighted the relevance of such analyses in the field overall. Similarly, the inclusion of energy usage as a metric in Study D puts a spotlight on this increasingly important side of AI. By emphasising these aspects despite their lesser prevalence in the given application-oriented and fringe field, this thesis contributes to a more comprehensive approach to AI.

Further potential While the TIR analysis pipeline was developed for network operators, it can provide a foundation beyond its use for DHS-centred TLD. Given the growing increase in TIR fault detection, it has the potential to be leveraged by companies offering general RS services, such as UAS-based analyses. Interesting scenarios range from photovoltaic [94] and power line [81] inspection to leak detection in oil and gas pipelines [103]. As the designed image analysis of this thesis is focused on DHS leaks, a transferral will likely require adaptation given differing anomaly characteristics, false alarm removal strategies, and acquisition condition requirements. Nevertheless, both the individual components and the overall procedure may provide benefit and guidance to these kinds of use cases.

7 Summary and Outlook

Amid the ongoing energy crisis, DHSs have the potential to reduce building-related emissions and to become a cornerstone of climate-neutral urban heating. However, the buried pipe networks have to contend with considerable losses owing to a lack of comprehensive leak monitoring strategy. TLD has emerged as a non-invasive, contact free, and technologically viable option for DHS assessment. Its large-scale implementation can be achieved by leveraging RS platforms such as UASs, though at the cost of tens of thousands of TIR images that need evaluation. While some research groups have begun developing automatic analysis procedures to this end, the required effort and expertise still remain a key obstacle to the approach's adoption. The present dissertation addresses the challenges of enabling UAS-based TLD by defining four key objectives: *Reliability* (developing effective algorithms), *Usability* (ensuring ease of use through automation), *Representativity* (basing developments on diverse data), and *Economic viability* (assessing viability in economic terms). Five studies are conducted to expand upon existing literature, contribute to the objectives, and work toward bridging the gap between scientific research and real-world implementation. Studies A – D focus on *Reliability*, *Usability*, and *Representativity*, while Study E is centred around *Economic viability*.

Study A places an emphasis on *Usability* by building an entirely automatic analysis procedure. Data preprocessing techniques are combined with a novel implementation of THT for anomaly detection and custom false alarm removal. Based on newly acquired UAS-based TIRs from Munich, the most adept photogrammetric preprocessing is identified. THT reliably detects thermal anomalies and leak feature-based filtering effectively reduces candidate lists. Remaining false alarms are found to stem from common warm objects in the urban landscape.

With a special focus on *Reliability*, Study B's main objective lies in finding the most suitable anomaly detection method for the analysis pipeline. After enhancing data preprocessing with VC, the THT method is compared with traditional CV techniques from literature – SM and LT. Based on a diversified dataset including TIRs from Karlsruhe, Study A's THT is found to be the most robust, particularly due to the included VC.

While Study B concentrates on traditional CV for anomaly detection, Study C instead leverages DL models. A novel, model- and task-agnostic training procedure is developed to handle the challenges of unconventional data type, lacking annotations, and class imbalance. Trained on the newly developed TASeg dataset, the SegFormer emerges as the most performant architecture. A comparison with Study B shows this TASeg model to be the most reliable for anomaly detection.

To address the other central analysis task of false alarm removal, Study D builds upon previous findings. Given the proficiency of DL, a multi-class U-Net model is developed to eliminate common false leak candidates based on the newly created, multispectral TUFSeg dataset. A comprehensive ablation study shows that using only TIR images and all enhancements achieves the overall best performance – once again highlighting the central role of data quality.

Focused on the objective of *Economic viability*, Study E conducts a new empirical study of German-centred DHSs, models the time-dependent costs of an exemplary leak, and estimates the expenses of UAS-based TLD. The first break-even analysis in DHS leak detection identifies TLD as viable option in economic terms as well, especially when including automatic TIR analysis. Given the identified practical shortcomings of traditional LD like integrated systems, the study also underscores the role TLD can play in closing this technological gap.

Combined, these five studies make significant strides in enabling the adoption of RS-based TLD as a large-scale method for DHS leak monitoring. In light of these contributions, several opportunities for enhancement remain. Future work can improve upon the developed analysis procedure – for instance by including temporal analyses to identify trends or enabling customisation depending on the data available to the operators. While the conducted image acquisitions and empirical study forged initial connections to the industry, further discussions and a closer collaboration could help identify additional requirements and improvements for the analysis. Given the UAS- and German-centric focus of this dissertation, a highly interesting avenue for future research would be assessing and enhancing algorithmic generalisability by expanding the data foundation – both in terms of geographic location and utilised acquisition platforms and parameters. Broadening the economic analyses to other regulatory and market contexts would also help assess the TLD's wider viability. Using new datasets, the two developed DL models could be continuously improved – for instance through active learning loops where new predictions are used as labels for retraining under minimal human supervision [76]. By dedicating further research to related aspects like drift detection – monitoring model performance to detect data shifts that cause degradation [72] –, federated learning – training a central model on private datasets that are not shared [19] –, and XAI, potential issues of trust and underperformance can be prevented and mitigated. On this basis, the developed automatic TIR analysis approach for DHS leak detection could be incorporated into UAS-based smart city approaches and thus contribute to managing sustainable cities of the future.

Bibliography

[1] Alkan, D. and Karasaka, L. (2023). "Segmentation of Landsat-8 Images for Burned Area Detection with Deep Learning". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-M-1-2023, pp. 455–461. DOI: 10.5194/isprs-archives-XLVIII-M-1-2023-455-2023.

[2] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Hauptbericht 2022 [Main report 2022]*. Report. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht (visited on 25 Aug. 2025).

[3] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Verfahren zur Zustandsermittlung von Fernwärmeleitungen und zur Feststellung / Einmessung von Abweichungen (Leckortung) [Methods for Assessing the Condition of District Heating Pipelines and for Locating Deviations (Leak Detection)]*. Technical report FW 435. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw-shop.de/regelwerk/fw-435c-verfahren-zur-zustandsermittlung-von-fernwarmeleitungen-und-zur-feststellung-einmessung-von-abweichungen-leckortung-druckfassung.html.

[4] Arbeitsgemeinschaft QM Fernwärme [Swiss Working Committee on District Heating] (2021). *Planungshandbuch Fernwärme [Handbook on Planning of District Heating Networks]*. Handbook. Version 1.3. Bern, Switzerland: Bundesamt für Energie [Swiss Federal Office of Energy]. URL: https://www.verenum.ch/Dokumente/PHB-FW_V1.3.pdf.

[5] Axelsson, S. (1988). "Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 26(5), pp. 686–692. DOI: 10.1109/36.7695.

[6] Bal, A. and Palus, H. (2023). "Image Vignetting Correction Using a Deformable Radial Polynomial Model". In: *Sensors* 23(3), p. 1157. DOI: 10.3390/s23031157.

[7] Bayomi, N. and Fernandez, J. E. (2023). "Eyes in the Sky: Drones Applications in the Built Environment under Climate Change Challenges". In: *Drones* 7(10), p. 637. DOI: 10.3390/drones7100637.

[8] Berg, A. and Ahlberg, J. (2014). "Classification of leakage detections acquired by airborne thermography of district heating networks". In: *8th IAPR Workshop on Pattern Reconition in Remote Sensing*. Stockholm, Sweden, pp. 1–4. DOI: 10.1109/prrs.2014.6914288.

[9] Berg, A., Ahlberg, J., and Felsberg, M. (2016). "Enhanced analysis of thermographic images for monitoring of district heat pipe networks". In: *Pattern Recognition Letters* 83, pp. 215–223. DOI: 10.1016/j.patrec.2016.07.002.

[10] Bishop, C. M. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York, United States: Springer. ISBN: 978-0-387-31073-2.

[11] Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen [The German Federal Ministry for Housing, Urban Development and Building] (2023). *Gesetz für die Wärmeplanung und zur Dekarbonisierung der Wärmenetze [Law for heat planning and decarbonization of heat networks]*. Germany. Enacted on 17 November 2023, effective from 1 January 2024. URL: https://www.bmwsb.bund.de/SharedDocs/gesetzgebungsverfahren/DE/kommunale-waermeplanung.html.

[12] Chaverot, M., Carré, M., Jourlin, M., Bensrhair, A., and Grisel, R. (2023). "Improvement of small objects detection in thermal images". In: *Integrated Computer-Aided Engineering* 30(4), pp. 311–325. DOI: 10.3233/ICA-230715.

[13] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation*. DOI: 10.48550/ARXIV.1706.05587. arXiv: 1706.05587.

[14] Chen, L., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). "Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation". In: *European Conference on Computer Vision (ECCV)*. Ed. by Ferrari, V., Hebert, M., Sminchisescu, C., and Weiss, Y. Vol. 11211. Munich, Germany: Springer, pp. 833–851. DOI: 10.1007/978-3-030-01234-2_49.

[15] Clouet, A. (2023). "The thermal imaging and sensing market since 2019". In: *Thermosense: Thermal Infrared Applications XLV*. Ed. by Avdelidis, N. P. Orlando, United States: SPIE, p. 34. DOI: 10.1117/12.2664172.

[16] Da Silva, I. and Segantine, P. C. L. (2025). *Geomatics Applied to Civil Engineering*. 2nd ed. Cham, Switzerland: Springer. ISBN: 978-3-031-75736-5. DOI: 10.1007/978-3-031-75737-2.

[17] Dempster, A. P. (1968). "A Generalization of Bayesian Inference". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 30(2), pp. 205–247. URL: https://www.jstor.org/stable/2984504.

[18] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale". In: *International Conference on Learning Representations (ICLR)*. Austria (virtual). URL: https://openreview.net/forum?id=YicbFdNTTy.

[19] Duda, L., Alibabaei, K., Vollmer, E., Klug, L., Kozlov, V., Berberi, L., Benz, M., Volk, R., Gutiérrez Hermosillo Muriedas, J. P., Götz, M., Sáinz-Pardo Díaz, J., López García, Á., Schultmann, F., and Streit, A. (2025). "Exploring Federated Learning for Thermal Urban Feature Segmentation - A Comparison of Centralized and Decentralized Approaches". In: *Computational Science and Its Applications (ICCSA)*. Ed. by Gervasi,

O., Murgante, B., Garau, C., Karaca, Y., Taniar, D., C. Rocha, A. M. A., and Apduhan, B. O. Vol. 15648. Istanbul, Turkey: Springer, pp. 285–302. DOI: 10.1007/978-3-031-97000-9_18.

[20] Euroheat & Power (2024). *DHC Market Outlook*. Report. Brussels, Belgium: Euroheat & Power. URL: https://api.euroheat.org/uploads/DHC_Market_Outlook_Insights_Trends_2023_81498577a7.pdf.

[21] Fahlstrom, P. and Gleason, T. (2012). *Introduction to UAV Systems*. 4th ed. Aerospace series. Chichester, United Kingdom: Wiley. ISBN: 978-1-119-97866-4.

[22] Friman, O., Follo, P., Ahlberg, J., and Sjokvist, S. (2014). "Methods for Large-Scale Monitoring of District Heating Systems Using Airborne Thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 52(8), pp. 5175–5182. DOI: 10.1109/TGRS.2013.2287238.

[23] Fuchs, H. and Frommhold, W. (1991). "Leckortung auf Fernwärmeleitungen [Leak localisation in district heating pipes]". In: *Fraunhofer IBP Mitteilungen [Fraunhofer IBP Communications]*. Neue Forschungsergebnisse, kurz gefasst [New research findings, in brief] 18(201). URL: https://www.ibp.fraunhofer.de/content/dam/ibp/ibp-neu/de/dokumente/ibpmitteilungen/1-400/201-300/201_IBPmitteilung.pdf.

[24] Gade, R. and Moeslund, T. B. (2014). "Thermal Cameras and Applications: A Survey". In: *Machine Vision and Applications* 25(1), pp. 245–262. DOI: 10.1007/s00138-013-0570-5.

[25] Garrard, C. (2016). *Geoprocessing With Python*. Shelter Island, United States: Manning Publications. ISBN: 978-1-61729-214-9.

[26] Gipiškis, R., Tsai, C.-W., and Kurasova, O. (2024). "Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey". In: *ICT Express* 10(6), pp. 1331–1354. DOI: 10.1016/j.icte.2024.09.008.

[27] Gonzalez, R. C. and Woods, R. E. (2017). *Digital image processing*. 4th ed. New York, United States: Pearson. ISBN: 978-0-13-335672-4.

[28] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. Cambridge, United States: MIT Press. ISBN: 978-0-262-03561-3. URL: https://www.deeplearningbook.org/.

[29] Guo, G., Liu, S., Jia, D., Wang, S., and Wu, X. (2021). "Simulation of a leak's growth process in water distribution systems based on growth functions". In: *AQUA - Water Infrastructure, Ecosystems and Society* 70(4), pp. 521–536. DOI: 10.2166/aqua.2021.021.

[30] Gurklienė, R., Hogland, W., Knutsson, H., Lukosevičius, V., Lundström, J., Ohlsson, M., Rogala, A., Rybarczyk, P., and Zajaczkowski, K. (2023). *BSAM data driven proactive maintenance handbook: Smart maintenance of district heating networks*. Kalmar, Sweden: Linnaeus University. ISBN: 978-91-8082-036-3.

[31] Hamilton, S. and Charalambous, B. (2020). *Leak Detection: Technology and Implementation*. 2nd ed. IWA Publishing. ISBN: 978-1-78906-085-0. DOI: 10.2166/9781789060850.

[32] He, Y., Deng, B., Wang, H., and Cheng, L. (2021). "Infrared machine vision and infrared thermography with deep learning: A review". In: *Infrared Physics & Technology* 116, p. 103754. DOI: 10.1016/j.infrared.2021.103754.

[33] Heating, W. C. Q. D. (2020). *Handbook on Planning of District Heating Networks*. Handbook. Version 1.0. Translation of version 1.2 (German, 2018). Bern, Switzerland: Swiss Federal Office of Energy. URL: https://www.verenum.ch/Dokumente/Handbook-DH_V1.0a.pdf.

[34] Heipke, C. and Tödter, J. (2020). *Drohnengestützte Thermografie als Basis der Asset- und Instandhaltungsstrategie von Fern- und Nahwärmenetzen [Drohne-based Thermography as an Asset- and Maintenancestrategy for District Heating Systems]*. Schlussbericht [Final report] IGF-Vorhaben Nr. 19768 N. Fernwärme-Forschungsinstitut in Hannover e.V. [District Heating Research Institute in Hannover E.V.] and Leibniz Universität Hannover [Leibniz University Hannover]. URL: https://www.fernwaerme.de/pdfdata/Schlussbericht_IGF_19768N.pdf.

[35] Herrmann, C., Ruf, M., and Beyerer, J. (2018). "CNN-based thermal infrared person detection by domain adaptation". In: *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*. Ed. by Dudzik, M. C. and Ricklin, J. C. Vol. 10643. Orlando, United States: SPIE, p. 8. DOI: 10.1117/12.2304400.

[36] Hlebnikov, A., Volkova, A., Dzuba, O., Poobus, A., and Kask, Ü. (2010). "Damages of the Tallinn District Heating Networks and Indicative Parameters for an Estimation of the Networks General Condition". In: *Environmental and Climate Technologies* 5(-1), pp. 49–55. DOI: 10.2478/v10145-010-0034-3.

[37] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022). "Explainable AI Methods - A Brief Overview". In: *xxAI - Beyond Explainable AI: International Workshop*. Ed. by Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W. Cham, Switzerland: Springer, pp. 13–38. DOI: 10.1007/978-3-031-04083-2_2.

[38] Hossain, K., Villebro, F., and Forchhammer, S. (2019). "Leakage Detection in District Heating Systems Using UAV IR Images: Comparing Convolutional Neural Network and ML Classifiers". In: *27th European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain: European Association for Signal Processing (EURASIP). DOI: 10.23919/EUSIPCO45326.2019.

[39] Hossain, K., Villebro, F., and Forchhammer, S. (2020). "UAV Image Analysis for Leakage Detection in District Heating Systems using Machine Learning". In: *Pattern Recognition Letters* 140, pp. 158–164. DOI: 10.1016/j.patrec.2020.05.024.

[40] Iglovikov, V., Mushinskiy, S., and Osin, V. (2017). *Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition*. DOI: 10.48550/arXiv.1706.06169. arXiv: 1706.06169.

[41] International Energy Agency (IEA) (2020). *World Energy Investment 2020*. Report. Paris, France: IEA. URL: https://www.iea.org/reports/world-energy-investment-2020.

[42] International Energy Agency (IEA) (2023). *World Energy Outlook 2023*. Technical report. Paris, France: IEA. URL: https://www.iea.org/reports/world-energy-outlook-2023.

[43] Itti, L., Koch, C., and Niebur, E. (1998). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), pp. 1254–1259. DOI: 10.1109/34.730558.

[44] Jayapalan, L., Murray, C., Chen, F., and Xie, Q. (2024). *Oracle® Spatial: Spatial GeoRaster Developer's Guide*. Guide F32256-13. Version 21c. United States: Oracle®. URL: https://docs.oracle.com/en/database/oracle/oracle-database/21/geors/spatial-georaster-developers-guide.pdf.

[45] Jones, H. and Sirault, X. (2014). "Scaling of Thermal Images at Different Spatial Resolution: The Mixed Pixel Problem". In: *Agronomy* 4(3), pp. 380–396. DOI: 10.3390/agronomy4030380.

[46] Kapil, R., Castilla, G., Marvasti-Zadeh, S. M., Goodsman, D., Erbilgin, N., and Ray, N. (2023). "Orthomosaicking Thermal Drone Images of Forests via Simultaneously Acquired RGB Images". In: *Remote Sensing* 15(10), p. 2653. DOI: 10.3390/rs15102653.

[47] Kapur, J., Sahoo, P., and Wong, A. (1985). "A new method for gray-level picture thresholding using the entropy of the histogram". In: *Computer Vision, Graphics, and Image Processing* 29(3), pp. 273–285. DOI: 10.1016/0734-189X(85)90125-2.

[48] Kinser, J. M. (2018). *Image Operators: Image Processing in Python*. 1st ed. Boca Raton, United States: CRC Press. ISBN: 978-1-498796-18-7.

[49] KMR Service GmbH (2021). *Zubehoer [Component parts]*. Technical report 4. Trollhagen, Germany: KMR Service GmbH. URL: https://kmr-fernwaerme.de/wp-content/uploads/4-Zubehoer-_hell-80_.pdf (visited on 11 Sept. 2025).

[50] Konstantin, P. and Konstantin, M. (2024). *Praxisbuch der Fernwärme und Fernkälteversorgung: Systeme, Netzaufbauvarianten, Kraft-Wärme und Kraft-Wärme-Kälte-Kopplung, Kostenstrukturen und Preisbildung [Practical guide to district heating and cooling: systems, network configurations, combined heat and power and combined heat, power and cooling, cost structures and pricing]*. 3rd ed. Berlin/Heidelberg, Germany: Springer. ISBN: 978-3-662-69526-5. DOI: 10.1007/978-3-662-69526-5.

[51] Latif, J., Shakir, M. Z., Edwards, N., Jaszczykowski, M., Ramzan, N., and Edwards, V. (2022). "Review on condition monitoring techniques for water pipelines". In: *Measurement* 193, p. 110895. DOI: 10.1016/j.measurement.2022.110895.

[52] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *IEEE International Conference on Computer Vision (ICCV)*. Montreal, Canada (virtual): IEEE, pp. 9992–10002. DOI: 10.1109/iccv48922.2021.00986.

[53] Ljungberg, S.-A. and Rosengren, M. (1987). "Aerial Thermography - A Tool For Detecting Heat Losses And Defective Insulation In Building Attics And District Heating Networks". In: *Thermosense IX: Thermal Infrared Sensing for Diagnostics and Control*. Vol. 780. Orlando, United States: SPIE, pp. 257–343. DOI: 10.1117/12.940525.

[54] Ljungberg, S.-A. and Rosengren, M. (1988). "Aerial and Mobile Thermography to Assess Damages and Energy Losses from Buildings and District Heating Networks - Operational Advantages and Limitations". In: *XVIth ISPRS Congress, Technical Commission VII: Interpretation of Photographic and Remote Sensing Data*. Ed. by Murai, S. Vol. XXVII, Part B7. Kyoto, Japan: ISPRS, pp. 348–359. URL: https://www.isprs.org/proceedings/XXVII/congress/part7/348_XXVII-part7.pdf.

[55] Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, Unites States: IEEE, pp. 3431–3440. DOI: 10.1109/cvpr.2015.7298965.

[56] Losi, E., Manservigi, L., Spina, P. R., and Venturini, M. (2024). "Data-driven approach for the detection of faults in district heating networks". In: *Sustainable Energy, Grids and Networks* 38, p. 101355. DOI: 10.1016/j.segan.2024.101355.

[57] Luhmann, T., Robson, S., Kyle, S., and Boehm, J. (2014). *Close-range photogrammetry and 3D imaging*. 2nd ed. De Gruyter textbook. Berlin, Germany: De Gruyter. ISBN: 978-3-11-030269-1. DOI: 10.1515/9783110607253.

[58] Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J. E., Hvelplund, F., and Mathiesen, B. V. (2014). "4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems". In: *Energy* 68, pp. 1–11. DOI: 10.1016/j.energy.2014.02.089.

[59] Lv, J., Shen, Q., Lv, M., Li, Y., Shi, L., and Zhang, P. (2023). "Deep learning-based semantic segmentation of remote sensing images: a review". In: *Frontiers in Ecology and Evolution* 11. DOI: 10.3389/fevo.2023.1201125.

[60] Manservigi, L., Bahlawan, H., Losi, E., Morini, M., Spina, P. R., and Venturini, M. (2022). "A diagnostic approach for fault detection and identification in district heating networks". In: *Energy* 251, p. 123988. DOI: 10.1016/j.energy.2022.123988.

[61] Mao, D., Wang, P., Fang, Y.-P., and Ni, L. (2024). "Understanding District Heating Networks Vulnerability: A Comprehensive Analytical Approach with Controllability Consideration". In: *Sustainable Cities and Society* 101, p. 105068. DOI: 10.1016/j.scs.2023.105068.

[62] Metadata Working Group (2010). *Guidelines For Handling Image Metadata*. Guideline. United States: Metadata Working Group. URL: https://s3.amazonaws.com/software.tagthatphoto.com/docs/mwg_guidance.pdf (visited on 11 Sept. 2025).

[63] Miccinesi, L., Beni, A., and Pieraccini, M. (2022). "UAS-Borne Radar for Remote Sensing: A Review". In: *Electronics* 11(20), p. 3324. DOI: 10.3390/electronics11203324.

[64] Murtazin, I., Kozhevnikov, M., and Starikov, E. (2021). "Development and application of methods of internal inspection of district heating networks". In: *International Journal of Energy Production and Management* 6(1), pp. 56–70. DOI: 10.2495/EQ-V6-N1-56-70.

[65] Neupane, B., Horanont, T., and Aryal, J. (2021). "Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis". In: *Remote Sensing* 13(4), p. 808. DOI: 10.3390/rs13040808.

[66] Olbrycht, R. and Więcek, B. (2015). "New approach to thermal drift correction in microbolometer thermal cameras". In: *Quantitative InfraRed Thermography Journal* 12(2), pp. 184–195. DOI: 10.1080/17686733.2015.1055675.

[67] Paardekooper, S., Lund, R. S., Mathiesen, B. V., Chang, M., Petersen, U. R., Grundahl, L., David, A., Dahlbæk, J., Kapetanakis, I. A., Lund, H., Bertelsen, N., Hansen, K., Drysdale, D. W., and Persson, U. (2018). *Heat Roadmap Europe 4: Quantifying the Impact of Low-Carbon Heating and Cooling Roadmaps*. Project report. Aalborg, Denmark: Aalborg University. URL: https://vbn.aau.dk/ws/portalfiles/portal/288075507/Heat_Roadmap_Europe_4_Quantifying_the_Impact_of_Low_Carbon_Heating_and_Cooling_Roadmaps..pdf.

[68] Parihar, A. S. and Singh, K. (2018). "A study on Retinex based method for image enhancement". In: *2nd International Conference on Inventive Systems and Control (ICISC)*. Coimbatore, India: IEEE, pp. 619–624. DOI: 10.1109/ICISC.2018.8398874.

[69] Pelski, S., Abbott, R., Annamalai, M., Beauregard, B., Chen, F., Chopra, R., Guo, D., Kotsovolos, S., Lamb, A., Lin, D., Mauro, J., Mavris, S., Mediouni, R., Meeks, J., Owens, D., Oxbury, S., Shepard, S., Stern, S.-A., Stuart, I., Toohey, R., Voss, B., Wang, G., Ward, R., and Yalavarthy, M. (2005). *Oracle® interMedia User's Guide*. Guide B14302-01. Redwood City, United States: Oracle®. URL: https://docs.oracle.com/cd/B19306_01/appdev.102/b14302.pdf (visited on 11 Sept. 2025).

[70] Pérez-Enciso, M. and Zingaretti, L. (2019). "A Guide on Deep Learning for Complex Trait Genomic Prediction". In: *Genes* 10, p. 553. DOI: 10.3390/genes10070553.

[71] Persson, U. and Werner, S. (2011). "Heat distribution and the future competitiveness of district heating". In: *Applied Energy* 88(3), pp. 568–576. DOI: 10.1016/j.apenergy.2010.09.020.

[72] Piano, L., Garcea, F., Gatteschi, V., Lamberti, F., and Morra, L. (2022). "Detecting Drift in Deep Learning: A Methodology Primer". In: *IT Professional* 24(5), pp. 53–60. DOI: 10.1109/MITP.2022.3191318.

[73] Pix4D (2020). *Pix4Dmapper*. Software. URL: https://www.pix4d.com/product/pix4dmapper-photogrammetry-software (visited on 28 Aug. 2025).

[74] Rafati, A. and Shaker, H. R. (2024). "Predictive maintenance of district heating networks: A comprehensive review of methods and challenges". In: *Thermal Science and Engineering Progress* 53, p. 102722. DOI: 10.1016/j.tsep.2024.102722.

[75] Rebala, G., Ravi, A., and Churiwala, S. (2019). *An Introduction to Machine Learning*. Cham, Switzerland: Springer. ISBN: 978-3-030-15728-9. DOI: 10.1007/978-3-030-15729-6.

[76] Ren, P., Xiao, Y., Chang, X., Huang, P.-Y., Li, Z., Gupta, B. B., Chen, X., and Wang, X. (2021). "A Survey of Deep Active Learning". In: *ACM Computing Surveys* 54(9). DOI: 10.1145/3472291.

[77] Rogalski, A., Martyniuk, P., and Kopytko, M. (2016). "Challenges of small-pixel infrared detectors: a review". In: *Reports on Progress in Physics* 79(4), p. 046501. DOI: 10.1088/0034-4885/79/4/046501.

[78] Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Ed. by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. Munich, Germany: Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[79] Ruck, J., Vollmer, E., Volk, R., and Vogl, M. (2024). *Detecting District Heating Leaks in Thermal Imagery: Comparison of Anomaly Detection Methods - Source Code and Datasets*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.11085776.

[80] Ruck, J., Vollmer, E., Volk, R., and Vogl, M. (2025). *Thermal Anomaly Segmentation Dataset - Thermal UAS-based Images from Germany with Annotations for Semantic Segmentation Model Training*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.14287864.

[81] Santos, T., Cunha, T., Dias, A., Moreira, A. P., and Almeida, J. (2024). "UAV Visual and Thermographic Power Line Detection Using Deep Learning". In: *Sensors* 24(17), p. 5678. DOI: 10.3390/s24175678.

[82] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* 128(2), pp. 336–359. DOI: 10.1007/s11263-019-01228-7.

[83] Sezgin, M. and Sankur, B. (2004). "Survey over image thresholding techniques and quantitative performance evaluation". In: *Journal of Electronic Imaging* 13(1), p. 146. DOI: 10.1117/1.1631315.

[84] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press. ISBN: 978-0691100425. DOI: 10.2307/j.ctv10vm1qb.1.

[85] Shan, X., Wang, P., and Lu, W. (2017). "The reliability and availability evaluation of repairable district heating networks under changeable external conditions". In: *Applied Energy* 203, pp. 686–695. DOI: 10.1016/j.apenergy.2017.06.081.

[86] Sledz, A. and Heipke, C. (2021). "Thermal Anomaly Detection Based on Saliency Analysis from Multimodal Imaging Sources". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-1-2021, pp. 55–64. DOI: 10.5194/isprs-annals-V-1-2021-55-2021.

[87] Sledz, A., Unger, J., and Heipke, C. (2018). "Thermal IR Imaging: Image Quality and Orthophoto Generation". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLII-1, pp. 413–420. DOI: 10.5194/isprs-archives-XLII-1-413-2018.

[88] Sledz, A., Unger, J., and Heipke, C. (2020). "UAV-based Thermal Anomaly Detection for Distributed Heating Networks". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B1-2020, pp. 499–505. DOI: 10.5194/isprs-archives-XLIII-B1-2020-499-2020.

[89] Sokolova, M. and Lapalme, G. (2009). "A systematic analysis of performance measures for classification tasks". In: *Information Processing & Management* 45(4), pp. 427–437. DOI: 10.1016/j.ipm.2009.03.002.

[90] Song, K., Zhao, Y., Huang, L., Yan, Y., and Meng, Q. (2023). "RGB-T image analysis technology and application: A survey". In: *Engineering Applications of Artificial Intelligence* 120, p. 105919. DOI: 10.1016/j.engappai.2023.105919.

[91] SZ DJI Technology Co. Ltd. (2018). *Matrice 600 Pro: User Manual*. Product information. Version 1.0. SZ DJI Technology Co. Ltd. URL: https://www.dji.com/de/matrice600-pro (visited on 11 Sept. 2025).

[92] SZ DJI Technology Co. Ltd. (2018). *Zenmuse XT 2: User Manual*. Product information. Version 1.0. SZ DJI Technology Co. Ltd. URL: https://www.dji.com/zenmuse-xt2 (visited on 11 Sept. 2025).

[93] Szeliski, R. (2011). *Computer Vision: Algorithms and Applications*. Texts in Computer Science. London, United Kingdom: Springer. ISBN: 978-1-84882-934-3. DOI: 10.1007/978-1-84882-935-0.

[94] Tanda, G. and Migliazzi, M. (2024). "Infrared thermography monitoring of solar photovoltaic systems: A comparison between UAV and aircraft remote sensing platforms". In: *Thermal Science and Engineering Progress* 48, p. 102379. DOI: 10.1016/j.tsep.2023.102379.

[95] Tereshchenko, T. and Nord, N. (2016). "Importance of Increased Knowledge on Reliability of District Heating Pipes". In: *Procedia Engineering* 146, pp. 415–423. DOI: 10.1016/j.proeng.2016.06.423.

[96] Torralba, A., Isola, P., and Freeman, W. T. (2024). *Foundations of Computer Vision*. Ed. by Bach, F. Adaptive Computation and Machine Learning series. Cambridge, United States: MIT Press. ISBN: 978-0-262-04897-2.

[97] Treier, S., Herrera, J. M., Hund, A., Kirchgessner, N., Aasen, H., Walter, A., and Roth, L. (2024). "Improving drone-based uncalibrated estimates of wheat canopy temperature in plot experiments by accounting for confounding factors in a multi-view analysis". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 218, pp. 721–741. DOI: 10.1016/j.isprsjprs.2024.09.015.

[98] Treier, S., Roth, L., Hund, A., Aasen, H., Levy Häner, L., Vuille-dit-Bille, N., Walter, A., and Herrera, J. M. (2025). "Analysis of variance and its sources in UAV-based multi-view thermal imaging of wheat plots". In: *Plant Phenomics* 7(2), p. 100046. DOI: 10.1016/j.plaphe.2025.100046.

[99] Tuikka, L. (2024). "Leak Detection Systems in Finnish District Heating Network". Bachelor thesis. Tampere, Finland: Tampere University of Applied Sciences. URL: https://www.theseus.fi/bitstream/handle/10024/872261/Tuikka_Lijun.pdf.

[100] United Nations (UN) (2015). *Paris Agreement*. Treaty FCCC/CP/2015/L.9/Rev.1. Adopted at the 21st Conference of the Parties (COP 21). Registered as Treaty No. XXVII-7-d in the UN Treaty Collection. Paris, France: UN Framework Convention on Climate Change (UNFCCC). URL: https://unfccc.int/sites/default/files/english_paris_agreement.pdf.

[101] United Nations (UN) (2015). *Transforming our world: the 2030 Agenda for Sustainable Development*. Resolution 70/1. New York, United States: UN. URL: https://sdgs.un.org/2030agenda.

[102] United Nations Environment Programme (UNEP) (2024). *Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector*. Tech. rep. Nairobi, Kenya: UNEP, Global Alliance for Building and Construction (GlobalABC). DOI: 10.59117/20.500.11822/45095.

[103] Usamentiaga, R. (2024). "Semiautonomous Pipeline Inspection Using Infrared Thermography and Unmanned Aerial Vehicles". In: *IEEE Transactions on Industrial Informatics* 20(2), pp. 2540–2550. DOI: 10.1109/TII.2023.3295409.

[104] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. Vol. 30. Long Beach, United States: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[105] Verband Fernwärme Schweiz [Association for District Heating in Switzerland] (2022). *Leitfaden Fernwärme / Fernkälte [Guide to district heating / cooling]*. Schlussbericht [Final report]. Version 1.3. Bern, Switzerland: Bundesamt für Energie [Swiss Federal Office of Energy]. URL: https://www.thermische-netze.ch/fileadmin/user_upload/Dokumente/Publikationen/Downloads/Leitfaden_Fernwaerme_Fernkaelte_03-2022.pdf.

[106] Vinogradova, K., Dibrov, A., and Myers, G. (2020). "Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34(10), pp. 13943–13944. DOI: 10.1609/aaai.v34i10.7244.

[107] Vollmer, E., Benz, M., Kahn, J., Klug, L., Volk, R., Schultmann, F., and Götz, M. (2025). "Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18, pp. 6206–6216. DOI: 10.1109/JSTARS.2025.3537330.

[108] Vollmer, E., König, S., Horstmann, V., Klug, L., Kahn, J., Volk, R., and Vogl, M. (2025). *Thermal Urban Feature Segmentation - Multispectral (RGB + Thermal) UAS-based images from Germany with annotations*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.10814413.

[109] Vollmer, E., Ruck, J., Volk, R., and Schultmann, F. (2024). "Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods". In: *Automation in Construction* 168, p. 105709. DOI: 10.1016/j.autcon.2024.105709.

[110] Vollmer, E., Ruck, J., Volk, R., and Schultmann, F. (2025). "Leak detection using thermal imagery: Deep learning versus traditional computer vision state-of-the-art". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 228, pp. 505–518. DOI: 10.1016/j.isprsjprs.2025.06.006.

[111] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[112] Vollmer, E., Volk, R., and Vogl, M. (2023). *Automatic analysis of UAS-based thermal images to detect leakages in district heating systems: Source code and exemplary dataset*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.7851726.

[113] Vollmer, M. and Möllmann, K.-P. (2018). *Infrared thermal imaging: fundamentals, research and applications*. 2nd ed. Weinheim, Germany: Wiley. ISBN: 978-3-527-69332-0.

[114] Wang, Z., Zhou, J., Ma, J., Wang, Y., Liu, S., Ding, L., Tang, W., Pakezhamu, N., and Meng, L. (2023). "Removing temperature drift and temporal variation in thermal infrared images of a UAV uncooled thermal infrared imager". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 203, pp. 392–411. DOI: 10.1016/j.isprsjprs.2023.08.011.

[115] WebODM Authors (2020). *WebODM*. Software. URL: https://www.opendronemap.org/webodm/ (visited on 28 Aug. 2025).

[116] Whitehead, K. and Hugenholtz, C. H. (2014). "Remote sensing of the environment with small unmanned aircraft systems (UASs), part 1: A review of progress and challenges". In: *Journal of Unmanned Vehicle Systems* 02(03), pp. 69–85. DOI: 10.1139/juvs-2014-0006.

[117] Wojdyga, K. and Chorzelski, M. (2017). "Chances for Polish district heating systems". In: *Energy Procedia* 116, pp. 106–118. DOI: 10.1016/j.egypro.2017.05.059.

[118] Woods, P. (2023). *An Introduction to District Heating and Cooling: Low carbon energy for buildings*. IOP Publishing. ISBN: 978-0-7503-5286-4. DOI: 10.1088/978-0-7503-5286-4.

[119] Xiang, H. and Tian, L. (2011). "Method for automatic georeferencing aerial remote sensing (RS) images from an unmanned aerial vehicle (UAV) platform". In: *Biosystems Engineering* 108(2), pp. 104–113. DOI: 10.1016/j.biosystemseng.2010.11.003.

[120] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Vancouver, Canada (virtual): Curran Associates, Inc., pp. 12077–12090. URL: https://proceedings.neurips.cc/paper_files/paper/2021/file/64f1f27bf1b4ec22924fd0acb550c235-Paper.pdf.

[121] Xiong, H., Cai, W., and Liu, Q. (2021). "MCNet: Multi-level Correction Network for thermal image semantic segmentation of nighttime driving scene". In: *Infrared Physics & Technology* 113, p. 103628. DOI: 10.1016/j.infrared.2020.103628.

[122] Xu, Y., Wang, X., Zhong, Y., and Zhang, L. (2016). "Thermal anomaly detection based on saliency computation for district heating system". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China: IEEE, pp. 681–684. DOI: 10.1109/IGARSS.2016.7729171.

[123] Yates, H. R. (2023). "Challenges and limitations of thermal infrared remote sensing with unoccupied aerial systems". Master thesis. Bozeman, Montana: Montana State University. URL: https://scholarworks.montana.edu/bitstreams/629c8673-852b-457e-b51a-ceb854e1ad64/download.

[124] Yu, L., Guo, Y., Zhu, H., Luo, M., Han, P., and Ji, X. (2020). "Low-Cost Microbolometer Type Infrared Detectors". In: *Micromachines* 11(9), p. 800. DOI: 10.3390/mi11090800.

[125] Yuan, W. and Hua, W. (2022). "A Case Study of Vignetting Nonuniformity in UAV-Based Uncooled Thermal Cameras". In: *Drones* 6(12), p. 394. DOI: 10.3390/drones6120394.

[126] El-Zahab, S. and Zayed, T. (2019). "Leak detection in water distribution networks: an introductory overview". In: *Smart Water* 4(1), p. 5. DOI: 10.1186/s40713-019-0017-x.

[127] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid Scene Parsing Network". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Unites States: IEEE, pp. 6230–6239. DOI: 10.1109/cvpr.2017.660.

[128] Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., and Zhang, L. (2019). "Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 151, pp. 207–222. DOI: 10.1016/j.isprsjprs.2019.02.021.

[129] Zhou, S., O'Neill, Z., and O'Neill, C. (2018). "A review of leakage detection methods for district heating networks". In: *Applied Thermal Engineering* 137, pp. 567–574. DOI: 10.1016/j.applthermaleng.2018.04.010.

[130] Žutautaite, I., Augutis, J., Krikštolaitis, R., Dundulis, G., Valinčius, M., and Rimkevičius, S. (2016). "Risk and reliability assessment of the district heating network methodology with case study". In: *Risk, Reliability and Safety: Innovating Theory and Practice*. Ed. by Walls, L., Revie, M., and Bedford, T. Boca Raton, United States: CRC Press, pp. 2578–2585. DOI: 10.1201/9781315374987-391.

Part II

Companion Articles

Overview of Companion Articles

Study A

Vollmer, E., Volk, R., & Schultmann, F. (2023). Automatic analysis of UAS-based thermal images to detect leakages in district heating systems. In: *International Journal of Remote Sensing*, 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586

Study B

Vollmer, E., Ruck, J., Volk, R., & Schultmann, F. (2024). Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods. In: *Automation in Construction*, 168, p. 105709. DOI: 10.1016/j.autcon.2024.105709

Study C

Vollmer, E., Ruck, J., Volk, R., & Schultmann, F. (2025). Leak detection using thermal imagery: Deep learning versus traditional computer vision state-of-the-art. In: *ISPRS Journal of Photogrammetry and Remote Sensing*, 228, pp. 505–518. DOI: 10.1016/j.isprsjprs.2025.06.006

Study D

Vollmer, E., Benz, M., Kahn, J., Klug, L., Volk, R., Schultmann, F. & Götz, M. (2025). Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering. In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 18, pp. 6206–6216. DOI: 10.1109/JSTARS.2025.3537330

Study E

Vollmer, E., Volk. R., & Schultmann, F. (2025). Assessing the Economic Viability of Thermography-based Leak Detection for District Heating Systems. Submitted to a scientific journal.

A Automatic Analysis of UAS-based Thermal Images to Detect Leakages in District Heating Systems

Abstract

The mostly subterranean nature of district heating system pipelines makes pinpointing any occurring leakages a challenge. Airborne thermography offers a means for widespread monitoring, allowing thermal anomalies to be identified within multitudes of infrared images. This paper details a program developed to automate the entire image analysis process using mainly open-source software. Thermal images are acquired via unmanned aircraft system (UAS), pre-processed, and georeferenced individually or combined to orthomosaics. The search space is minimized to areas around the pipelines. Regions of interest are determined by image segmentation via tailored triangle histogram thresholding. The majority of resulting false alarms are removed by comparison with characteristic traits and results classified by their severity. The algorithm is applied to images newly acquired in Germany as part of a case study. The implemented methodology allows for a reduction of between 92 and 99% of thermal anomalies to a manageable amount of potential leakages for network operators to view. The use of orthomosaiking software in this context, though helpful in coalescing data, is found to lack robustness, precision and therefore reliability. Despite some limitations, the developed program is able to confidently detect and categorize leakages of varying severity and can be used directly by network operators. Future research will focus on further data pre-processing to eliminate thermal drift and remove the remaining false alarms, which mostly pertain to common urban features.

Abbreviations

CRS coordinate reference system

DHS district heating system

DTM digital terrain model

GDAL Geospatial Data Abstraction Library

GIS geographic information system

GNSS global navigation satellite system

GSD ground sampling distance

IMU inertial measurement unit

LOD1 level of detail 1

LoG Laplacian of Gaussian

LWIR long-wavelength infrared

ODM OpenDroneMap

OSM OpenStreetMap

RJPEG Radiometric Joint Photographic Experts Group

SWM Stadtwerke München [Munich's municipal utilities company]

TIR thermal infrared

UA unmanned aircraft

UAS unmanned aircraft system

A.1 Introduction

District heating systems (DHSs) are networks of underground pipelines, which distribute thermal energy in the form of hot water or steam to end users [19, 34]. They are capable of incorporating new energy sources and therefore play an important role in the integration of renewables in the heat sector [19, 34]. Compared to local heat generators, DHSs are eco-friendlier, highly energy and resource efficient, provide means for easy pollution control, and are accompanied by a reduced administrative effort [9, 19].

However, after decades of use, such subterranean pipelines fatigue. Leakages develop owing to corrosion, the deterioration of insulation material, or motion in the ground [23]. If left unattended, these leakages can turn into public safety hazards, eroding the surrounding soil and causing the ground to collapse [9]. The results include damages of considerable cost to infrastructure, the environment, involved personnel, and system efficiency [38]. Any loss of energy or fluid with such a potentially catastrophic impact needs to be mitigated.

Effective monitoring and maintenance are key to ensuring these systems stay efficient and safe. The subterranean nature of the networks, however, complicates the matter. In- and outlet flow sensors can generally identify system changes, but pinpointing exact leakage locations becomes a difficult and cost-intensive undertaking when the pipes need to be dug free for inspection [9, 38]. Therefore, various alternatives for large-scale network monitoring have emerged, including the use of remote sensing and infrared thermography.

This method is based on the premise that a pipe leakage causes an increase in temperature at the surface above, which may then be detected as an anomaly within thermal images [3]. However, manually screening the vast number of images resulting from such an aerial thermography process would be extremely laborious and time-consuming for network operators [9]. As a result, some researchers have already presented methods for automated analyses of thermal images and leakage detection.

A.1.1 Related work

After Axelsson [3] and Ljungberg and Rosengren [21] suggest using airborne thermography for large-scale monitoring of district heating systems, Friman et al. [9] are the first to describe a method to automatically analyse images and detect pipeline leakages. They acquire thermal infrared (TIR) images via manned flight and georectify them using recorded GPS and inertial measurement unit (IMU) measurements. The search area is minimized by masking the images with a network blueprint and additional buffer in geographic information system (GIS) format. The anomaly detection problem is solved by identifying those pixels which exceed a certain threshold of the pixel intensity probability distribution. False alarms caused by buildings are removed through segmentation via watershed transform. Berg et al. [5] expand on this methodology by extracting and evaluating distinguishing image features and using machine learning to classify these previously found detections as true or false. Building segmentation is improved upon by implementing a building mask created using OpenStreetMap (OSM).

Xu et al. [36] implement saliency computation models based on colour, orientation, and intensity features on TIRs acquired during manned flights. Georeferencing the images allows the search space to be minimized to areas close to the DHS. This is achieved with a network blueprint in GIS format manually buffered using the commercial software ArcGIS. Saliency maps of the masked TIR images are created with a simplified Itti model and leakages detected by threshold segmentation. Zhong et al. [39] expand on this method by fusing a local and global saliency map to include a comparison to local pixel neighbourhoods as well as overall feature rarity. The authors use maximum entropy segmentation to identify leakages in the final saliency map. Thermal images are acquired during both manned and unmanned aircraft system (UAS) flights. When pipeline GIS data is unavailable, the authors assume the DHS to be located underneath roads and segment streets from RGB images as an alternative.

Sledz et al. [28] manually create orthomosaics[1] from UAS-based images using the software Agisoft Metashape and mask these with buffered DHS GIS data. Anomalies are detected as elliptical thermal hot spots by means of an Laplacian of Gaussian (LoG) blob detector. Blobs found by the detector are divided into regions and the temperature difference from warmest region to surrounding area is thresholded to identify which depict a leakage.

[1] Orthomosaics are georectified mosaics, meaning mosaics that are correctly mapped and have had their inherent distortions removed [1].

Hossain et al. [17] demonstrate a machine learning approach on images acquired via UAS to find potential leakages. A region extraction algorithm is used for pre-procesing, which detects warm regions and large gradients in an image, then combines both to an output image. A human operator identifies and labels the thus identified regions as leakages or not using ground truth data and expert knowledge. Some images are subsequently used to train a deep learning convolutional neural network and several different conventional machine learning classifiers.

A.1.2 Aim and scope

This paper showcases novel implementations of various methodology to the end of identifying leakages in district heating networks. The problem at hand is solved while simultaneously focusing on automation and generalisation of the thermal image analysis procedure. The work thereby features a number of contributions to the fields of remote sensing, automation and image analysis.

While the above mentioned papers present different approaches, few elaborate on the extent of achieved automation or any actual software created for a widespread application. Oftentimes, third-party software is mentioned to have manually been used for steps like georeferencing or adding buffers. Some steps are not detailed at all and the means of creating and cost of running any mentioned or built software is unknown. Neither source code nor any thermal image data sets are provided, thus prohibiting tests, comparisons, or improvements. All these factors substantially impede the use or recreation of these methods and make the adoption of the proposed methodology near impossible.

The first contribution of this work therefore lies in the detailed description of a fully automated, implementable and replicable program for DHS leakage detection. In contrast to related work, the source code will be provided with the publication of this paper and is available on Zenodo [33]. The focus of the program design lay on using mostly open-source software to allow as large a user-ship as possible. This new form of implementation requires novel adaptations as well as the introduction of new methods.

Furthermore, this paper showcases various new methodological implementations. Following the example of only the most recent papers in this field, the thermal images are acquired via UAS. This is preferable to manned aircraft because flight planning is much more flexible, cheaper, and the process yields images with a higher ground sampling distance (GSD) [28]. However, the close proximity to the ground complicates the georeferencing process because buildings have a more prominent distortive effect. The authors of the previously discussed work do not detail the means by which they realized georeferencing or did not automate the process at all, and they do not address the problem of said distortions.

Another key contribution of this work therefore lies in the incorporation of three different approaches of handling the georeferencing problem automatically, varying in complexity and precision. The first is an individual image georeferencing method by means of affine transformation matrix using global navigation satellite system (GNSS) and IMU data and its

adaption to remove general distortions by including building and topography data. Additionally, a novel indicator for false attitude recordings is identified and according automatic sorting out developed. This allows for new insights to be gathered concerning best practice rules for image acquisition via UAS. The other two implemented georeferencing methods demonstrate and compare the utilisation of open- and closed-source photogrammetry software for automated orthomosaic generation. Additionally, because photogrammetry software itself can only handle input images of high quality, another contribution of this work is the development of an automatic thermal image quality indicator using the Canny edge detector. Finally, the novel use of these three different georeferencing approaches in parallel allows for a first direct comparison between these methods in their automatically implemented form.

Further contributions of this work revolve around the leakage identification itself. Previously proposed detection methods in the afore-described papers either failed to yield satisfactory results on the given real-world data[2] or proved to be inapplicable owing to a temperature drift in the given case study images. Therefore, the focus of this work lay in identifying a generally applicable method independent of data drifts and without requiring extensive preliminary effort. This new method for image segmentation implements the triangle histogram algorithm for thresholding as well as essential adaptions for its use in this new application. False positives are sorted out based on a novel combination of three characteristic leakage traits, as separately described in aforementioned work. Additionally, a simple means of result classification by severity via relative temperature difference is introduced. Lastly, the results are automatically prepared for user-friendly reviewing.

The automation of these aforementioned steps in their entirety as a focus of this work limits the required user interaction to the beginning (for data input) and the end (for result review). The program and method functionality are demonstrated and discussed using newly acquired images for a case study in Germany, which were used as a basis for design and testing. Contrary to predecessors and in the spirit of open data exchange, a dataset of thermal images will be made publicly available and can be viewed on Zenodo [33].

A.1.3 Outline

The rest of this paper is organized as follows. Section A.2 discusses the required data for the program to function. This includes details on the thermal image acquisition process, required prevalent conditions, and necessary auxiliary data. Section A.3 describes the five key methodological steps implemented for leakage detection within the framework of the developed program. Section A.4 details its application on images of a new case study, while Section A.5 discusses the results and limitations. Section A.6 concludes the paper with a summary and outlook.

[2] Preliminary testing showed saliency maps and LoG blob detectors to be ineffective or require the setting of multitudes of dataset-dependent parameters

A.2 Thermal image acquisition and auxiliary program data

Required data for program usage are thermal images acquired via UAS, GIS data of the DHS, and surface information, which include a level of detail 1 (LOD1) model of buildings in the area and a digital terrain model (DTM).

A.2.1 Thermal images and their acquisition

Image acquisition via UAS is preferable as it enables a flexibility, cost and ground sampling distance unobtainable by manned aircraft [28]. Depending on the size of a DHS, it can take several days and the subdivision of the network into multiple regions to cover the entire area [9]. Certain conditions must prevail during all flights to ensure sufficient image quality. These can be summed up as follows:

- Image acquisition should take place at night, preferably a few hours before dawn so as to ensure most irrelevant heat sources caused by sun and people have been eliminated [9, 16].
- Outdoor air temperatures should not exceed 10°C. In the northern hemisphere, the acceptable time frame is given as October to April. [16]
- The ground should be free of foliage or snow to give the thermal camera an unhindered view of the ground above the DHS. Not only may leakages otherwise be overlooked, but also snow reflectance can falsify the thermal imagers' temperature recording [9, 16].
- It should not be snowing, raining, or foggy during image acquisition and wind speeds should be comparatively low.[3] Water particles absorb some of the IR radiation before it can reach the thermal camera, resulting in grainy images lacking contrast and detail. Long-lasting rain and high winds cause the ground surface to cool down and puddles to form, which falsify the recorded temperature data. [9, 16]

Several characteristics of the DHS in question need to be considered as well. The pipes cannot be placed too deep below the surface for thermography to work reliably. Heipke and Tödter [16] set the limit at 1 m and stress that the depth, just like the temperature of the transported medium, should stay constant throughout the system. Additionally, the method may not function reliably on unconventional types of pipe casings, meaning these must be tested in advance.[4] Lastly, local laws pertaining to UAS usage on private or company property might call for special permits or prevent data acquisition in specific areas. [16]

[3] Heipke and Tödter [16] define 2 m/s as the maximum wind speed yet exceeded that limit frequently during their own image acquisition.

[4] Heipke and Tödter [16] name the following as unproblematic: fibre cement, plastic, protective, steel, and flexible pipes, independent of overhead or underground placement.

The recorded images must contain extractable thermal data and specific metadata such as a GNSS coordinate, flight height, camera parameters, and attitude information in the form of yaw, pitch, and roll angles of flight and gimbal. A fitting file format is therefore Radiometric Joint Photographic Experts Group (RJPEG) as these images contain thermal matrices with a temperature value for every recorded pixel.

A.2.2 Auxiliary data

Friman et al. [9] already recognized the advantage of reducing the search space within the images to areas close to the DHS. It allows distant objects, which otherwise would cause false alarms,[5] to be removed from consideration before leakage detection has even begun. This same step is therefore also implemented within the developed program and requires a blueprint of the pipeline network in GIS format, specifically a shapefile. These are obtainable from the local authorities or network operators [9].

This step can only be performed if the images themselves are georeferenced in advance. Georeferencing refers to the conversion of the image coordinate system to a chosen geographic coordinate system, thus correctly placing an image on a map [18]. Images acquired during UAS flights suffer from high amounts of relief displacement [35], which is the distortion and positional error induced by objects of different heights in the camera's field of view [14]. To mitigate this somewhat, building and topography data are taken into account. A LOD1 model of surrounding buildings should be provided in shapefile format, topographic information in the form of a DTM of points and corresponding ground elevation values in text file format.

A.3 Methodology

Using a similar methodology to Friman et al. [9]'s, the leakage detection process is divided into five main steps within the program.

1. preliminary image processing
2. image georeferencing and/or orthomosaic generation
3. search space minimisation to areas above and around the DHS
4. thermal anomaly detection via image segmentation
5. leakage identification by reduction of false alarms

[5] For instance faulty building insulation, streetlamps, people, vehicles, and metal surfaces

The procedure is visualized as a flowchart in Figure A.1. The acquired thermal images are analysed in sets, which are groups of images collected in one flight. Pre-processing occurs both in preparation for the subsequent step as well as for later leakage identification. The images are placed on a map or alternatively combined to generate orthomosaics using third-party software. By overlaying these with a DHS pipeline blueprint the search space can be minimized to areas above and around the network, thus excluding those of no importance and ensuring considerably fewer false alarms ensue. Areas of interest within the resulting masked images are defined as those pixels that exceed a specific threshold. These thermal anomalies are evaluated by size, shape, and temperature to identify those which are true leakages and classify the results into three categories of severity. The leakage positions are additionally evaluated.

The program is designed to run on a ground workstation with at minimum 8 GB of RAM, an Intel i5 processor, and either a Windows 10 Pro or Linux Ubuntu 22.04 operating system. For the purpose of automation, the open-source programming language Python is utilized in the versions 3.6.7 and 3.9.1. Python has several advantages for its use in this application, such as its widespread availability, large support network, and variety of thirdparty modules in addition to extensive support libraries. Other programs are also easily accessible via Python, such as the aerial image mapping tools OpenDroneMap (ODM) [24] and Pix4D [26] that are employed in this case (Section A.3.2). The code can be viewed in full on Zenodo [33].

It is run via command line and utilizes user prompts to attain all relevant inputs, analysing as many sets of images as desired. The final results are saved into excel files and different diagrams depict the stages of the process. Users therefore need only interact with the program at the beginning of a run-through by providing all required data. The results can then be viewed at the end to determine which identified leakages merit closer observation or even direct intervention.

A.3.1 Image pre-processing

Preliminary image processing encompasses the calculation of the raw thermal data, the definition of a set's temperature limits in preparation for subsequent steps, as well as a quality assessment of the entire set. The temperature data is imperative for later leakage identification and can only be obtained from the original images.

For later data conversions, a credible temperature range for each set of images is determined. The maximum and minimum of all extracted values cannot represent this range because thermal cameras can incorrectly record a temperature given adverse surrounding conditions or specific material types.[6] Relative minimum and maximum temperature limits

[6] Metal objects, for instance, are depicted with much colder temperatures than are actually measured on their surface. [10]

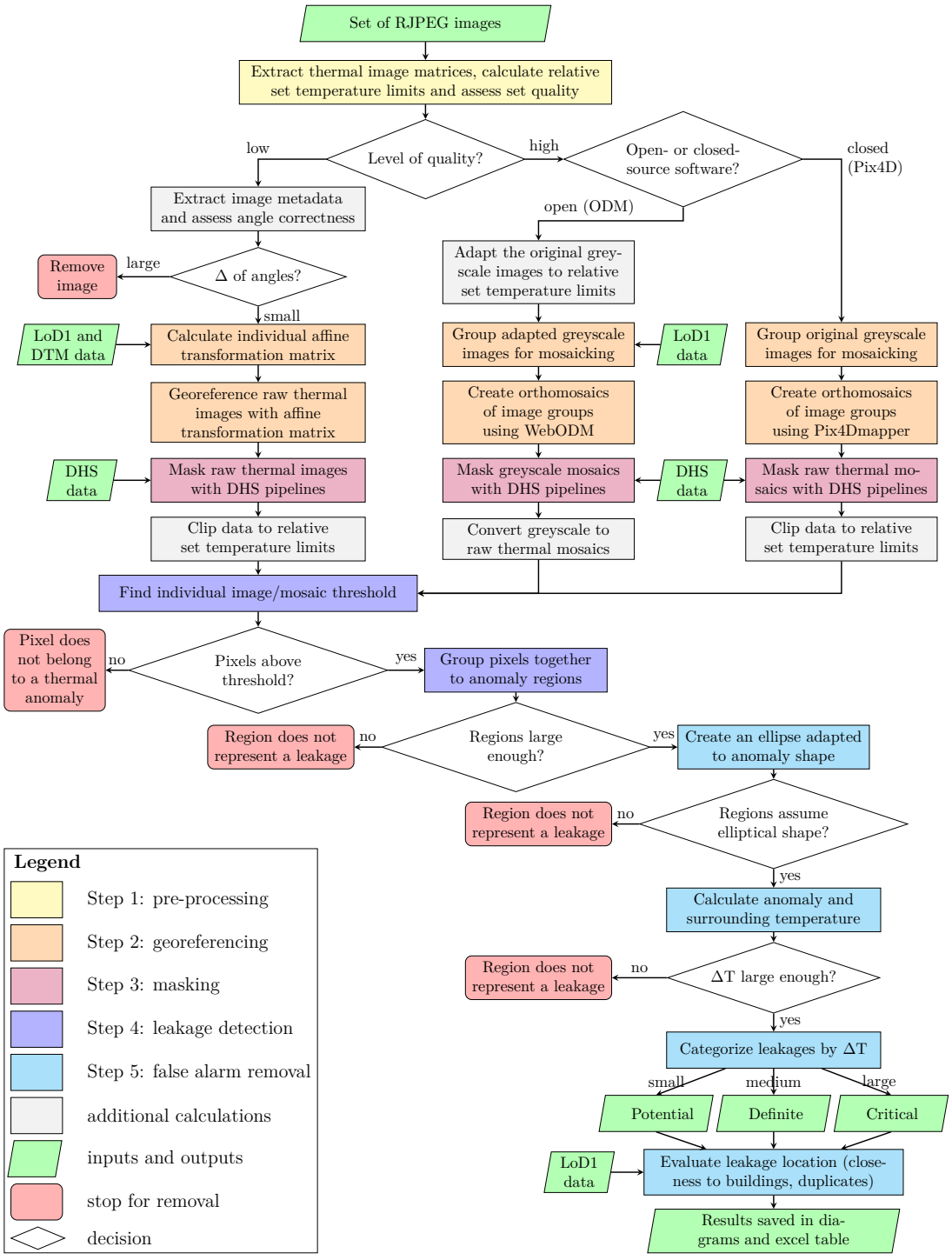


Figure A.1: Leakage detection procedure implemented in the developed Python program.

of a set of images are instead defined using the arithmetic mean μ and standard deviation σ of all temperature values:

$$T_{\text{set}_{\text{min}_{\text{rel}}}} = \mu_{\text{set}} - 3\sigma_{\text{set}} \tag{A.1}$$

$$T_{\text{set}_{\text{max}_{\text{rel}}}} = \mu_{\text{set}} + 9\sigma_{\text{set}} \tag{A.2}$$

In a normal distribution, the area between $\mu - 3\sigma$ and $\mu + 3\sigma$ encompasses 99.7% of all data [22]. Although a histogram of typical set values does not directly equate such a distribution, they are akin in shape. Therefore, these values are chosen as initial limits. The upper limit is extended and defined by triple the distance to the set's mean temperature value because the usage of this data for leakage detection purposes inherently requires a focus shift to higher temperature ranges. Leakages may display a more than 20°C higher temperature in comparison to their surroundings [9] and thus extend well beyond such an average set temperature.

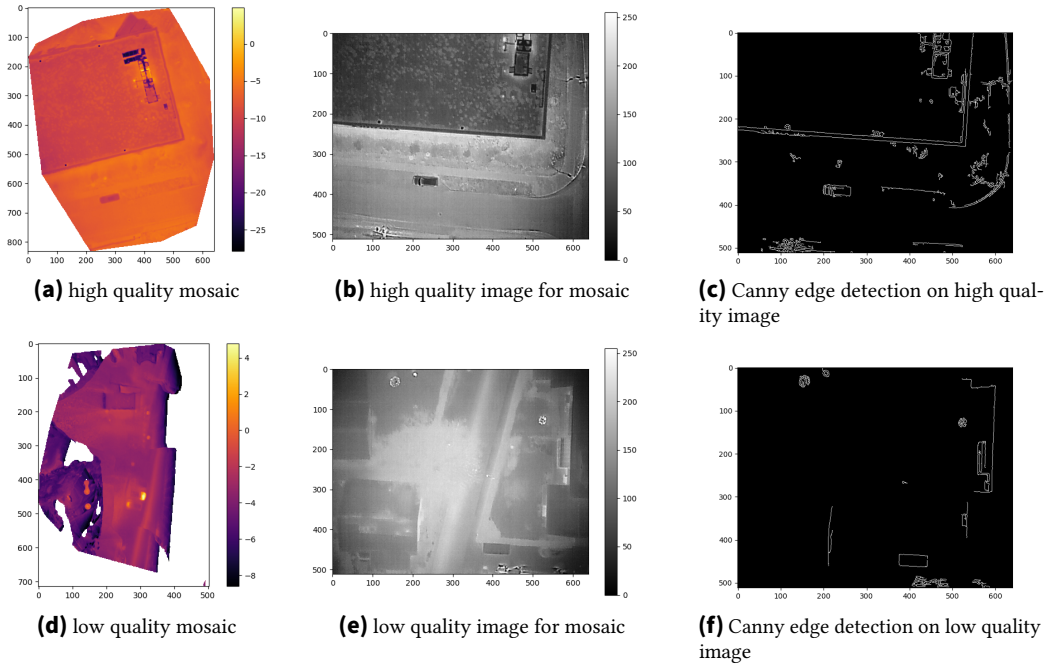
Due to the vast amount of data in a set, the values of all temperature matrices cannot be handled simultaneously. Instead, the mean and standard deviation are computed for each thermal image individually and then combined. The combined arithmetic mean of all temperature matrices is simply the average of all single image means. To calculate the combined standard deviation, Burton [6]'s code using analysis of variance is applied. Owing to an extensive image overlap, not all image pixels are considered equally. However, the mathematical error is minor and the assumption of using all temperature data still valid for determining a general range of relevant values.

Lastly, the general image quality of a set is evaluated because it defines the way in which georeferencing can be implemented. An analysis of both photogrammetry software shows that low quality images do not yield usable results (Figure A.2). Low quality hereby refers to having few distinguishable features, in other words being particularly grainy or depicting large amounts of vegetation. Resulting mosaics are blurry and contain large gaps of data missing (Figure A.2d).

The quality of a set is determined using the Canny edge detector, an algorithm that can optimally balance precise edge localisation and noise influence [7, 31]. It returns a binary image displaying the dominant edges as white pixels, the pixel count of which ultimately defines set quality. The algorithm is applied to every image (Figure A.2c, A.2f) to determine a set's mean edge pixel count. Each set is then classified as being of either high or low quality via thresholding. The optimal threshold for this case study was found to be 4000 pixels.

A.3.2 Image georeferencing and orthomosaic creation

In order to minimize the search space to areas of relevance, the images are georeferenced first by positioning them on a map and thus linking internal coordinate system to geographic counterpart [20]. The step is crucial to overall accuracy, as only precisely



(a) high quality mosaic **(b)** high quality image for mosaic **(c)** Canny edge detection on high quality image

(d) low quality mosaic **(e)** low quality image for mosaic **(f)** Canny edge detection on low quality image

Figure A.2: Comparison of mosaics generated from high and low quality images. Quality is defined by the existence of distinguishable features or lack thereof and visualized using the Canny edge detector algorithm applied to the images.

positioned images allow for exact masking results. The difficulty, as mentioned in Section A.2.2, lies in mitigating the relief displacement and geometric distortions inherent in remote sensing applications [14]. Three methods are implemented within the program: (I) georeferencing of individual images, (II) using aerial image mapping software WebODM [24] and (III) Pix4Dmapper [26]. Both applications create mosaics of several images at a time while simultaneously removing the inherent distortions in a process called geo- or orthorectification [1, 31]. Since the mapping software requires high image quality for successful processing, the first method comes into play when the images have a low quality.

The low quality images are georeferenced individually by calculating an affine transformation matrix for each [11]. Applying this matrix will shift, scale, and rotate the image raster data to its correct position within a chosen coordinate reference system (CRS) [2]. It can be calculated with a set of geographical coordinates, the GSD, and the drone heading angle [11]. For this, the following parameters are extracted from the image metadata: focal length f, sensor width s_w, image width i_w and height i_h, flight height h, longitude lon and latitude lat of the image centre point, and yaw or so-called heading angles κ of both unmanned aircraft (UA) and camera. The intrinsic camera parameters shown in Figure A.3

(left side) are used to calculate the size of the image's footprint g_w and g_h on the earth's surface using the following equations [25]:

$$GSD = \frac{s_w \times h \times 100}{f \times i_w} \quad \left[\frac{mm}{pxl}\right] \tag{A.3}$$

$$g_w = \frac{GSD \times i_w}{100} \quad [m] \quad \text{and} \quad g_h = \frac{GSD \times i_h}{100} \quad [m] \tag{A.4}$$

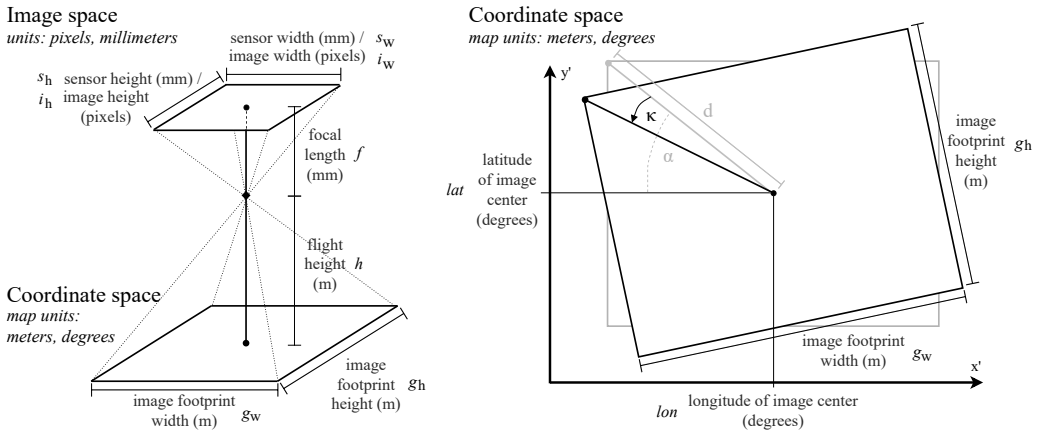


Figure A.3: Visualisation of the image parameters required for georeferencing. Left side: parameters relevant for calculating the GSD and image footprint. Right side: combination with the UA heading angle to calculate ground coordinates of the image corners. [25]

The coordinates of each image corner are calculated by combining an image's footprint with its centre point coordinate's easting and northing values[7] and heading angle κ[8] as depicted in Figure A.3 (right side). The following trigonometric equations [22] exemplify the procedure for the upper left corner:

$$d = \sqrt{\left(\frac{g_w}{2}\right)^2 + \left(\frac{g_h}{2}\right)^2} \qquad \text{and} \qquad \alpha = \arctan\left(\frac{g_h}{g_w}\right) \tag{A.5}$$

$$x'_{UL} = Easting_{UL} = Easting_{center} - d \times \cos(\alpha - \kappa) \tag{A.6}$$

$$y'_{UL} = Northing_{UL} = Northing_{center} + d \times \sin(\alpha - \kappa) \tag{A.7}$$

By calculating the coordinates of the other three image corners in analogous fashion, Python's Geospatial Data Abstraction Library (GDAL) [12] is capable of constructing the affine transformation matrix.

[7] Latitude and longitude are translated from the geographical coordinate system (unit: degrees, minutes and seconds) into easting and northing values in a Cartesian CRS (unit: meters).

[8] Ordinarily, complex angle transformations would be necessary to convert the recorded navigational attitude yaw, pitch, and roll to the photogrammetric equivalents' omega, phi, and kappa used in georeferencing [4]. These can be omitted here because the implemented 3-axis gimbal [30] stabilizes the UA, ensuring negligibly small roll and pitch angles (Prof. Grant Petty, personal communication, April 14, 2021). Owing to nadir UA alignment, the yaw angle is directly assumed to be κ.

Some adjustments are necessary for the application of this method on images recorded via UAS. Relief displacement caused by the small flight height is prominent in images recorded above buildings and mitigated by subtracting the building height from the flight height when the centre coordinate lies above a building. A LOD1 model of the buildings is used for this purpose. Additionally, ground elevation information from a given DTM is incorporated, as terrain following was not part of the flight program.

A fail-safe is included to account for incorrect flight attitude, as the recorded angles are not always accurate (Prof. Grant Petty, personal communication, April 14, 2021). As both the UA and camera save a yaw angle per image, inaccurate measurements can be identified by a large delta between the two.[9] In these cases, the images are removed from consideration because the calculated corner points lead to incorrectly placed images.

Alternatively, orthomosaics are generated from high quality thermal images by either or both WebODM [24] and Pix4Dmapper [26]. These are chosen owing to their capacity for automation and advantages (Table A.1).

Table A.1: Overview and comparison of photogrammetry software WebODM and Pix4Dmapper (information from OpenDroneMap [24], Pix4D [26], Pix4D [27], and Groos et al. [15])

	WebODM	**Pix4Dmapper**
Relevant inputs	JPEG images	JPEG, RJPEG images
	including all original metadata	
Useful outputs	orthomosaic	orthomosaic, thermal index map
Installation effort	**high** - prerequisites: Docker, Git, if Windows: 10 Pro OS	**medium** - prerequisites: Pix4Dengine, Python version 3.6.7
Python access	**directly**	**indirectly** - via Pix4Dengine
Cost	**none**	**high** - licenses for Pix4Dmapper (1500€ for research purposes) and -engine

The images are divided into groups (by size and proximity to surrounding buildings) prior to processing. For WebODM, an error handling mechanism is included to increase the general groups of between 15 and 30 images to a greater size if no usable mosaic is generated from the original group. The groups for Pix4Dmapper are generally larger at around 30 images as those result in the highest quality mosaics. It should be noted that image groups cannot be extended endlessly as the resulting mosaics will display a too low resolution and quality for effective leakage detection [9], which is why the fail-safes are limited in their extension capabilities.

WebODM requires greyscale equivalents of the raw thermal images as inputs, which are calculated using the relative set temperature limit defined during pre-processing (Section A.3.1). This is necessary because, unlike Pix4Dmapper which can generate

[9] An 8° difference is experimentally determined as the threshold for the given UA Matrice 600 Pro [29].

thermal index maps, WebODM is incapable of creating orthomosaics from the thermal RJPEGs themselves.

A.3.3 Minimising the search space

Georeferenced images and orthomosaics are overlaid with the network blueprint in GIS format to remove pixels that are too far away from the DHS. Overlaying image raster data with network pipeline vector data is achieved with Python's 'Rasterio' [13], a library also used to extend the DHS shapefile by a 3.5 m buffer on all sides. Figure A.4 exemplifies the described process using the georeferencing results of the previous step. The WebODM mosaic data are finally converted back to thermal values using the same relative set temperature limits and all other image and mosaic data are clipped to the same range.

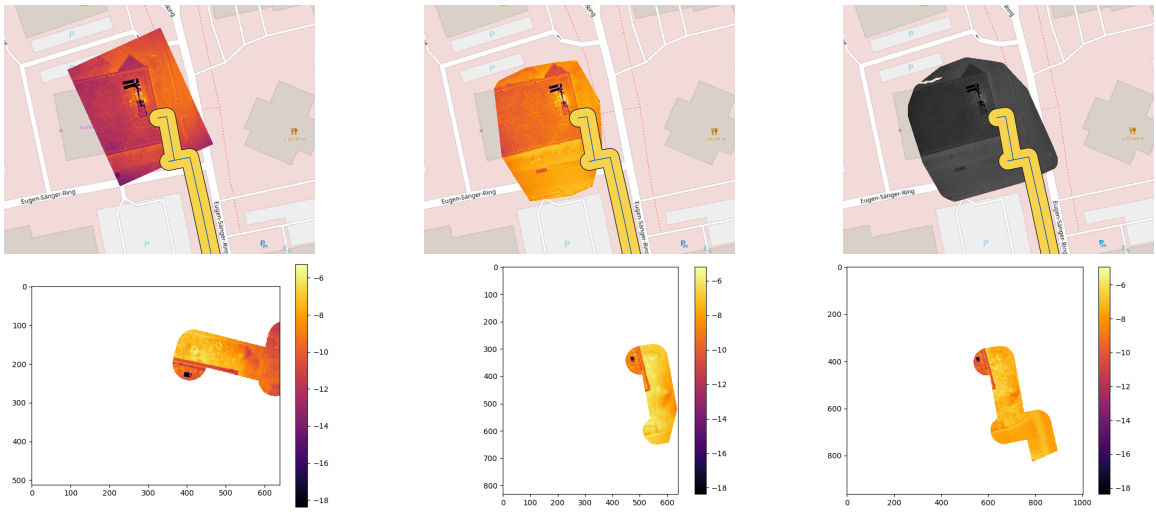


Figure A.4: Comparison of masking and post-processing procedures of individual image georeferencing (first), Pix4Dmapper (second) and WebODM (third column).

A.3.4 Detection of thermal anomalies

The masked images or mosaics are segmented to find pixels of interest and thus thermal anomalies. Similarly to Friman et al. [9], pixels of interest are hereby assumed to reside in the upper tail of a histogram of all values. Instead of choosing the threshold as an upper percentile at random, the triangle threshold histogram algorithm is implemented to identify a fitting limit. Owing to a temperature calibration drift within the case study data, the method cannot be used to identify an overall set threshold and is instead repeated for each individual image. Figure A.5 exemplifies this non-linear drift in a dataset of case study images, showing an increase of temperatures without cause.

The algorithm uses the hypotenuse that links the histogram peak P_1 to the right-side edge P_2 of the set temperature range (Figure A.6). The longest orthogonal connection

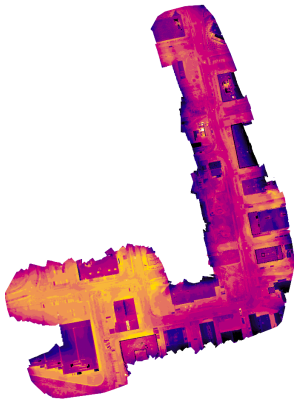


Figure A.5: Temperature drift in a case study set, with acquisition from top right to bottom left. Warm areas do not stem from underground heat (SWM).

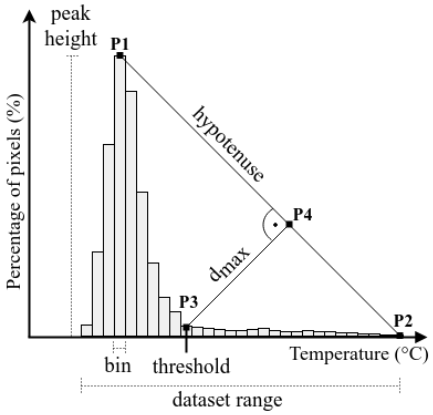


Figure A.6: Visualisation of the triangle histogram thresholding method and involved parameters [37].

d_{max} between hypotenuse and histogram bin is determined iteratively and identifies the threshold as the corresponding bin's temperature value. [37]

Here, an adjustment is made to account for histograms consisting of multiple distinctive peaks when large uniform areas of different temperatures are depicted. This can happen when a building and street take up near equal areas within an image. The peak is then not simply defined as the bin of maximum height, but instead as the warmest amongst all distinctive local maxima. Only if no other peaks of considerable size are detected is the bin with the maximum percentage selected.

Experimental analyses showed that the threshold may at times be too conservative, choosing too many pixels as being of interest. This can happen in very varied images, which include a broad range of temperatures occurring multiple times. For this reason, the pixel percentage of the chosen threshold bin is taken into account. If the percentage exceeds 1%, the threshold is moved further to the right to a bin that does not exceed said experimentally determined limit. The Appendix A.I exemplifies these described amendments to the traditional triangle histogram thresholding method, thus demonstrating the reasons for how it is tailored to match the specific requirements of thermal leakage detection in the urban context.

The described method results in an array for each image consisting of pixels labelled according to their relation to the found threshold. The pixels of interest are grouped to regions of interest. This is accomplished by means of Python's 'scikit-image' [32], a library that calculates various parameters for pixels with identical labels. First, the labelled arrays are assessed to determine whether more than one region of interest exists by thresholding the so-called extent (ratio of labelled pixels to all within their bounding box). If the pixels are all in the same area, they are said to comprise a single region, while sparse and widely spread out pixels are recognized as multiple clusters. The clusters are identified using the flood fill algorithm. Since the algorithm only connects direct neighbours, small regions are additionally joined with larger ones if they are in close proximity to one another.

A.3.5 False alarm removal

Even after search space minimisation and segmentation, many false alarms can be found amongst the detected thermal anomalies. Therefore, only those anomalies that satisfy specific size, form, and temperature constraints are classified as leakages.

1. Size: A visual analysis of the thermal images and correspondence with local municipal utilities companies show that very small thermal anomalies[10] cannot represent actual pipeline leakages and are therefore sorted out.
2. Form: The form constraint is defined by elliptical or circular shape.[11] The thermal anomaly should, to a certain degree, exhibit such a form to be considered a leakage. The anomaly's centre pixel coordinate, major and minor axis lengths, and rotational angle are calculated to define the ellipse's size, shape and orientation [22]. The ratio of anomaly pixels to the amount constituting the ellipse is thresholded with an experimentally found limit to determine which thermal hotspots can be interpreted as leakages.
3. Temperature: True leakages are identified in a comparison of anomaly temperature to its surroundings.[12] A ring is constructed at a distance around an anomaly to calculate the average temperature of the surrounding area. The anomaly temperature is defined by the mean of all pixels in small hotspots, the warmest 50 in medium-sized, and the top 100 pixels in large anomalies. If the temperature difference exceeds 5°C[13] the anomaly is declared a leakage.

The identified leakages are categorized as either potential, definite, or critical depending on the severity of said temperature difference. Potential leakages have a temperature delta between 5°C and 10°C, definite ones display a difference of 10°C to 15°C, while critical leakages are defined as exceeding the 15°C mark.

To remove redundancy and give insight into further potential false alarms, a location analysis is preformed. The leakage positions are compared to surrounding buildings as well as each other. This helps determine whether a found region of interest is situated on a building roof (e.g. caused by faulty insulation or a warm chimney) or if any duplicates are present (e.g. the same thermal anomaly is displayed within multiple subsequent images). Additionally, anomalies in close proximity to a building may be caused by vents or entrances [9]. To identify leakages on top of or close to buildings, the coordinate of each leakage centre point is compared to the buildings LOD1 model.

[10] In the case of a UA's flight height, a conservative threshold is chosen at 10 pixels.

[11] This form is based upon Sledz et al. [28]'s observation that an elliptical shape best depicts leakages. Berg et al. [5] consider form as an important factor, but only take circularity into account as a feature for distinguishing leakages. However, as Friman et al. [9] state, a leakage is not always perfectly round and may also be partially covered, making an ellipse the more obvious choice.

[12] Berg et al. [5] propose a similar method of comparing the intensity of a hotspot and its surroundings. However, using temperature data has the advantage of enabling postliminary leakage categorisation with empirical values.

[13] Sledz et al. [28] define a delta of 5°C already sufficient to indicate a leakage.

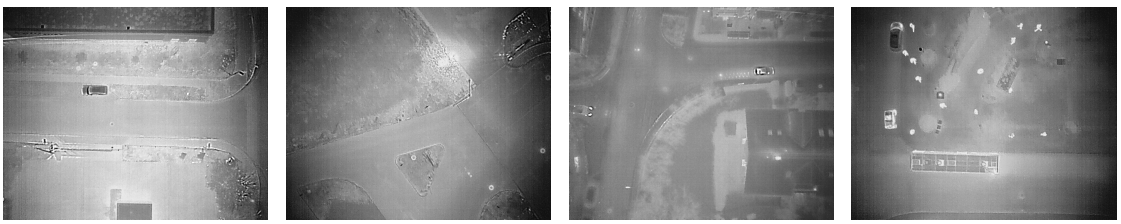
In theory, the procedure works well. However, practical application has shown the calculated geotransform (Section A.3.2) to lack precision, thus prohibiting the inclusion of this positional information in the leakage detection process itself. Appendix A.II exemplifies how in particular the calculated mosaic placement can deviate from its true positioning. Furthermore, true leakages may appear near buildings as pipeline networks are often situated in their close proximity [9]. Therefore, this information is provided as supplementary material to the list of leakages to aid in facilitating the result evaluation.

A.4 Results

A.4.1 Case study description

The case study examines parts of the DHS of suburban areas close to Munich in Germany. It encompasses several regions including the towns Taufkirchen, Ottobrun, and Neubiberg, which in total contain approximately 43 km of DHS pipelines. The water transported by these pipelines ranges from 80°C to 130°C. The installation depth is unknown. The level of urbanization varies throughout the given region, being mostly suburban but including many parks and forests. The area was split into 49 zones, with every flight requiring approximately 25 min. Image acquisition took place at night-time between 7 pm and 6 am in December of 2019. The ambient temperature ranged from 2° to -5°C with maximum wind speeds at 20 km/h. The process produced approximately 55,000 images.

While many of the resulting images have a high quality, there were some disadvantages to the chosen time frame, which lead to low quality, temperature drift or the depiction of features that falsify leakage detection. In some images, snow coverage is distinctly visible while ongoing snowfall during the acquisition of others produced extremely grainy pictures. The comparatively early starting time allowed for a large amount of human activity, which resulted in the presence of people as well as vehicles exuding heat. These scenarios are exemplified in Figure A.7. Such problematic occurrences naturally reduce the efficiency of any leakage detection algorithm owing to either an increase of false alarms or loss of valuable information.



(a) Good weather conditions **(b)** Good weather conditions, leakage visible **(c)** Problematic weather conditions (snowfall) **(d)** Problematic timing (high level of activity)

Figure A.7: Select case study TIR images examples of high and low quality showcasing various aspects of image acquisition.

DJI's Matrice 600 Pro UA was utilized for nearly fully automated image acquisition. The routes for image acquisition were pre-programmed into the control software and only required a human operator for takeoff. The UA's flight height was consistent at 60 m above the starting point. An image overlap of approximately 88% was achieved. The camera technology was provided in form of the Zenmuse XT2, a combination of FLIR's Duo Pro R thermal and visual camera and DJI's gimbal. The thermal camera contains an uncooled VOx microbolometer and therefore works in the long-wavelength infrared (LWIR) range between 7.5 and 13.5 µm. It has a focal length of 13 mm, a sensor width of 10.88 mm, and a spatial resolution of 640×512 pixels. [8, 29, 30]

The thermal camera can output images in JPEG, TIFF or FLIR's own RJPEG file format [8]. Because RJPEG contains extractable thermal data saved to each image pixel, they were chosen as the preferred output. Every image stored all relevant metadata including the camera parameters, GNSS coordinate, relative height, and positional angles. It should be noted that the recorded absolute altitude values were incorrect, as they placed the images below the ground. For this reason, the relative values were utilized in combination with a DTM. Ordinary RGB images (in JPEG format) were recorded as well, although they are not utilized. Both the DTM and LOD1 model were provided by the city of Munich.

A.4.2 Case study results

Four sets out of the 49 (Figure A.8) were evaluated to demonstrate the program's functionality, allude to its effectiveness, and discuss potential for improvement. They were chosen for the presence of confirmed ground truth data and varying image quality (Table A.2).

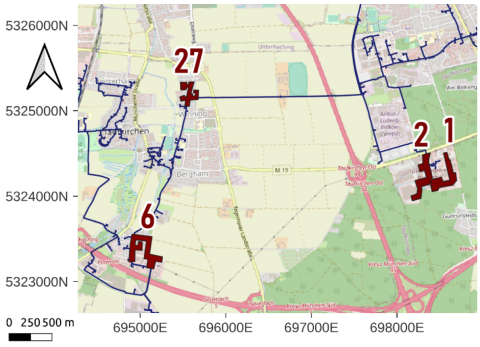


Figure A.8: Locations of four case study sets

set #	1	2	6	27
date	02.12.	03.12.	02.12.	11.12.
time	11.15-11.50 pm	12.05-12.40 am	10.00-10.45 pm	10.30-11.00 pm
image count	875	868	982	640
Canny pixel count	8250	10706	8444	2163
quality	high	high	high	low

Table A.2: General information on the four select case study sets.

Applying the developed leakage detection algorithm to the four select sets lead to the results given in Table A.3. Between 92 and 99% of false alarms were sorted out using the various methods described in Section A.3.5. The remaining false positives were mostly found to stem from common urban features, such as warm vehicles, buildings or manholes.[14]

[14] Munich's municipal utilities company SWM states some manholes may indicate DHS leakages and should therefore not be disregarded completed, although the majority are marked as false alarms in Table A.3.

Individual image evaluation leads to the largest amount of detected leakages, mainly due to the high image overlap causing regions of interest to be identified multiple times. The leakage location evaluation helps determine these cases and remove duplicates. A comparative analysis of leakage groupings shows the chosen methodology to be effective in grouping together true positives, meaning identical thermal anomalies.[15] The false positive rate, in other words the amount of leakages defined as being a replica although they are not, equalled approximately 8% in both mosaics and individual images. The location information also gives insight into false alarms that are caused by buildings. An evaluation of the computed leakage locations showed that thermal anomalies close or above buildings were correctly found in 58% of cases within mosaics and 68% within single images. Of the anomalies determined to be in proximity to buildings, a small percentage[16] was found to stem from objects close by (such as vehicles) instead of the building itself.

This requires different computational effort depending on the method involved. Single image georeferencing takes the least amount of time and depends solely on the amount of images in a set. Orthomosaic generation is more time-intensive owing to the need to access the third party software. WebODM outstrips both alternatives by far because of the included fail-safe mechanism, which reruns faulty image groups.

A.5 Discussion

The results of the developed leakage detection algorithm were compared to a manual assessment of the select case study sets, an overview of which is given in Table A.4. This manual evaluation was performed on a mosaic created from all images within a set, as little actual ground truth data was available. The algorithm returns only a small portion of the manually identified regions of interest, the reasons for which are provided in last row of Table A.4.

In approximately 70% of the cases (denoted by “ΔT") manually defined areas of interest were detected by the algorithm, yet sorted out owing to a low temperature difference to their surroundings. The vast majority of manually identified thermal hotspots therefore did not differ enough from their setting to be classified as leakages. Almost 15% of the visually identified leakages (denoted by “mask") were eliminated from consideration due to their increased distance from the DHS pipelines. The algorithm is therefore found to greatly support human analysis by removing those manually identified regions which do not fulfil key leakage criteria. The other missed leakages shed light upon two shortcomings of the leakage detection algorithm in its current implementation. The first is caused by the automation of the georeferencing and mosaicking step (denoted by ’no mos’ or ’mos wrong’), the second by the thresholding method being applied individually to images and mosaics (denoted by “T_{th}") due to the mentioned calibration drift in the case study sets (Appendix A.II).

[15] In mosaics approximately 85% and in individual images 96% of thermal hotspots
[16] 7% within mosaics and 10% within single images

Table A.3: Overview of the results of the leakage detection algorithm applied to four select sets
Abbreviations: pot = potential, def = definite, crit = critical, bldg = building, mh = manhole, mos = mosaic, veh = vehicle

set #	**1**			**2**			**6**			**27**
method	**Pix4D mapper**	**Web ODM**	**single images**	**Web ODM**	**Web ODM**	**single images**	**Web ODM**	**Web ODM**	**single images**	**single images**
leakages before removal	170	1488	4439	150	2083	5014	137	2275	5922	5402
leakages of fitting size	73	171	727	75	218	541	61	226	942	972
leakages of fitting shape	31	55	376	30	62	200	33	101	495	541
leakages of high ΔT	7	14	117	12	29	90	10	30	141	58
unique leakages	6 pot: 6	9 pot: 6, def: 2, crit: 1	17 pot: 13, def: 3, crit: 1	7 pot: 4, def: 2, crit: 1	12 pot: 7, def: 4, crit: 1	13 pot: 7, def: 2, crit: 4	8 pot: 5, def: 3	20 pot: 17, def: 2, crit: 1	41 pot: 37, def: 3, crit: 1	15 pot: 15
false positives	2x veh, 3x mh, 1x mos	1x veh, 3x mh, 1x bldg, 2x lamp	2x veh, 1x mh, 10x bldg, 2x lamp	1x veh, 1x mh, 1x bldg, 2x mos	3x mh, 5x bldg, 1x metal	3x mh, 6x bldg, 1x metal	2x veh, 2x mh, 4x bldg	8x veh, 1x mh, 8x bldg	9x veh, 2x mh, 23x bldg, 2x metal	5x veh, 6x bldg, 2x lamp
actual leakages	0	pot: 1	pot: 2	pot: 1, crit: 1	pot: 2, crit: 1	pot: 2, crit: 1	0	pot: 2	pot: 5	pot: 1
run time [h]	1.57	4.49	1.25	1.75	4.34	1.26	1.99	3.96	1.48	0.93
$\sum$ run time [h]		**6.08**			**6.02**			**5.91**		**0.93**

Table A.4: Comparison of the leakage detection algorithm results with the manual evaluation of the select case study sets.
Abbreviations: pot - potential, def - definite, crit - critical, mos - mosaic, th - threshold
Reasons for oversight: incorrect manual identification: ΔT - delta too small, mask - too far from DHS
incorrect removal by algorithm: T_{th} - incorrect threshold, mos - mosaic failure

set #	**1**			**2**			**6**			**27**
method	**Pix4D mapper**	**Web ODM**	**single images**	**Pix4D mapper**	**Web ODM**	**single images**	**Pix4D mapper**	**Web ODM**	**single images**	**single images**
unique leakages	pot: 6	pot: 6, def: 2, crit: 1	pot: 13, def: 3, crit: 1	pot: 4, def: 2, crit: 1	pot: 7, def: 4, crit: 1	pot: 7, def: 2, crit: 4	pot: 5, def: 3	pot: 17, def: 2, crit: 1	pot: 37, def: 3, crit: 1	pot: 15
true leakages	0	pot: 1	pot: 2	pot: 1, crit: 1	pot: 2, crit: 1	pot: 2, crit: 1	0	pot: 2	pot: 5	pot: 1
manually identified leakages		pot: 4			pot: 3, crit: 1			pot: 5, def: 2		pot: 5
manually identified and true leakages	0	0	0	pot: 1, crit: 1	pot: 2, crit: 1	pot: 2, crit: 1	0	0	pot: 1	0
reasons manually identified leakages are missed	1x ΔT, 1x mask, 1x T_{th}, 1x no mos	2x ΔT, 1x mask, 1x T_{th}	2x ΔT, 1x mask, 1x T_{th}	1x ΔT, 1x mos wrong	1x ΔT	1x ΔT	4x ΔT, 2x T_{th}, 1x mos wrong	5x ΔT, 1x mask, 1x T_{th}	4x ΔT, 1x mask, 1x T_{th}	4x ΔT, 1x T_{th}

Automating the georeferencing step - be it as orthomosaics or individually - means not all images can be processed. Table A.5 shows the extent of this fact. As discussed in Section A.3.2, some images include incorrect heading angles and must therefore be removed from consideration during individual georeferencing. Owing to the high image overlap of nearly 90%, this only poses a problem when large amounts of consecutive images fail the angle evaluation - an extremely rare situation ("single" columns in Table A.5). Figure A.9 depicts the source of most incorrect angles: they occur when the UA abruptly changes direction. This observation can be used as a newly identified best practice rule and should be taken into account during future image acquisition and flight planning to ensure the usability of all images.

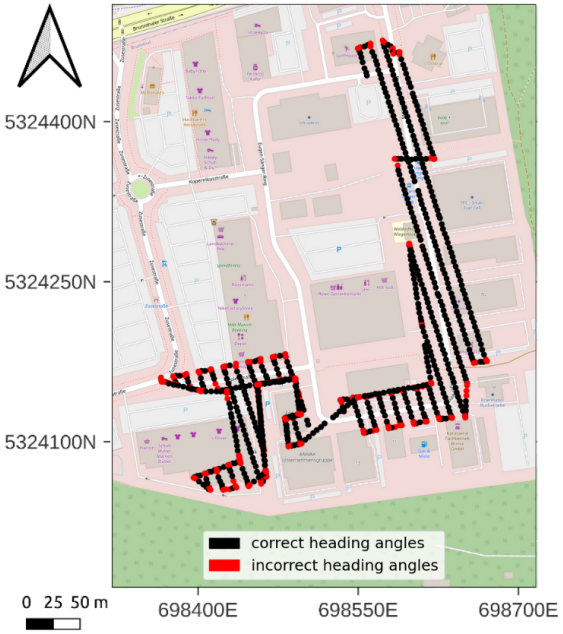


Figure A.9: GNSS coordinates of all images in a case study set displayed on an OSM background. Images with incorrect heading angles are red.

This is more problematic in the automated mosaicking process. Both WebODM and Pix4Dmapper were at times found to fail to generate mosaics, leading to entire image groups being passed over. This happens in WebODM despite implemented quality and failure control mechanisms when an image group still yields no usable results after considerable size extension. In Pix4Dmapper, even extending a previously failed image group to double its amount of images may not yield any result and forces the original group to be skipped. This explains why the automated Pix4mapper software misses most images (Table A.5) and one of the manually identified leakages in Set 1 (Table A.4). While theoretically more adept at removing distortions and image vignetting, evaluation of the case study datasets showed the georeferencing of both WebODM and Pix4Dmapper mosaics to be imprecise (Appendix A.II).

Despite all specified processing options, the mosaics sometimes failed to achieve high quality. In the selected image sets, the WebODM mosaics included holes or small areas

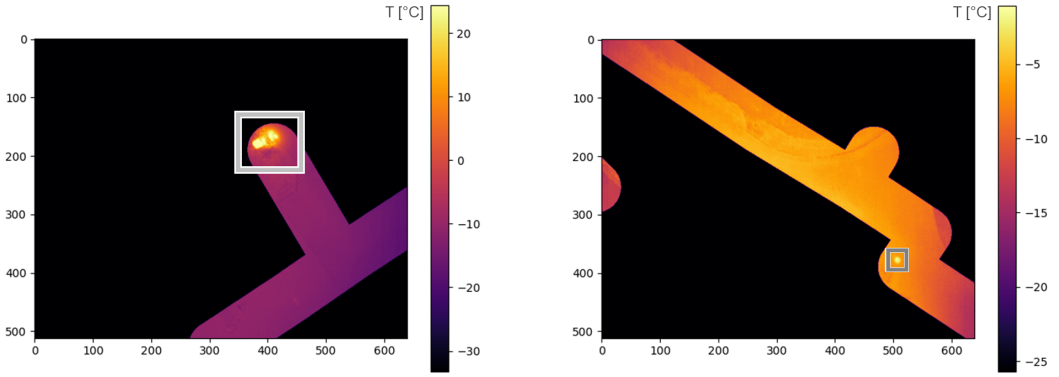
missing in approx. 7% of the cases. Pix4Dmapper displayed the more serious peculiarity of sometimes producing near (approx. 17%) or completely (approx. 12%) unusable mosaics. While a cause for this might be the previously mentioned conflicting image metadata, it highlights the necessity for a more indepth testing, particularly of Pix4Dmapper, to ensure the usability for applications such as leakage detection. Additionally, the importance of high image quality must be stressed: The optimal acquisition conditions described in Section A.2.1 should always be adhered to. The comparative analyses of all implemented georeferencing methodology has shown the single image georeferencing process, while not as adept at removing distortions, to be the most consistent, robust and reliable. Moreover, due to the large number of images and various viewpoints of similar scenery, the scaling distortions do not weigh quite as heavily. It also requires the least processing time (Table A.3).

The second shortcoming of the implemented leakage detection algorithm comes from the necessity to individually threshold images and mosaics due to the observed calibration drift (Section A.3.4). While this individual segmentation method works well in detecting locally concentrated leakages of various sizes, it cannot identify more widespread variants that encompass large areas. Extending the algorithm to perform both individual as well as the combined thresholding demonstrated by Friman et al. [9] could solve this problem and allow for the detection of long pipeline sections of interest. Any temperature drift similar to this case study's would, however, cause a vast increase in false positives. Ultimately, an extensive analysis of the thermal camera is required to ensure such a drift does not occur during future acquisitions. Alternatively, calibration references such as black-bodies with known temperatures can be placed within the flight zone to calculate it (Prof. Grant Petty, personal communication, April 14, 2021).

Owing to the algorithm design and non-automatic removal of duplicates as well as little actual ground truth data available, an exact comparison to methods from related work

Table A.5: Comparison of the select case study sets illustrating the methodology dependant amounts of unevaluated image data. Files refer to either mosaics or thermal images. Different file counts pertain to the originally given or planned number of files versus those resulting after georeferencing. Percentages are calculated in terms of images, although overlapping mosaics are not considered.

set #	**1**			**2**			**6**			**27**
method	Pix 4D	Web ODM	sin-gle	Pix 4D	Web ODM	sin-gle	Pix 4D	Web ODM	sin-gle	sin-gle
original file count	28	75	875	28	82	868	32	109	982	640
resulting file count	22	75	694	26	81	651	30	111	796	448
missing images	21%	0%	21%	7%	0.8%	25%	6%	0.9%	18.9%	30%
relevant missing images	21%	0%	0%	7%	0.8%	0.5%	6%	0.9%	0%	0.2%



(a) Critical leakage (Set 2) with a temperature difference of 36.4°C

(b) Potential leakage (Set 6) with a temperature difference of 5.1°C

Figure A.10: Example leakages detected within the select case study image sets

cannot properly be performed. However, despite the revealed problems, the implemented program is capable of identifying the most critical and important leakage in the entire case study within Set 2, which ruptured two weeks after image acquisition. It was detected within all variations of the georeferencing process, characterized by a 36.4°C temperature difference to its surroundings, and therefore categorized correctly as critical in all cases (Figure A.10A.10a). Various smaller, potentially existent leakages were reliably identified as well (Figure A.10A.10b). Most importantly, the developed algorithm was shown to be capable of automatically evaluating thousands of images, detecting the important leakages and returning concise and easily manageable results completely autonomously.

Finally, some limitations of the program in its current form need to be addressed. For adequate functionality, thermal images must be acquired in a similar fashion as those of the case study. For instance, terrain following should not be implemented because the elevation data used to calculate the flight height would otherwise have an adverse effect and cause imprecision. Changes in flight height or choice of thermal camera affect the correctness of the thermal image quality assessment, buffer size around the DHS network, and false alarm removal by anomaly size and may therefore cause subpar functionality.

Incorporating building data into false alarm identification requires extremely precise georeferencing. Because none of the three presented methods can offer this to the desired extent, the location evaluation step could not be included in the overall process. Further enhancement is necessary to do so.

A.6 Conclusion

A.6.1 Summary

The aim of this study was to develop and demonstrate a simple, automated way for network operators to monitor and assess their DHSs. Aside from making use of the cost- and

time-efficient UAS-based acquisition process, this meant the development of an program suitable for automated thermal image analysis. The designed software is functional, completely automated, and capable of correctly identifying leakages of varying importance. Considerable value was placed upon the development of a freely implementable program to ensure the program's future usability and allow for effortless modifications of both code and general methodology.

The developed algorithm comprises five fundamental steps modified for automation. The images are pre-processed, three variations of georeferencing are applied and compared, results are masked with the DHS blueprint, and thermal anomalies are identified within the minimized search space. Various false positives are sorted out based on characteristic size, shape, and temperature of thermal anomalies, and the final results are categorized as leakages of different severity. Lastly, information is supplied on leakage locations, their proximity to buildings, and potentially existing duplicates.

The results are presented in an easily comprehensible and transparent fashion using excel data sheets and the plotting of leakage locations on maps to give operators an effortless and quick overview. Additionally, the identification of identical leakages and anomalies in proximity to buildings can help reduce the amounts of remaining false alarms.

Result evaluation exposed two drawbacks of the algorithm in its current implementation, with suggestions for improvement discussed in Section A.5. It highlighted the importance of capturing high quality images by adhering to best-practice acquisition rules (Section A.2.1). The camera should be tested to ensure no discernible temperature drift occurs. Additionally, flight speeds or patterns should be altered to record correct heading angles and avoid sharp curves. In order to alleviate the misinterpretation of common urban features such as vehicles and people, it would be prudent to acquire images during the night and early morning.

A.6.2 Outlook

An important extension to a leakage detection program would be the inclusion of temporal analyses as described by Berg et al. [5]. Repeating the image acquisition process after one or more years under similar conditions allows for a comparison to the initial state and gives further insight into the general network status and its degradation. Critical changes can be identified and initially overlooked leakages discovered. [5]

Additional insight can also be gained by integrating simultaneously recorded RGB images as suggested by Sledz et al. [28], which would allow for a parallel analysis of both thermal and optical images. For instance, it would be possible to remove common false alarms in the given case study data sets caused by streetlamps or manholes through their identification within the corresponding optical images. However, since the optimum time frame is image acquisition at night-time, the inherent RGB information that can be extracted from optical images will be limited.

Future work may also profit from the integration of network characteristics like age, repairs, material information, and pipeline depth. Including these can help improve the quality of leakage forecasts. Data gained from collaborations with local utility companies would also allow for more precise analyses by including any identified true ground truth information, something that studies so far lack.

One of the main gaps in current literature is a true comparison of already developed solutions. The lack of published thermal datasets and source code makes this impossible up to now. We hope that by sharing our data as part of this paper's publication, the open-source approach will become common-place amongst these kinds of case studies and thus allow for genuine analyses and a collaborative exchange of knowledge.

Acknowledgements

The thermal images for the case study in Germany were acquired in collaboration with the Air Bavarian GmbH and Munich's municipal utilities company Stadtwerke München.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

CRediT author statement

Elena Vollmer: Conceptualisation, Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Rebekka Volk**: Conceptualisation, Data Curation, Writing - Review & Editing, Supervision. **Frank Schultmann**: Writing - Review & Editing, Supervision. All authors have agreed to the published version of the manuscript.

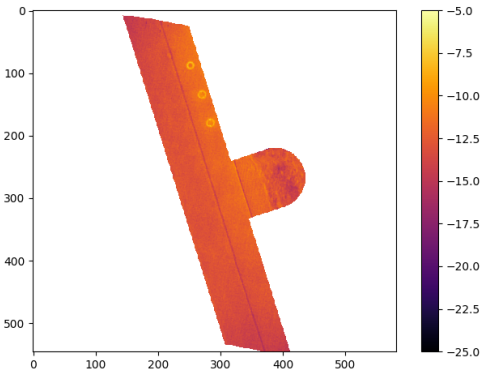
Appendices

Appendix I

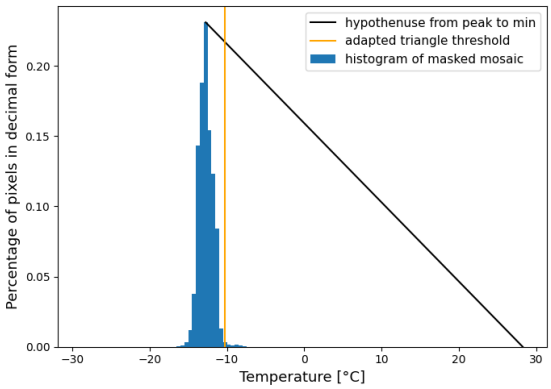
The following figures exemplify the amendments made to the basic triangle histogram thresholding method, showing why these are necessary in form of a cursory, visual ablation study. Each figure depicts in (*a*) a thermal image after mapping and masking, in (*b*) a distribution of the thermal data and threshold automatically selected by the thresholding

method, and in (*c*) the results of applying that threshold to the image to obtain a binary matrix with white pixels being those of interest.

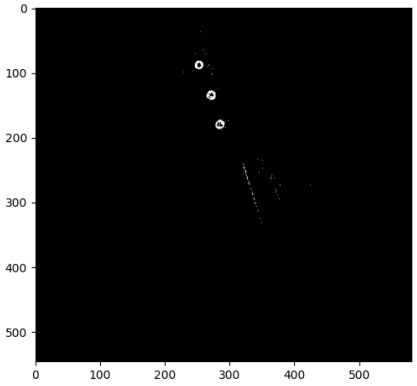
Figure A.11 shows a regular thermal mosaic and its threshold, which is used without further changes to create the resulting binary image in A.11c. No amendments are necessary for the majority of images such as this one.



(a) Standard thermal mosaic (temperature colourbar in °C)



(b) Histogram and triangle threshold

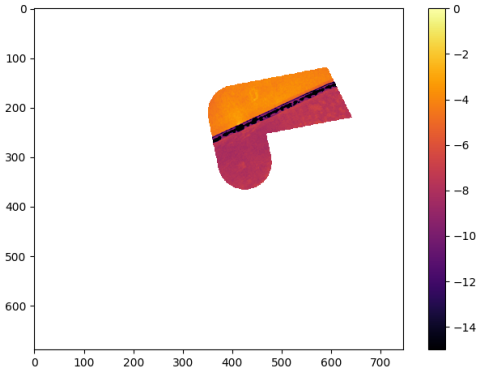


(c) Resulting binary thresholded mosaic

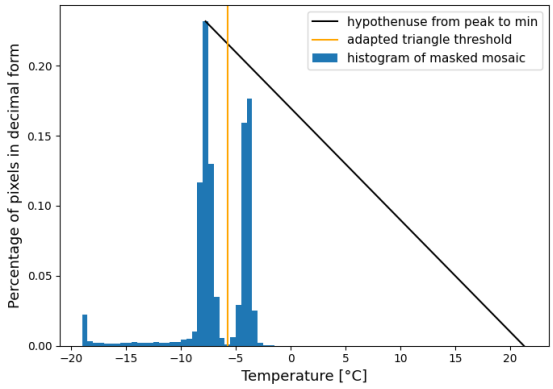
Figure A.11: Triangle histogram thresholding demonstrated using a standard thermal mosaic

Figure A.12 showcases an example where the standard choice of threshold lies between peaks, making an amendment necessary to choose the correct threshold according to peak distribution. Figures A.12d and A.12e display the effects of the methodological change on the choice of threshold and subsequent pixels of interest.

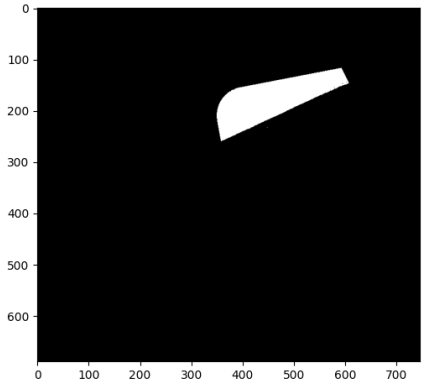
Figure A.13 demonstrates a too conservatively chosen threshold and the implemented correction to mitigate the problem. Figures A.12d and A.12e display the effects of the methodological change on the choice of threshold and subsequent pixels of interest.



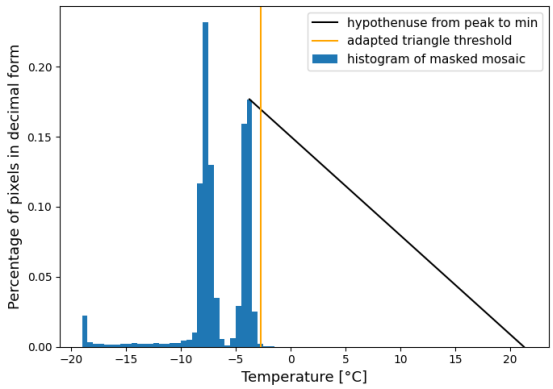
(a) Thermal mosaic with two distinct areas (temperature colourbar in °C)



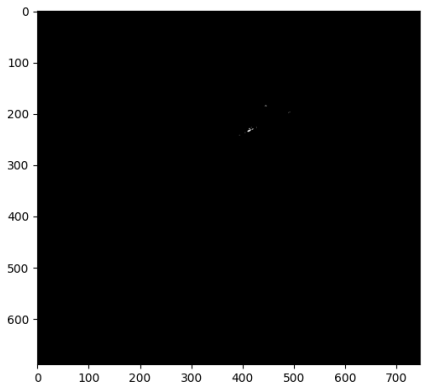
(b) Histogram and triangle threshold of thermal mosaic without peak consideration



(c) Resulting binary thresholded mosaic with too many pixels of interest

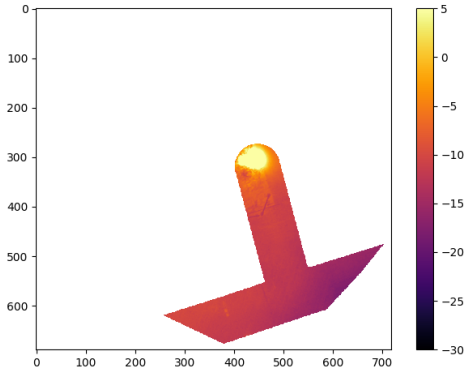


(d) Histogram and triangle threshold of thermal mosaic with peak consideration

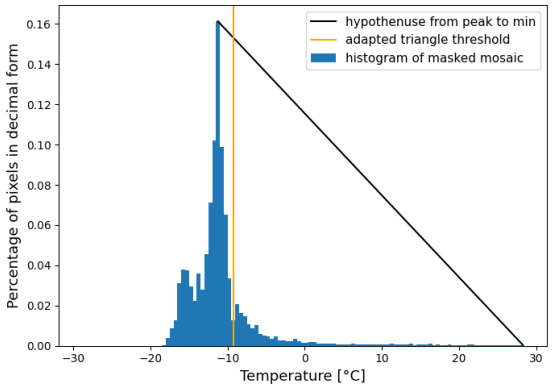


(e) Amended resulting binary thresholded mosaic

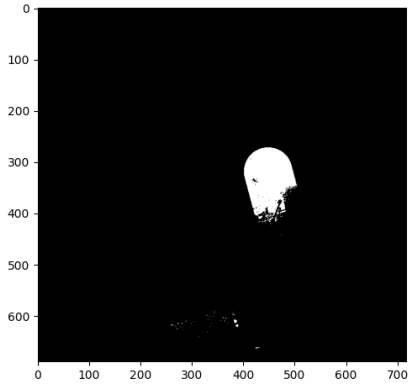
Figure A.12: Triangle histogram thresholding demonstrated using a thermal mosaic with two distinct areas causing two peaks. The peak adjustment allows the selection of only relevant pixels within the warmer region.



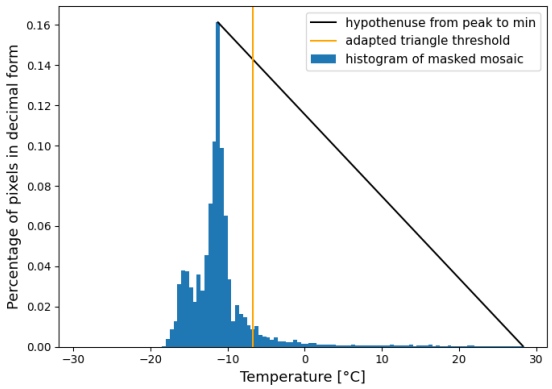
(a) Varied thermal mosaic with a large leakage (temperature colourbar in °C)



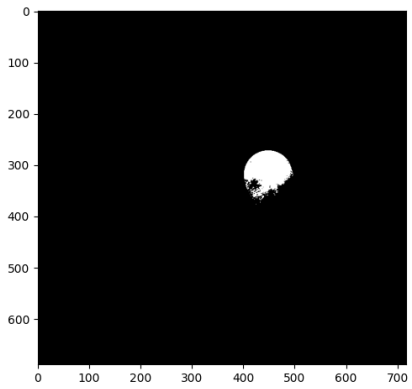
(b) Histogram and triangle threshold of thermal mosaic without consideration of bin magnitude



(c) Resulting binary thresholded mosaic with too many pixels of interest



(d) Histogram and triangle threshold of thermal mosaic with consideration of bin magnitude

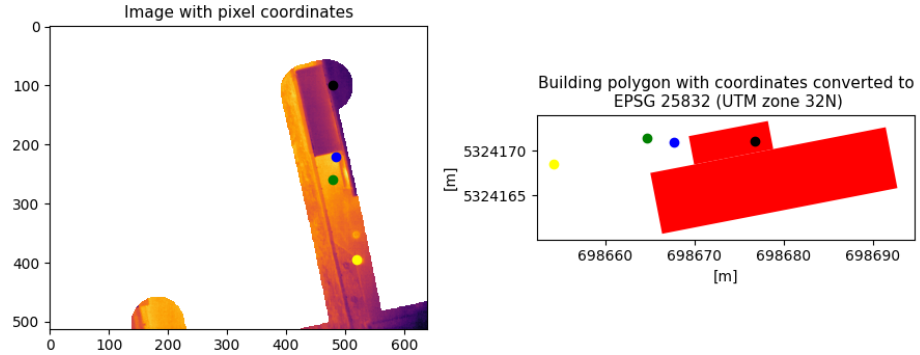


(e) Amended resulting binary thresholded mosaic

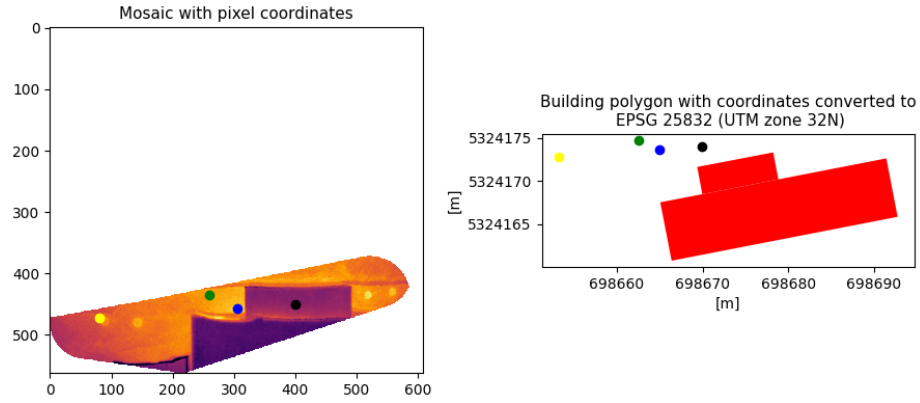
Figure A.13: Triangle histogram thresholding demonstrated using a varied thermal mosaic with a large leakage causing a too conservative threshold. The adjusted threshold ensures the region of interest is reduced to the appropriate size.

Appendix II

Figure A.14 comparatively visualises how the geotransform calculated during A.14a individual image georeferencing and georectification via A.14b WebODM or A.14c Pix4Dmapper can deviate from the true positioning. Particularly in orthomosaics it is found to be too inconsistent to use reliably in classification. Figure A.14 demonstrates this with example points, marked in relation to the depicted building[17], to illustrate the shift in comparison to the building's equivalent geographically accurate polygon.



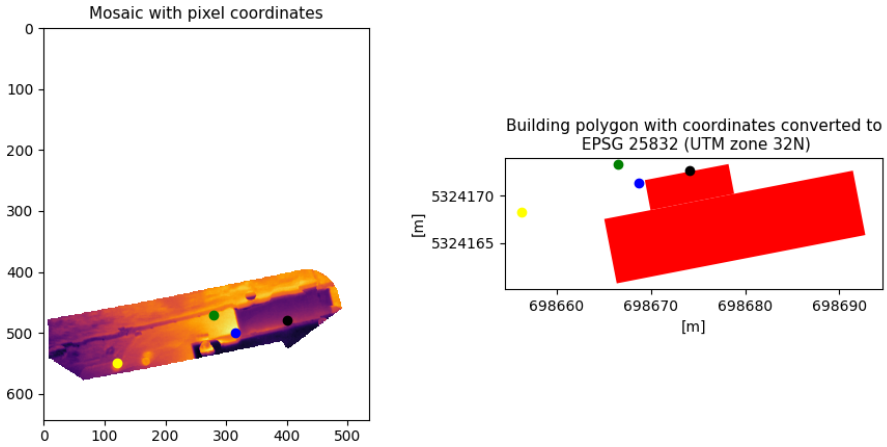
(a) Georeferencing discrepancies in an exemplary georeferenced single image



(b) Georeferencing discrepancies in an exemplary WebODM orthomosaic

Figure A.14: Possible discrepancies in georeferencing exemplified for all variations of georeferencing. The left side shows the georeferenced image or mosaic with points plotted in relation to the depicted building. The right side draws those same geographic points in relation to the building's equivalent geographic polygon. Legend: black - on building, blue - at building border, green - close to building, yellow - far from building

[17] In the Universal Transverse Mercator (UTM) projection, zone '32N' (denoting European territory), and European Terrestrial Reference System 1989 (ETRS89) map datum.



(c) Georeferencing discrepancies in an exemplary Pix4D orthomosaic

Figure A.14: Possible discrepancies in georeferencing exemplified for all variations of georeferencing (continued). The left side shows the georeferenced image or mosaic with points plotted in relation to the depicted building. The right side draws those same geographic points in relation to the building's equivalent geographic polygon.
Legend: black - on building, blue - at building border, green - close to building, yellow - far from building

Bibliography

[1] ArcGIS Pro (2021). *Documentation: Generate an Orthomosaic using the Orthomosaic Wizard*. URL: https://pro.arcgis.com/en/pro-app/latest/help/data/imagery/generate-an-orthomosaics-using-the-orthomosaic-wizard.htm (visited on 8 Feb. 2022).

[2] ArcGIS Pro (2021). *Documentation: Overview of georeferencing*. URL: https://pro.arcgis.com/en/pro-app/latest/help/data/imagery/overview-of-georeferencing.htm (visited on 25 May 2021).

[3] Axelsson, S. (1988). "Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 26(5), pp. 686–692. DOI: 10.1109/36.7695.

[4] Bäumker, M. and Heimes, F.-J. (2002). "New Calibration and Computing Method for Direct Georeferencing of Image and Scanner Data Using the Position and Angular Data of an Hybrid Inertial Navigation System". In: *Proceedings of OEEPE Workshop on Integrated Sensor Orientation*.

[5] Berg, A., Ahlberg, J., and Felsberg, M. (2016). "Enhanced analysis of thermographic images for monitoring of district heat pipe networks". In: *Pattern Recognition Letters* 83, pp. 215–223. DOI: 10.1016/j.patrec.2016.07.002.

[6] Burton, D. A. (2013). *Composite Standard Deviations*. Cary, North Carolina, USA. URL: http://www.burtonsys.com/climate/composite%7B%5C_%7Dsd.php%7B%5C#%7Dpython (visited on 23 Apr. 2021).

[7] Distante, A. and Distante, C. (2020). *Handbook of Image Processing and Computer Vision: From Image to Pattern*. Vol. 2. Cham, Switzerland: Springer International Publishing. ISBN: 978-3-030-42373-5. DOI: 10.1007/978-3-030-42374-2.

[8] FLIR Systems, Inc. (2017). *Duo Pro R: User Guide*. Product information. Version 1.0. FLIR® Systems, Inc. URL: https://www.flir.com/products/duo-pro-r/.

[9] Friman, O., Follo, P., Ahlberg, J., and Sjokvist, S. (2014). "Methods for Large-Scale Monitoring of District Heating Systems Using Airborne Thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 52(8), pp. 5175–5182. DOI: 10.1109/TGRS.2013.2287238.

[10] Gade, R. and Moeslund, T. B. (2014). "Thermal Cameras and Applications: A Survey". In: *Machine Vision and Applications* 25(1), pp. 245–262. DOI: 10.1007/s00138-013-0570-5.

[11] Garrard, C. (2016). *Geoprocessing With Python*. Shelter Island, United States: Manning Publications. ISBN: 978-1-61729-214-9.

[12] GDAL/OGR contributors (2022). *GDAL/OGR Geospatial Data Abstraction software Library*. Open Source Geospatial Foundation. DOI: 10.5281/zenodo.5884351. URL: https://gdal.org (visited on 20 Feb. 2022).

[13] Gillies, S. et al. (2013). *Rasterio: geospatial raster I/O for Python programmers*. Mapbox. URL: https://github.com/rasterio/rasterio (visited on 30 May 2021).

[14] Government of Canada (2015). *Geometric Distortion in Imagery*. Canada. URL: https://www.nrcan.gc.ca/maps-tools-publications/satellite-imagery-air-photos/remote-sensing-tutorials/satellites-sensors/geometric-distortion-imagery/9401 (visited on 19 May 2021).

[15] Groos, A. R., Bertschinger, T. J., Kummer, C. M., Erlwein, S., Munz, L., and Philipp, A. (2019). "The Potential of Low-Cost UAVs and Open-Source Photogrammetry Software for High-Resolution Monitoring of Alpine Glaciers: A Case Study from the Kanderfirn (Swiss Alps)". In: *Geosciences* 9(8), p. 356. DOI: 10.3390/geosciences9080356.

[16] Heipke, C. and Tödter, J. (2020). *Drohnengestützte Thermografie als Basis der Asset- und Instandhaltungsstrategie von Fernwärmenetzen*. Schlussbericht 19768 N. IGF, Fernwärme-Forschungsinstitut in Hannover e.V., and Leibnitz Universität Hannover.

[17] Hossain, K., Villebro, F., and Forchhammer, S. (2020). "UAV Image Analysis for Leakage Detection in District Heating Systems using Machine Learning". In: *Pattern Recognition Letters* 140, pp. 158–164. DOI: 10.1016/j.patrec.2020.05.024.

[18] Jayapalan, L., Murray, C., Chen, F., and Xie, Q. (2024). *Oracle® Spatial: Spatial GeoRaster Developer's Guide*. Guide F32256-13. Version 21c. United States: Oracle®. URL: https://docs.oracle.com/en/database/oracle/oracle-database/21/geors/spatial-georaster-developers-guide.pdf.

[19] Krist, H., Reinelt, B., Frisch, A., Kreidenweis, S., Dalsass, A., and Hofmann, A. (2017). *Wärmenetze in Kommunen - In zehn Schritten zum Wärmenetz, Leitfaden*. Tech. rep. [Heating networks in municipalities - ten steps to a heat network, a guide]. 4th. Kempten, Germany: Bayerisches Landesamt für Umwelt (LfU) and Bayerisches Staatsministerium für Wirtschaft, Landesentwicklung und Energie (StMWi).

[20] Leidner, J. L. (2016). "Georeferencing: from texts to maps". In: *International Encyclopedia of Geography: People, the Earth, Environment and Technology*. Ed. by Richardson, D., Castree, N., Goodchild, M. F., Kobayashi, A., Liu, W., and Marston, R. A. Oxford, UK: Wiley, pp. 1–10. DOI: 10.1002/9781118786352.wbieg0160.

[21] Ljungberg, S.-A. and Rosengren, M. (1988). "Aerial and Mobile Thermography to Assess Damages and Energy Losses from Buildings and District Heating Networks - Operational Advantages and Limitations". In: *XVIth ISPRS Congress, Technical Commission VII: Interpretation of Photographic and Remote Sensing Data*. Ed. by Murai, S. Vol. XXVII, Part B7. Kyoto, Japan: ISPRS, pp. 348–359. URL: https://www.isprs.org/proceedings/XXVII/congress/part7/348_XXVII-part7.pdf.

[22] Merziger, G., Mühlbach, G., Wille, D., and Wirth, T. (2013). *Formeln und Hilfen: Höhere Mathematik*. 7th. [Formulas and Equations: Higher Mathematics]. 7th. Barsinghausen, Germany: Binomi. ISBN: 978-3-923923-36-6.

[23] Olsson, M. E. (2001). "Long-term thermal performance of polyurethane-insulated district heating pipes". PhD thesis. Göteborg, Sweden: Chalmers University of Technology.

[24] OpenDroneMap (2020). *WebODM*. Software. URL: https://www.opendronemap.org/webodm/ (visited on 25 May 2021).

[25] Pix4D (2019). *What is accuracy in an aerial mapping project?* URL: https://www.pix4d.com/blog/accuracy-aerial-mapping (visited on 25 Jan. 2021).

[26] Pix4D (2020). *Pix4Dmapper*. Software. URL: https://www.pix4d.com/product/pix4dmapper-photogrammetry-software (visited on 25 May 2021).

[27] Pix4D (2021). *Pix4Dmapper: Processing steps*. URL: https://support.pix4d.com/hc/en-us/articles/115002472186-Processing-steps (visited on 25 May 2021).

[28] Sledz, A., Unger, J., and Heipke, C. (2020). "UAV-based Thermal Anomaly Detection for Distributed Heating Networks". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B1-2020, pp. 499–505. DOI: 10.5194/isprs-archives-XLIII-B1-2020-499-2020.

[29] SZ DJI Technology Co. Ltd. (2018). *Matrice 600 Pro: User Manual*. Product information. Version 1.0. SZ DJI Technology Co. Ltd. URL: https://www.dji.com/de/matrice600-pro.

[30] SZ DJI Technology Co. Ltd. (2018). *Zenmuse XT 2: User Manual*. Product information. Version 1.0. SZ DJI Technology Co. Ltd. URL: https://www.dji.com/zenmuse-xt2.

[31] University Consortium for Geographic Information Science (2020). *GIS&T DC-30 - Georeferencing and Georectification*. URL: https://gistbok.ucgis.org/bok-topics/georeferencing-and-georectification (visited on 19 May 2021).

[32] Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). "scikit-image: image processing in Python". In: *PeerJ* 2, e453. DOI: 10.7717/peerj.453.

[33] Vollmer, E., Volk, R., and Vogl, M. (2023). *Automatic analysis of UAS-based thermal images to detect leakages in district heating systems: Source code and exemplary dataset*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.7851726.

[34] Werner, S. (2017). "District Heating and Cooling in Sweden". In: *Energy* 126, pp. 419–429. DOI: 10.1016/j.energy.2017.03.052.

[35] Whitehead, K. and Hugenholtz, C. H. (2014). "Remote sensing of the environment with small unmanned aircraft systems (UASs), part 1: A review of progress and challenges". In: *Journal of Unmanned Vehicle Systems* 02(03), pp. 69–85. DOI: 10.1139/juvs-2014-0006.

[36] Xu, Y., Wang, X., Zhong, Y., and Zhang, L. (2016). "Thermal anomaly detection based on saliency computation for district heating system". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China: IEEE, pp. 681–684. DOI: 10.1109/IGARSS.2016.7729171.

[37] Zack, G. W., Rogers, W. E., and Latt, S. A. (1977). "Automatic measurement of sister chromatid exchange frequency". In: *The journal of histochemistry and cytochemistry: official journal of the Histochemistry Society* 25(7), pp. 741–753. DOI: 10.1177/25.7.70454.

[38] El-Zahab, S. and Zayed, T. (2019). "Leak detection in water distribution networks: an introductory overview". In: *Smart Water* 4(1), p. 5. DOI: 10.1186/s40713-019-0017-x.

[39] Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., and Zhang, L. (2019). "Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 151, pp. 207–222. DOI: 10.1016/j.isprsjprs.2019.02.021.

B Detecting District Heating Leaks in Thermal Imagery: Comparison of Anomaly Detection Methods

Abstract

District heating systems offer means to transport heat to end-energy users through underground pipelines. When leakages occur, a lack of reliable monitoring makes pinpointing their locations a difficult and costly task for network operators. In recent years, aerial thermography has emerged as a means to find leakages as hot-spots, with several papers proposing image analysis algorithms for their detection. While all publications boast high performance metrics, the methods are constructed around very different datasets, making a true comparison impossible.

Using a new set of aerial thermal images from two German cities, this paper implements, improves, and evaluates three anomaly detection methods for leakage detection: triangle-histogram-thresholding, saliency mapping, and local thresholding with filter kernels. The approaches are integrated into a software pipeline with globally applicable pre- and postprocessing, including vignetting correction. While all methods reliably detect thermal anomalies and are suitable for automated leakage detection, triangle-histogram-thresholding is the most robust.

Abbreviations

DHS district heating system

DR detection rate

FN false negative

FP false positive

GUI graphical user interface

IoU intersection over union

LT local thresholding

ML machine learning

P precision

R recall

RGB red green blue

SM saliency mapping

THT triangle-histogram-thresholding

TIR thermal infrared

TN true negative

TP true positive

UA unmanned aircraft

UAS unmanned aircraft system

VC vignetting correction

B.1 Introduction

B.1.1 Context

In the face of anthropogenic global warming, political and societal efforts are increasingly directed towards the buildings sector as one of the major contributors to climate change [24]. Accounting for approximately 30 % of the world's energy demand, building operation – specifically heating – is primarily responsible for these sizeable requirements [24]. For this reason, the German government has recently enacted new legislation which mandates the development of a nationwide heat supply strategy and requires all municipalities to devise comprehensive plans for climate-neutral heating [4]. The approach mirrors long-standing laws in Scandinavia, where centralised technologies – most notably district heating systems (DHSs) – are paving the way towards more sustainable cities [1].

DHSs are networks of mostly subterranean pipelines which connect energy-generating facilities with end-energy users to supply heat. When it comes to providing buildings with energy, they can offer a viable solution for densely populated areas and an alternative to individual, building-wise fossil-fuel based approaches [12]. In Denmark, for instance, such networks currently already supply two-thirds of the population with 89 % climate-neutral heat [1]. Various designs have been implemented in numerous countries throughout the past millennia. However, constant usage inevitably causes material fatigue and thus leakages to occur. In Germany, network losses have remained a constant drain on system efficiency, annually amounting to between 10 % and 14 % since 2000 [1]. This can be attributed to the fact that DHSs often lack a form of integrated monitoring and even newer networks only provide rough location estimates for leakages [9, 30]. However, performing

timely repairs is not only vital for reasons of efficiency: Pipeline leakages may precipitate serious and costly damage to the system, surrounding infrastructure, and environment if left unrepaired [30]. Therefore, finding alternative economical and reliable inspection techniques is crucial to ensuring minimal losses in a technology that has the potential of meeting our cities' future heating demands by sustainable means.

To this end, a thermography-based approach has emerged for DHS monitoring [30]. It is centred around Ljungberg and Rosengren [16]'s and Axelsson [2]'s finding that a heated medium leaking into pipeline surroundings will cause a localised spike in temperature at the surface and can thus be identified as a hot-spot in thermal infrared (TIR) images. As a remote sensing technique, this method warrants no direct contact to the networks themselves and does not rely on built-in or pre-existing technology, making it generally applicable to a broad range of DHS types. It chiefly requires the collection of TIRs above pipeline areas, a procedure which is simplified greatly by use of unmanned aircraft systems (UASs). These permit the acquisition of images with a high ground resolution in a time-efficient, flexible manner – potentially even in combination with smart city approaches [6]. However, the resulting tens of thousands of TIRs require some form of automatic analysis for the method to become financially viable to network operators as a means of DHS monitoring [9, 30]. Moreover, the selection of a robust leakage detection method is essential for success [30]: On the one hand, it must work conservatively so as not to miss important anomalies, on the other be selective enough to provide a manageable list of candidates to operators [25]. Several research groups have designed effective methodological approaches to perform the automatic detection task using custom case studies. However, on account of considerable differences in utilised data and study conditions, these have – thus far – been incomparable. Therefore, this study aims to identify a true state-of-the-art method and comparison in robust automatic anomaly detection for DHS leakage detection among existing in literature.

B.1.2 Related work

Published approaches generally follow similar procedures to output a list of leakage candidates with as few false alarms as possible. After acquiring the data at night so as to reduce thermal reflectance and irrelevant hot-spots, the images commonly undergo photogrammetric processing. By combining the georeferenced data with geographical DHS information, the images can be cropped to pipeline surroundings, thus removing anomalies outside the analysis scope. Publications diverge in their choice of anomaly detection method for image binarisation and subsequent false-alarm removal steps.

Friman et al. [9] implement a histogram-based method, identifying pixels of interest as a defined percentile of the warmest within a set of images. A watershed transform helps remove buildings and associated hot-spots. To reduce the potential for misclassification, Berg et al. [3] instead use building data from OpenStreetMap and additionally integrate a feature-based machine learning (ML) classifier to improve false alarm reduction.

Sledz et al. [20] apply a Laplacian of Gaussian blob detector and cluster merging by temperature categories to find elliptical hot-spots. They generate digital surface models to help sort out false alarms above surface level.

Xu et al. [27] and Zhong et al. [31] implement Itti et al. [13]'s saliency mapping (SM) based approach derived from the human visual system. The output, a combination of various feature maps, is binarised via maximum entropy segmentation. Some shortcomings of this approach include its non-discriminatory saliency definition (cold and warm regions are equally conspicuous) and its normalisation (neighbouring anomalies may be eliminated).

Therefore, Sledz and Heipke [19] instead modify Itti et al. [13]'s SM method by including a Max-operator to specify only warm regions as being of interest and a normalisation limited to a percentile-defined interval. For binarisation, the results are combined with simultaneously acquired red green blue (RGB) images using Dempster [8]'s and Shafer [18]'s evidence theory. This method requires detailled RGBs, meaning all data must be acquired with a dual camera during the day and not be cropped to the DHS.

Hossain et al. [10] and Hossain et al. [11] similarly do not mask their TIR images. They perform anomaly detection by local thresholding (LT) of various combined filter outputs. The results of edge and local maxima detectors are binarised via threshold and joined using a logical AND operator. For false alarm reduction, the authors adopt a convolutional neural network based classification approach and demonstrate its superiority to various conventional ML methods, including Berg et al. [3]'s.

Most recently, Vollmer et al. [25] showcase an automation of all steps in one image analysis pipeline, including previous manual georeferencing. They implement an adapted triangle-histogram-thresholding (THT) based on Zack et al. [29] for image binarisation. False alarms are removed by size, shape, and temperature difference ΔT to surroundings and classified by ΔT severity.

B.1.3 Objectives and contribution

The variety of existing approaches – all of which are said to excel at the detection task – inherently give rise to our research question: Which method is best suited for TIR analysis to help network operators identify leakages in DHSs? This, however, is difficult to answer for several reasons.

Table B.1 shows how various aspects of data acquisition deviate between assorted research groups. Differences in aircraft, flight height, sensor type, and TIR image resolution make a direct comparison impossible. Additionally, the amount of data vary greatly, with some researchers basing their method development on several cities, some only on a single one. A further obstacle is the lack of shared data and code, without which neither method nor images can be easily reviewed or transferred to new studies. Only Vollmer et al. [25] have made both their code and a dataset available online [26].

In this paper, we aim to answer the posed question for the first time and find the most suitable and robust anomaly detection method for DHS leakage detection in existence.

Table B.1: Overview of data used in the anomaly detection studies presented in chapter B.1.2.

Publication	Geographical information	Aerial vehicle	Flight height [m]	Infrared camera / sensor	Image count	Image resolution [pixels]
Friman et al. [9] Berg et al. [3]	15 Swedish and Norwegian cities	airplane	800	FLIR SC7000	>50,000	640x512
Xu et al. [27]	Gävle, Sweden	airplane	-	FLIR X8000sc	-	1280x1024
Zhong et al. [31]	Gävle, Sweden	UAV DJI S1000	120	FLIR Tau 2 640	-	640x512
	Datong, China	UAV DJI S1000	150	FLIR Tau 2 640	-	640x512
Hossain et al. [10]	7 Danish cities	UAV	-	-	27,050	-
Hossain et al. [11]	12 Danish cities	UAV	-	FLIR Tau 2 640	243,082	640x512
Sledz et al. [20] Sledz and Heipke [19]	Hannover, Germany	UAV DJI M200	40	DJI Zenmuse XT2	290	640x512
Vollmer et al. [25]	Munich, Germany	UAV DJI M600	60	DJI Zenmuse XT2	3,365	640x512

To do so, we select three approaches from Section B.1.2, which we refine with essential adaptations, novel enhancements, and parameter grid search to identify the best possible variation of each algorithm. By utilising a newly developed case study, we are able to directly compare the different approaches and assess their potentials in an unprecedented manner. To enable a true comparison, the methods are embedded in an analysis pipeline similar to Vollmer et al. [25] for identical data pre- and post-processing. This, too, is enhanced with a novel, universally applicable vignetting correction (VC) which significantly improves performance of, for instance, Vollmer et al. [25]'s THT. An exceedingly detailed evaluation – including quantitative, qualitative, and holistic assessment – supports the selection of an overall best method. Following open science principles, both code and datasets will be published alongside this study [17].

The paper is divided into five parts. Section B.2 covers the general pre- and postprocessing steps and all implemented approaches. This encompasses necessary adaptations to enable the algorithms to work with the provided data, as well as novel enhancements to optimise performance and create the best possible variant for each approach. Section B.3 presents the case study providing the foundation for all methodological development and evaluation. The data is an extension of the images used by Vollmer et al. [25], including new imagery from Munich as well as Karlsruhe to create a more substantial basis for analyses. Section B.4 lays the foundation for a sound evaluation, while all methods are assessed quantitatively, qualitatively, and in the overall context of the leakage detection pipeline in B.5. Section B.6 draws conclusions from the study, details limitations, and presents an outlook for future work.

B.2 Implemented methodologies

The following anomaly detection methods are implemented in this study: 1. Vollmer et al. [25]'s triangle-histogram-thresholding (THT) approach, inspired by Friman et al. [9], 2. Hossain et al. [11]'s local thresholding (LT) method based on filter kernels, 3. Sledz and Heipke [19]'s adaptation of Itti et al. [13]'s saliency mapping (SM). These three approaches

reflect key directions that presented studies have branched out into, the most promising of which are chosen. Only methods that process images individually are considered[1] to offset potentially occurring UAS-based acquisition effects like thermal drift. Heuristic-based approaches are selected owing to their lower requirement for labelled datasets and fewer parameters to be optimised, which makes them more efficient and practical when annotated data is scarce as is the case in this instance. While the aforementioned studies act as implementation guidelines, all approaches require adaptation to the data[2] and / or to optimise performance.

The anomaly detection methods are embedded into an image analysis pipeline, fashioned after Vollmer et al. [25]. The pipeline evaluates a set of images – hereafter referred to as a dataset – acquired in a single flight under similar conditions with the same camera (see Table B.1) in three steps:

1. Preprocessing (Section B.2.1): General image enhancement, georeferencing, and masking the images with DHS pipeline information
2. Anomaly detection (Section B.2.2): Binarising images into foreground (pixels of interest) and background
3. Leakage identification (Section B.2.3): Grouping of pixels into regions of interest and sorting out false alarms

B.2.1 Image preprocessing

Preprocessing helps enhance and prepare data for algorithm application. In the context of leakage detection, this includes reducing the search space to areas of interest. All datasets are processed according to Vollmer et al. [25] by clipping to a mean- and standard deviation-based interval (to reduce measurement errors), translating recorded intensity values to temperature arrays, georeferencing the images by estimating image-wise affine transformation matrices, and masking them with geographical DHS pipeline information. After applying these steps, every TIR UAS image T has an associated full temperature array T_u (unmasked), an affine transformation matrix A_{geo}, and a masked array T_m. This procedure is improved by including a novel, globally applicable VC (Section B.2.1.1), preceding all other steps.

B.2.1.1 Vignetting correction (VC)

In thermography, the "vignetting" effect refers to a radial distortion where image corners and edges exhibit colder values than the centre. Despite thermal cameras including automatic non-uniformity correction, the case study's TIRs significantly suffer from vignetting

[1] For this reason, Friman et al. [9]'s dataset-based thresholding is not used.
[2] The urban setting and detail of this case study's imagery effectuates a high number of false alarms, potentially posing a greater challenge than the original paper's data.

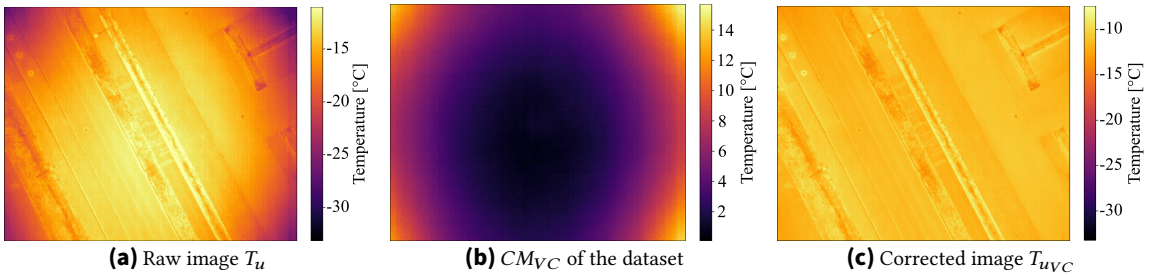
– an observation that aligns with field tests conducted by Yuan and Hua [28]. Various factors, particularly temperature and wind speeds, impact the effect's severity. As a thermal camera is found to require around 30 min to stabilise, the authors suggest capturing homogeneous images after each flight for correction. [28]

However, using calibration images has its disadvantages. It requires finding suitable scenes in the field, thus increasing acquisition effort, and precludes the correction of existing data. Therefore, a novel, universally applicable VC is developed in this study which requires no additional imagery.

The simplified approach is based on Yuan and Hua [28]'s observation that a pixel-by-pixel temperature average over all corrected images of a 30 minute acquisition window produces a near homogeneous image. This suggests that temperatures equalise across a flight and that differences between averages stem from systematic measurement errors. PWM is defined as an array of pixel-wise average temperatures of all uncorrected TIRs within a dataset. The approximated correction mask CM_{VC} is calculated as the relative difference between those pixels and PWM's minimum (Eq. B.1). A raw thermal image T_u can be corrected to $T_{u_{VC}}$ via Eq. B.2[3], as visualised in Fig. B.1. Vignetting masks such as Fig. B.1b vary greatly depending on the acquisition conditions, so it is paramount to calculate one for each dataset.

$$CM_{VC} = PWM - min(PWM) \tag{B.1}$$

$$T_{u_{VC}} = T_u + CM_{VC} \tag{B.2}$$



(a) Raw image T_u **(b)** CM_{VC} of the dataset **(c)** Corrected image $T_{u_{VC}}$

Figure B.1: Visualisation of the implemented VC. Combining a raw, unmasked TIR (B.1a) with the dataset's correction mask (B.1b) returns the corrected image (B.1c).

B.2.2 Anomaly detection

This section details the anomaly detection methods used for binarisation – in other words to divide the image into background and pixels of interest. Each approach is first outlined according to its implementation in literature, after which the novel adaptations and

[3] While this may potentially falsify T_u's absolute temperature values, only relative temperature differences are used in the context of leakage detection.

enhancements developed over the course of this study are described. The reasons for any required adaptations and the value of the contrived improvements are illustrated. Parameters are defined in Section B.4.3, where optimal method variants are found.

B.2.2.1 Triangle-histogram-thresholding (THT)

Original methodology by Vollmer et al. [25] First implemented by Vollmer et al. [25] in the context of leakage detection, the adapted THT approach functions as this study's baseline. As suggested by Friman et al. [9], a binarising threshold with which to segment an image can be found by using a histogram, a graphical representation of data distribution and frequency. The value range of the data in question is divided into intervals, with each of the ensuing so-called bins forming a class to which data points are assigned according to value. In the case of image data, this means pixels are distributed according to their intensities, culminating in columns of varying height depending on value prevalence. Naturally, pixels at the upper end of the intensity or temperature histogram are of particular interest for the identification of thermal anomalies.

Friman et al. [9] choose a threshold simply by defining the value as a specific upper percentile of a histogram generated from an entire dataset. To prevent effects like thermal drift from impacting binarisation, Vollmer et al. [25] instead calculate a histogram and image-wise threshold per masked image T_m using an approach based on Zack et al. [29].

Fig. B.2 shows how a histogram is created based on the temperature distribution of a masked image, with each class covering an interval of $0.5°C$. Point P1, determined by the centre of the tallest column, and P2, the upper end of the dataset range, define a right-angled triangle. The threshold is determined as the centre P3 of whichever bin has the greatest distance d_{max} to the hypotenuse, measured for each class by the orthogonal line connecting it's apex to the triangle at P4.

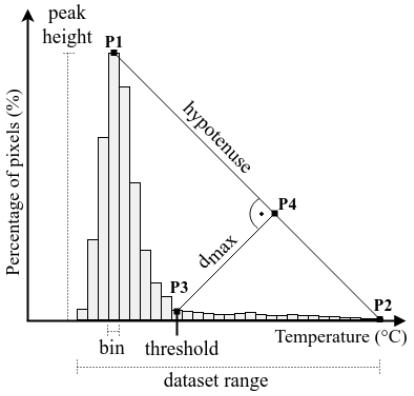


Figure B.2: Visualisation of the THT method [25]

The method is adapted to the given context of leakage detection. In histograms with multiple peaks, the warmest local maximum (furthest to the right) determines P1. To avoid segmenting overlarge anomaly areas, the selected threshold is adjusted until the associated relative frequency is less than 1 %.

Adaptations for this study. As the method was developed on similar data, no further enhancements are made aside from those described in Section B.2.1.

B.2.2.2 Local thresholding (LT)

Original methodology by Hossain et al. [10, 11] While Hossain et al. [10, 11]'s anomaly detection can also be described as thresholding-based, the authors apply a series of filter kernels to find local instead of image-wise ones with a region extraction algorithm. In a first step, the warmest image regions are found by comparing pixels to their surroundings. Eq. B.3 is applied to the unmasked array T_u to get a binary segmentation I_{warm} based on local maxima. A pixel (i, j) is selected if its temperature exceeds a combination of arithmetic mean $\mu(i, j)$ and standard deviation $\sigma(i, j)$ of the pixel's neighbourhood – with $\alpha = 1$ in Hossain et al. [11]. These surroundings are defined as a $(2 \cdot r + 1) \times (2 \cdot r + 1)$ square with radius $r = 100$ pixels in Hossain et al. [11].

$$I_{warm}(i, j) = \begin{cases} 1, & \text{if } \mu(i, j) + \alpha \cdot \sigma(i, j) < T_u(i, j) \\ 0, & \text{else} \end{cases} \tag{B.3}$$

A second filter combining vertical and horizontal Sobel edge detectors calculates a gradient image G. Hossain et al. [11] apply it under the assumption that the spread of hot water underground is characterised by diffuse temperature distributions. Consequently, gradients should not exceed a certain magnitude, as defined by the gradient-based binarisation I_{grad}.

$$I_{grad}(i, j) = \begin{cases} 1, & \text{if } \mu(i, j) + 0.5 \cdot \sigma(i, j) > G(i, j) \\ 0, & \text{else} \end{cases} \tag{B.4}$$

Both segmentation masks are combined by logical AND operator. Smaller regions and image noise are then removed via 5×5 majority voting kernel.

Adaptations for this study. An analysis in the context of this case study reveals several method problems. Hossain et al. [11]'s hypothesis regarding gradients does not hold true, as gradient-based filtering eliminates true leakages. Additionally, temperature-based binarisation has the unfortunate tendency to identify any larger, warm areas (including i.e. sidewalks) as regions of interest, contradicting the common, locally confined appearance of leakages. Thus, crucial changes are made to adapt the approach (Fig. B.3).

Instead of a single temperature-based mask generated with $r = 100$, masks are created for every radius $r \in R, R = \{r_1, ..., r_n\}$. This means one is also computed at image level with a global threshold defined by the p_{th} percentile of all pixel temperatures. Resulting binarisations are again combined by logical AND operator. For an anomaly to be included in the final mask, it must be present in all filter radii, ensuring relevant pixels of interest in both local and global context. Analogous to the original approach, noise is minimised by applying a majority vote filter kernel.

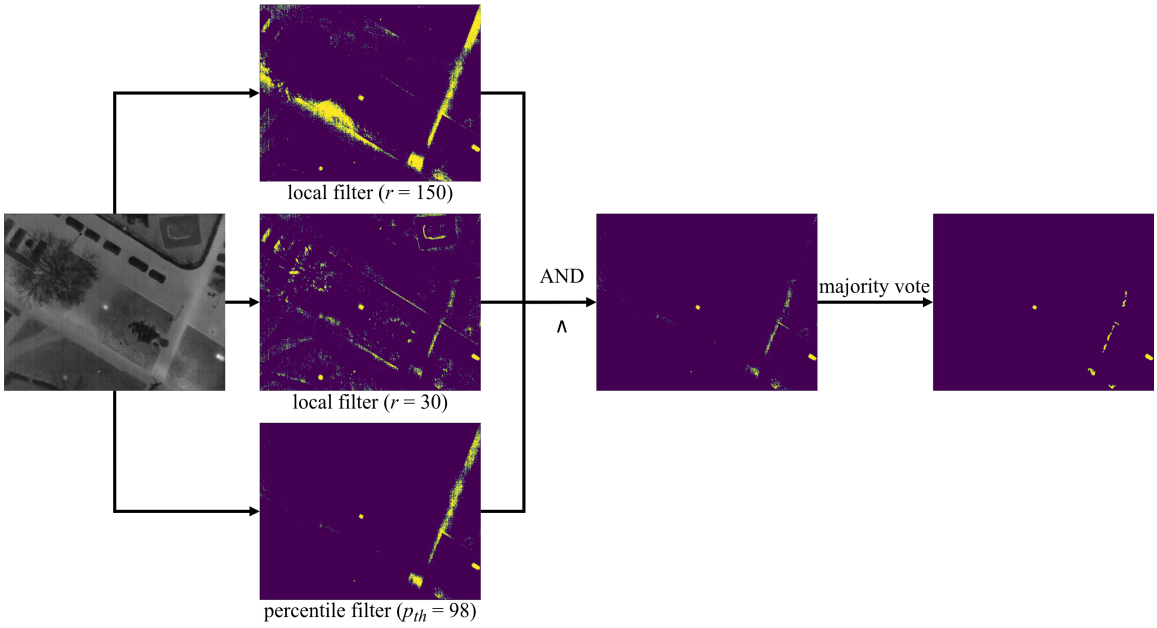


Figure B.3: Visualisation of the adapted LT process.

B.2.2.3 Saliency mapping (SM)

Original methodology by Itti et al. [13], Xu et al. [27], and Zhong et al. [31] Saliency analyses model the human brains' attention directing mechanisms for visual stimuli to find conspicuous image regions [7]. As thermal anomalies stand out in TIRs, Xu et al. [27] are the first to propose this method for leakage detection in DHSs. Based on Itti et al. [13], their algorithm returns a saliency map per image, where each pixel's value represents how strongly it stands out in the overall image context. This comprises three steps [13]: 1. Compute a set of feature maps for intensity $I(c, s)$, color $RG(c, s)$ & $BY(c, s)$, and orientation $O(c, s, \theta)$ via Gaussian image pyramids and centre-surround across-scale subtractions $\ominus$, 2. Combine feature maps into a conspicuity map for intensity $\bar{I}$, color $\bar{C}$, and orientation $\bar{O}$ through across-scale addition $\oplus$, 3. Create the saliency map via normalisation and summation.

Multiscale feature extraction relies on Gaussian image pyramids to subsample an input image into $\sigma \in \{0, ..., 8\}$ spatial scales. Feature maps are derived from performing point-by-point subtractions $\ominus$ between finer and coarser scales. Specifically, this means pixels at centre scale $c \in \{2, 3, 4\}$ are compared with their positional equivalents at surround scale $s = s + \delta, \delta \in \{3, 4\}$. While the original RGB-based method uses 6 intensity, 12 color, and 24 orientation feature maps, Zhong et al. [31] recognise that color can be omitted as TIRs are in greyscale. Intensity maps are directly calculated as the delta between c and s, while orientation maps adapt the scaled images by applying Gabor filters with various orientations $\theta \in \{0°, 45°, 90°, 135°\}$.

Every feature map F is normalised via Eq. B.5 to an interval $[0, M]$ dependent on F's maximum M, thereby eliminating amplitude variations and ensuring comparability. By including a term that subtracts the average of all local maxima $\bar{m}$ from the global one M, maps with prominent peaks are favoured over uniform ones. These outputs are combined into two conspicuity maps, which - via normalisation and summation - form the saliency map.

$$N(F) = \langle F \rangle_0^M \cdot (M - \bar{m})^2 \tag{B.5}$$

Xu et al. [27] and Zhong et al. [31] propose various adaptations for the given leakage detection task, like the combination of local and global maps. However, tests on this study's data reveal significant issues and a substantially lower accuracy than their reported 90 %. Reasons for this are unclear, though Hossain et al. [10] report similarly unsatisfactory results and attribute the deviations to the simplistic nature of Zhong et al. [31]'s imagery.

To binarise the saliency map, Zhong et al. [31] implement an adaptive thresholding technique called maximum entropy segmentation [15]. A saliency map constitutes a range of intensity values, each of which has an associated probability. The map is divided into a foreground F and background B at a threshold t, defined by probability-dependent distributions. Combining the entropies for F and B results in a function $\psi(t)$, which, when maximised, returns a segmentation with maximum information content.

Adapted methodology by Sledz and Heipke [19] Building on Itti et al. [13]'s approach, Sledz and Heipke [19] make two key adaptations:

1. Feature map calculation: Saliency maps identify *all* conspicuous regions, which contradicts the specific search for hot-spots. To suppress unwanted negative (cold) values, the maximum operator is used:

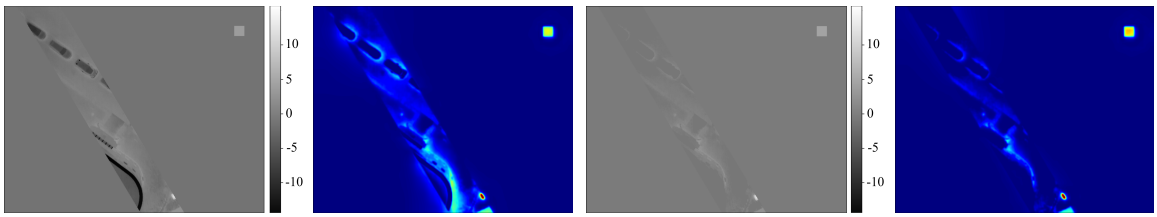
$$I(c, s) = max\left(I\left(c\right) \ominus I\left(s\right), 0\right) \tag{B.6}$$

2. Normalisation: While standard saliency analyses promote global maxima and suppress local ones, all thermal anomalies can be of interest for leakage detection. Feature map normalisation is therefore enhanced by limiting the interval range to $[percentile(F, p_{min}), percentile(F, p_{max})]$, with p_{min} and p_{max} the percentiles of F to be used. Defining $p_{max} < 100$ promotes local peaks, while values of $p_{min} > 0$ help suppress noise.

Sledz and Heipke [19] binarise the saliency maps by implementing Dempster [8]'s and Shafer [18]'s evidence theory, which combines simultaneously daytime-acquired TIRs and RGBs. As this case study consists solely of thermal data (captured at night), this approach is not directly applicable here.

Adaptations for this study. An in-depth analysis highlights a shortcoming of Sledz and Heipke [19]'s normalisation. Saliency maps vary greatly depending on p_{max}, with each image having its own optimal parameter definition. A low p_{max} emphasises local hot-spots (even in the presence of global ones), but overvalues irrelevant regions in images without anomalies. A high p_{max} risks undervaluing local hot-spots and may still highlight irrelevant regions where no anomalies exist. The method is therefore enhanced in several ways:

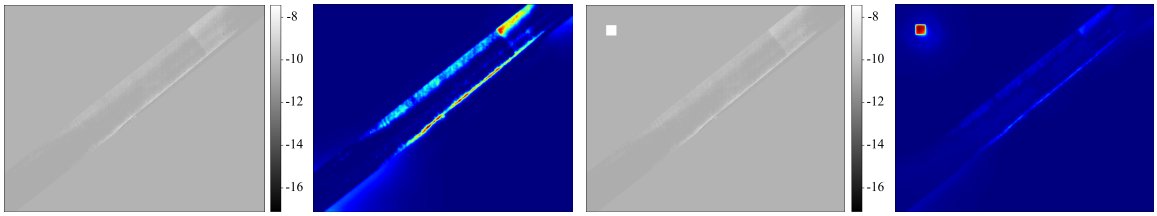
1. The approach is applied to the masked T_m to reduce false alarms. Defining masked pixels as the mean of unmasked ones prevents salient seams.
2. Saliency maps favour high temperature gradients. To prevent salient seams around cold objects, images are clipped to $[percentile(I, p_{clip}), \infty]$, limiting the lower bounds. Fig. B.4 shows the impact this has.



(a) Input image (without clipping) **(b)** Saliency map (without clipping) **(c)** Input image ($p_{clip} = 40$) **(d)** Saliency map ($p_{clip} = 40$)

Figure B.4: Comparison of generated saliency maps with and without active clipping

3. Where images lack hot-spots, irrelevant regions are systematically overestimated. Placing a $w \times h$-sized reference square into the masked area of every image (with a ΔT higher temperature) guarantees at least one artificial anomaly to counteract the effect. Fig. B.5 illustrates this.



(a) Input image (without ref square) **(b)** Saliency map (without ref square) **(c)** Input image (with ref square) **(d)** Saliency map (with ref square)

Figure B.5: Influence of a reference point on the generated saliency maps

A qualitative assessment of Zhong et al. [31]'s maximum entropy segmentation shows it provides largely robust results. However, in images lacking significant anomalies, the suggested threshold may be too low, meaning irrelevant pixels are included. In images with salient regions, the determined threshold is often too high, rejecting pixels of interest. The threshold s_{th} is therefore instead defined by equations centred around the maximum entropy threshold s_{ME}, the saliency of the previously introduced reference square s_{ref}, and a minimum threshold value s_{min} (to ensure no irrelevant anomalies are detected):

$$s_{\text{th}} = \min\left(\max\left(s_{ME}, s'_{min}, s_{ref} - \Delta s_{neg}\right), s_{ref} + \Delta s_{pos}\right) \tag{B.7}$$

$$s'_{\text{min}} = \min(s_{min}, s_{ref} - \Delta s_{min}) \tag{B.8}$$

Using the artificial reference anomaly, the equations implement two mechanisms to counteract unwanted effects: 1. Δs_{neg} and Δs_{pos} define a corridor around s_{ref}, in which s_{th} has to reside, 2. The minimal saliency threshold s_{min} can be decreased if the value is at least Δs_{min} smaller than s_{ref}.

B.2.3 Leakage identification

Section B.2.2's algorithms return binary segmentation masks with fore- and background pixels, from which anomalous regions have to be extracted.

B.2.3.1 Clustering pixels to regions

Foreground pixels are clustered into regions of interest via a multi-step procedure. First, all connected foreground pixels are grouped together and assigned an individual label. Two pixels are considered neighbours if they border one another vertically, horizontally, or diagonally. Owing to the nature of the binary segmentation cut-off, anomalies occasionally manifest as multiple clusters in close proximity to one another. Analogous to Vollmer et al. [25], all labelled regions are therefore classified based on their size as small ($\leq$ 10 pixels) or large ($>$ 10 pixels). An extended bounding box is drawn around all larger regions. If a small cluster lies entirely within the bounds of such a box, it is assigned that one's label, thus combining regions enclosed by another.

B.2.3.2 Classifying regions

A key indicator of anomaly relevance is the temperature difference ΔT between it and its immediate surroundings. Following Vollmer et al. [25], ΔT is determined by subtracting the surroundings S from the anomaly A's temperature, where T_A is defined as the average of all anomaly pixel values. If sufficiently large, T_A corresponds to the average of the anomaly's 50 or 100 warmest pixels, which prevents ΔT from being underestimated. The ambient temperature T_S is calculated by expanding the hot-spot's convex hull outward to create a surrounding ring and averaging the ring's values. To avoid falsifying T_S, all pixels belonging to other anomalies or outside of the area left by pipeline masking are excluded from the calculation.

The relevant order of magnitude of ΔT varies between studies. Sledz et al. [20] assume a required delta of at least 5 °C based on literature research, while Berg et al. [3] report a confirmed leak with only 3 °C under specific ambient conditions. Vollmer et al. [25] use a multi-step categorisation including, among others, a 10 °C or more limit based on

information from local municipal companies. This study finds instances of a ΔT lower than 5 °C not to represent leaks and also implements a categorisation into four discrete classes: uncritical ($\Delta T < 5°C$), moderate ($5°C \leq \Delta T < 10°C$), pronounced ($10°C \leq \Delta T < 15°C$), or critical ($15°C \leq \Delta T$). We refrain from further categorisation (by classifier or geographical positions) as the purpose of study lies in comparing the anomaly detection methods themselves.

B.3 Case study

B.3.1 Data

The case study comprises 3750 UAS images of two DHSs from the German cities of Munich and Karlsruhe. The water temperature in these DHSs lies between 80 °C and 130 °C depending on the season. Both studied areas have a predominantly suburban character, although the level of urbanisation and the development types differ, including single and multi-family home residential areas, commercial areas, and green spaces such as parks and forests. The inspected regions around Munich include the municipalities of Taufkirchen, Ottobrunn, and Neubiberg and constitute the larger part of the case study. Of 49 acquired datasets, 5 were selected for their high quality and depiction of diverse urban landscapes and leakage candidates. The images were acquired between 8 p.m. and 1 a.m. in December 2019, with outside temperatures of −5 °C to 2 °C. Including data from a second city such as Karlsruhe helps diversify the study, highlights the existence of city-specific features, and allows a comprehensive evaluation of the developed algorithms. The images were recorded in January and March 2022, at 0 °C to 3 °C outdoor temperatures.

Table B.2: Overview of case study dataset acquisition details.

	Munich (MU)					**Karlsruhe (KA)**	
UA	DJI Matrice 600 Pro (hexacopter)					DJI Matrice 300 RTK (quadrocopter)	
# images	2638					1112	
	MU1	**MU2**	**MU6**	**MU15**	**MU16**	**KA1**	**KA2**
Date	02.12.2019	03.12.2019	02.12.2019	04.12.2019	10.12.2019	16.01.2022	01.03.2022
Time	11.15 - 11.50 PM	00.05 - 00.40 AM	10.00 - 10.45 PM	00.17 - 00.24 AM	08.47 - 09.05 PM	03.13 - 03.45 AM	01.33 - 02.03 AM
# images	681	651	795	205	306	496	616

All flight routes were based on known DHS pipelines' positions. Both utilised Matrice unmanned aircrafts (UAs) supports automated flight along previously defined routes, with only take-off requiring manual handling. While nimble, the Matrice 300 RTK UA [23] is more susceptible to wind than the 600 Pro [21], making exact georeferencing more difficult. Acquisition of nadir images took place at 60 m altitude. A flight speed of 3 m/s ensured an 88 % image overlap, reducing the risk that leaks are overlooked. The utilised camera system – a Zenmuse XT2 by DJI and Teledyne FLIR LLC [22] – combines an optical 4K camera sensor for capturing 4000×3000-sized RGBs with a FLIR infrared sensor. The latter

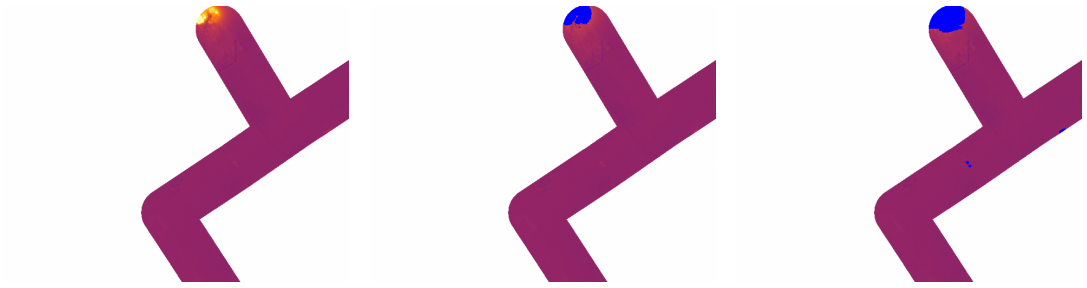
is an uncooled VOx microbolometer with a $7.5 - 13.5\ \mu m$ wavelength range, 13 mm focal length, and 640×512 resolution [22]. Images are stabilised via DJI's integrated gimbal [22]. While this setup provides both TIR and RGB images, only the thermal data and associated metadata, such as GPS positioning, are used in this case study.

B.3.2 Hard- and software

The processing pipeline is implemented in Python `v3.10` and designed to run on all common desktop operating systems. Unless stated otherwise, an Apple M1 Pro processor (8× Performance-cores, 2× Efficient-cores) and 16 GB of RAM under MacOS 13.4 was used. The grid search was performed on a "Thin" computing node of the bwUniCluster2.0, a high-performance computing cluster operated by the state of Baden-Wuerttemberg, Germany.

B.4 Preparing the evaluation

Comparing the anomaly detection methods from Section B.2.2 proves difficult owing to the lack of an objectively correct ground truth binarisation. Whether or not a region should be classified as an anomaly depends on a variety of factors, such as absolute temperature, local environment temperature, and its size. Additionally, it is impossible to define a definitively correct anomaly contour. As Fig. B.6 demonstrates, one could assign only the warmest pixels to an anomaly (as in B.6b) or include less hot, neighbouring areas (as in B.6c) – both fundamentally valid options. Even with strict annotation guidelines, manual labels will remain subject to some uncertainty. While the exact border between hot-spot and background has no serious impact on an algorithm's suitability to detect leakages, the choice of method parameters can greatly vary the resulting segmentation. This, in turn, impacts metric calculations, which strictly compare such masks to the defined ground truth.



(a) Preprocessed image T_m **(b)** Option 1 **(c)** Option 2

Figure B.6: Examples of possible segmentation masks for the same image.

The evaluation is therefore designed to minimise the influence of these factors and allow for a well-founded assessment of various method properties. A custom evaluation dataset is created to find the best parameter combinations for each algorithm via grid search. For

the aforementioned reasons and to ensure consistent labelling throughout, this dataset is generated according to the following guidelines: 1. A single expert performs all annotating to maintain uniformity, 2. A novel, custom-built labelling tool is employed to facilitate the procedure. This graphical user interface (GUI) utilises a temperature-based slider to generate a basic mask and a paintbrush tool for manual corrections to ensure the final segmentation mask remains between the two extremes from Fig. B.6. Both the created masks as well as the developed GUI labelling tool are published to Zenodo [17].

B.4.1 Evaluation dataset

A total of 290 images are selected from the datasets presented in Section B.3.1 and manually annotated with a custom labelling tool. The data is divided into training, validation, and test sets – or "splits". A high spatial overlap prevents images from being allocated entirely at random, as duplicates across splits would distort the results. All images therefore start

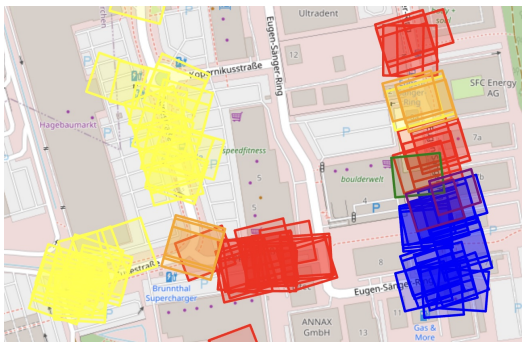


Figure B.7: Visualisation of the splitting procedure. Images in train are blue, validation red, and test yellow. Others are removed due to overlap.

	train	val	test
# images	172	52	45
MU1	38	23	
MU2			41
MU6	34	7	2
MU15	13		
MU16	4	10	2
KA1	41	12	
KA2	42		

Table B.3: Overview of the evaluation dataset.

out as part of the training set. Random ones are selected and moved to either validation or test split, together with all that share an overlap. This procedure is repeated until the target split size values are reached. Any remaining overlap at split boundaries is resolved by gradually removing images from the respective sets via a heuristic greedy algorithm, resolving as many conflicts simultaneously as possible. Special case handling ensures the desired size ratio of each split is maintained. Fig. B.7 visualises the procedure for an exemplary area, while Table B.3 details the generated splits. *MU2* is solely included in the test split, as evaluating on "unseen" data is imperative. Images of the same dataset cannot be considered entirely unrelated due to congruent acquisition conditions.

B.4.2 Metrics

The common and custom binary semantic segmentation metrics shown in Table B.4 are used to evaluate algorithm suitability. A predicted segmentation mask **P** is compared to

Table B.4: Overview of the selected evaluation metrics.

Metric	Description	Definition
Recall (R)	Proportion of foreground pixels correctly assigned to the foreground	$\frac{TP}{TP + FN}$
Precision (P)	Proportion of actual foreground pixels among all pixels assigned to the foreground	$\frac{TP}{TP + FP}$
Intersection over union (IoU) / Jaccard coefficient	A measure of the similarity between **P** and **G**	$\frac{TP}{TP + FP + FN}$
F_β score, specifically F_2 score	A combination of P and R, which for $\beta = 1$ is the harmonic mean of the two. R takes precedence over P in leakage detection[4], so F_2 is more suitable. [14]	$(1 + \beta^2) \cdot \frac{P \cdot R}{(\beta^2 \cdot P) + R}$
Detection rate (DR)	Proportion of anomalies in **G** that are also in **P**, where an anomaly is detected if its bounding box in **P** has a >30 % overlap with one in **G**[5].	
Detection rate 30 (DR_{30})	Corresponds to the DR of anomalies larger than 30 pixels.	

the ground truth **G** on pixel level using true positive (TP), true negative (TN), false positive (FP), and false negative (FN) values.

On their own, precision (P) and recall (R) cannot judge semantic segmentation mask quality. A meaningless model will achieve maximum R by assigning all pixels to the foreground or a high P by defining only a small number of unambiguous areas as the foreground. Therefore, intersection over union (IoU) and F_β have become key metrics for a holistic model evaluation. Next to mask quality, the amount of recognised ground truth anomalies must be assessed, which is why custom detection rate (DR) metrics are included.

B.4.3 Parameter grid search

A grid search is performed for each method to specify the various parameters influencing segmentation outputs and thus find an optimal variant for later comparisons. Grid searches originate from ML, where they are used for hyperparameter tuning [5]. A set of options is defined for each variable. The model is trained for all parameter combinations, evaluated on the validation split, and the optimal combination chosen based on a metric. This procedure can be adapted to identify suitable parameters for conventional anomaly detection methods by omitting model training.

In this study, grid search is applied to the parameter-rich SM and LT algorithms using 50 randomly chosen images from the training split, as this permits a feasible runtime. Table B.5 shows the investigated parameters and their values, resulting in 20,000 combinations for SM and 448 for LT. THT does not require a grid search as it can be used as per Vollmer et al. [25]. Appendix B.I highlights the range the mentioned performance metrics assume on account of the parameter grid search via statistical analysis.

[4] Because anomaly existence takes precedence and FPs are removed in subsequent steps.
[5] 30 % coverage ensures enough of the anomaly is included, also for candidate inspection.

Table B.5: Overview of the parameters used in the grid search.

Method	Parameter	Description	Values
SM	ΔT	temperature delta between background and reference square	3, 4, 5, 6, 7
	p_{clip}	clipping percentile	30, 40
	(p_{min}, p_{max})	normalisation interval	(9999, 5), (9999, 20), (20, 5), (40, 5), (60, 5)
	$(\Delta s_{neg}, \Delta s_{pos})$	permissible interval for the selected threshold	(0, 100), (1, 99.99), (1, 99.98), (1, 99.97), (1, 99.96)
	s_{min}	minimum permissible threshold	70, 80, 90, 98, 110, 120, 130, 145, 160, 175
	(w, h)	width and height of the reference square	(15, 15), (25, 15), (25, 25), (35, 35)
	$\lvert \ominus \rvert$	whether to use the absolute of the centre-surround difference	true, false
LT	R	set of radii to be used for local filters	(10, 150), (20, 150), (30, 150), (30, 200), (40, 150), (40, 160), (50, 150), (30, 40, 60, 100, 150)
	p_{th}	percentile for the percentile filter	10, 90, 95, 96, 97, 98, 99
	α	multiplier for determining the threshold value	1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7

Four suitable parameter combinations are found for both methods by optimising Section B.4.2's evaluation metrics. The *MaxIoU* configuration maximises IoU. Since this can favour high P over R, the configurations *MaxIoU@85* and *MaxIoU@90* ensure an R of at least 85 % or 90 %. The final configuration, $MaxF_2$, maximises F_2, which is comparable to setting a minimum R limit. Table B.6 lists the identified, optimal parameter values.

Table B.6: Grid search results for selected parameters and four metric optimisations.

Method	Parameter	MaxIoU	MaxIoU@85	MaxIoU@90	$\mathbf{MaxF_2}$
SM	ΔT	7	7	7	7
	p_{clip}	40	40	40	40
	(p_{min}, p_{max})	(0, 100)	(0, 100)	(1, 99.98)	(0, 100)
	$(\Delta s_{neg}, \Delta s_{pos})$	(9999, 5)	(9999, 5)	(9999, 20)	(9999, 5)
	s_{min}	90	70	80	70
	(w, h)	(15, 15)	(25, 15)	(15, 15)	(25, 15)
	$\lvert \ominus \rvert$	False	False	False	False
LT	R	(30, 40, 60, 100, 150)	(30, 40, 60, 100, 150)	(50, 150)	(50, 150)
	p_{th}	99	98	98	99
	α	1.3	1.2	1.3	1.3

B.5 Method evaluation and comparisons

B.5.1 Quantitative evaluation

The optimal method variants from Section B.4.3 are quantitatively evaluated using the metrics described in B.4.2. THT is included both in its original form [25] and this study's

version, which incorporates VC (Section B.2.1.1). The evaluation results (Table B.7) are subject to variance due to the small split sizes and potentially ambiguous categorisation of image regions. The differences between the validation and test set results highlight this fact, a possible explanation for which is the test set's more suburban nature and large-scale, confirmed leakage. The validation split was randomly sampled from all available data and contains a greater proportion of typical urban anomalies such as warm cars. Generally, all analysed methods can reliably detect anomalies if suitable parameters are selected. As expected, configurations that achieve a higher R or DR often score lower in P and IoU.

Table B.7: Results of the leakage detection method variants evaluated on validation and test sets. Best results are in bold.

Mode	Config	Validation						Test					
		IoU	F_2	R	P	DR	DR_{30}	IoU	F_2	R	P	DR	DR_{30}
THT	with VC	59.8	77.5	79.5	70.7	88.6	88.2	47.8	**72.8**	**79.5**	54.5	79.8	85.3
	without VC	54.0	65.3	62.4	**80.1**	78.1	79.6	37.0	50.3	48.1	61.6	37.8	42.2
SM	MaxIoU	**60.3**	80.0	83.4	68.5	88.6	92.5	**55.0**	67.4	65.2	77.7	72.3	80.4
	MaxIoU@85	57.1	81.2	88.1	61.8	**94.3**	94.6	53.3	67.6	66.4	73.0	75.6	82.4
	MaxIoU@90	52.5	**80.3**	**90.2**	55.6	93.3	**95.7**	46.0	71.8	79.2	52.3	**85.7**	**92.2**
	MaxF_2	57.1	81.2	88.1	61.8	**94.3**	94.6	53.3	67.6	66.4	73.0	75.6	82.4
LT	MaxIoU	51.6	62.7	59.6	79.4	71.4	79.6	35.2	43.4	39.0	**78.1**	63.0	67.6
	MaxIoU@85	52.8	71.9	74.0	64.8	84.8	93.5	37.7	49.4	46.4	66.7	76.5	80.4
	MaxIoU@90	49.7	76.0	84.1	54.9	89.5	94.6	43.3	57.4	55.5	66.4	79.0	83.3
	MaxF_2	51.8	63.7	73.5	63.7	78.1	84.9	43.3	53.7	49.9	76.4	66.4	71.6

THT achieves high DRs of 88.6 % and 79.8 % on both validation and test splits, with a simultaneously high IoU score of 59.8 % and 47.8 % respectively. Comparing results to THT without VC highlights the importance of pre-processing for a reliable leakage detection, especially for a high DR. The SM method achieves an IoU of 60.3 % on the validation and 55.0 % on the test set for the *MaxIoU* configuration. However, the DR on the test set is comparatively low at 72.3 %. The *MaxIOU@90* configuration boasts the overall highest DR of 85.7 %, though it is accompanied by a significant IoU decrease, indicating a higher amount of FP detections. Out of all methods, LT performs worst in terms of IoU, with additionally low DRs across all configurations. A 55.5 % R for *MaxIoU@90* on the test set is particularly striking. As the same configuration achieves over 90 % on the training data, this method's robustness for different acquisition areas and framework conditions must be called into question.

B.5.2 Qualitative evaluation

While the quantitative analysis greatly depends on the specific variants and their segmentation outputs, a qualitative analysis can highlight characteristic method properties and differences. The generated segmentation masks will therefore be compared in an exemplary fashion using select images. Fig. B.8 shows examples of common scenarios in their original form, preprocessed according to Section B.2.1, and with the various anomaly detection algorithms applied. For SM and LT, the *MaxIoU* and *MaxIoU@90* configurations

are taken into account. Insights from this analysis can only be generalised to a certain extent, as they are influenced by various factors.

Fig. B.8.1 shows the segmentation results for an image containing a confirmed leak with an exceptionally high surface temperature as well as smaller hot-spots. THT is able to identify the significantly colder, yet still anomalous manhole covers in the lower part of the image. In contrast, both SM configurations do not identify the manholes at all. This suppression of less significant anomalies when warmer ones occur is concerning. The LT configurations only classify the warmest leakage pixels as being anomalous. This may explain the low R on the test set, as a larger area was annotated.

A closer look at the segmentation masks highlights a tendency of SM configurations to predict large-scale detections. This is particularly true for images containing sections of buildings or their facades, as demonstrated by Fig. B.8.2. Similar behaviour cannot be observed in any other method. Fig. B.8.2 also shows how selecting a uniform, image-wise threshold – as done by THT – can be problematic for complex imagery. Only the warmest anomaly is detected here, while some of the smaller hot-spots on the cold building roof are not classified in spite of their comparatively high temperature difference. Fig. B.8.3 shows an example where THT defines such a high threshold that the warm manhole cover detected by all other methods is missed.

Overall, all implemented methods are generally capable of detecting ano-malies with a significant ΔT to their surroundings. Inconsistencies can be observed primarily in SM and LT methods, which cause an increased number of false-positives or the non-detection of relevant anomalies in individual images. Among the examined methods, THT is most consistent overall, with especially larger anomalies being reliably detected. Only complex images – where the selection of a uniform threshold for the entire image does not enable a sufficiently precise differentiation – can be considered problematic.

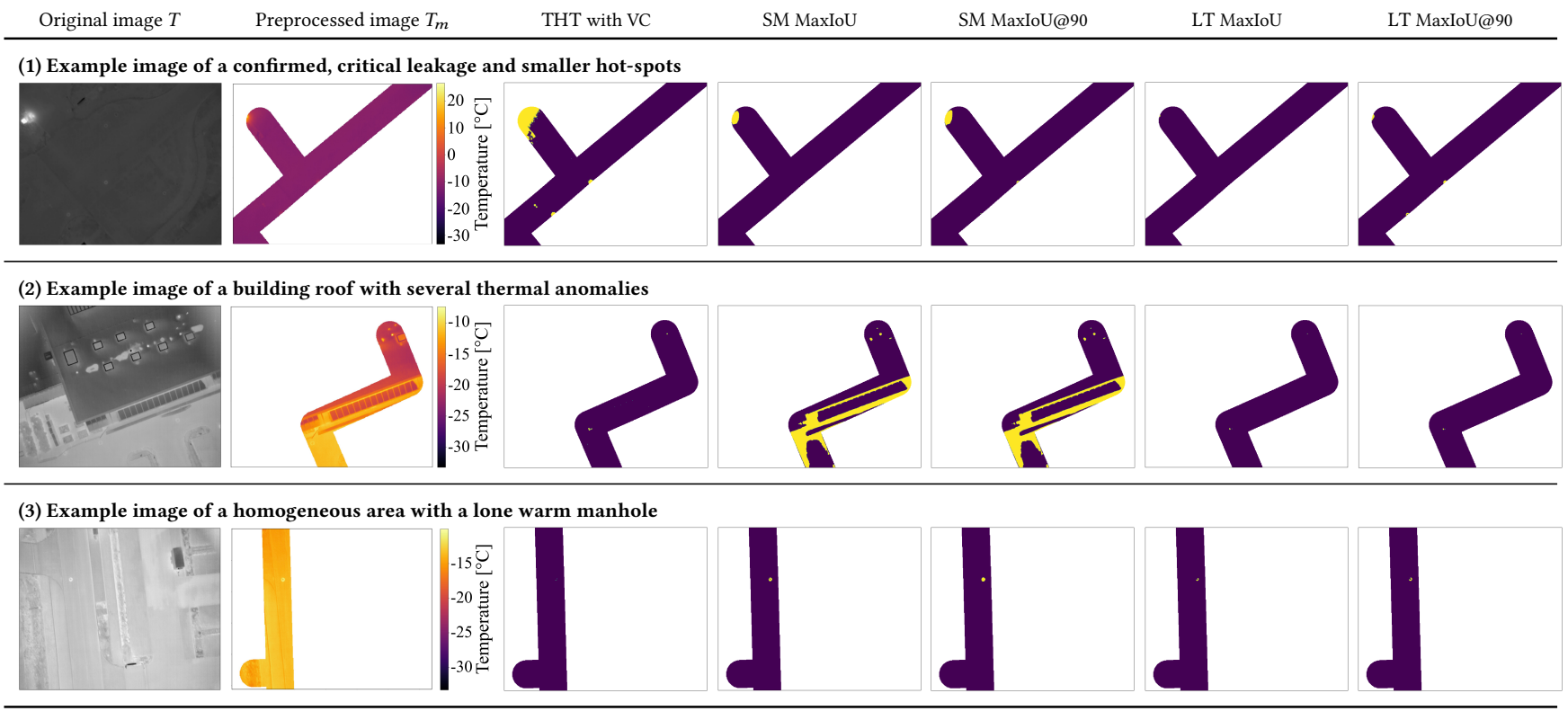


Figure B.8: Visualisation of segmentation mask results of the different configurations for three example scenarios.

B.5.3 Evaluation of the analysis pipeline

So far, the algorithms were evaluated as standalone components of the analysis pipeline. However, the entire processing pipeline must be considered to assess their suitability for leakage detection in DHSs. A challenge in doing so is the small amount of confirmed leakages in this study's datasets and literature, meaning conclusions drawn about method reliability are somewhat restricted. The pipeline from Section B.2 is run on one dataset per city: *MU2* and *KA1*. The former includes a confirmed, very critical leakage, while the latter has a very urban character and therefore offers a great variety of heat sources (meaning FPs) for method assessment.

Table B.8: Leakage detection pipeline evaluation results for the datasets *MU1* and *KA1*

		MU2			***KA1***		
		THT with VC	**LT** MaxIoU90	**SM** MaxIoU85	**THT** with VC	**LT** MaxIoU90	**SM** MaxIoU85
# of anomalies		709	2209	1128	647	1586	668
average anomaly area		195.4	92.8	257.4	202.7	223.1	158.0
Classified by ΔT	uncritical	561	2066	951	567	1506	593
	moderate	105	105	148	62	66	63
	pronounced	18	15	10	18	14	12
	critical	25	23	19	0	0	0
# of relevant anomalies		148	143	177	80	80	75
Classified by type (manually)	leakage	20	19	19	0	0	0
	manhole	82	64	57	51	50	51
	car	46	60	101	10	10	10
	other				19	20	14

Table B.8 summarises the results after anomaly clustering (see Section B.2.3.1) and classification by temperature difference ΔT into four categories (see Section B.2.3.2). All relevant anomalies – meaning moderate or higher ($\Delta T > 5\ ^{\circ}C$) – are manually classified to identify the top two occurring types of urban features. For $MU2$, these are found to be leakages and manholes; for $KA1$, manholes and cars.

Conspicuously, the different method results differ significantly for *MU2*. The number of identified anomalies, for instance, is much lower in THT than LT. However, as becomes apparent through temperature-based classification, the vast majority of these additional anomalies lie below the 5 °C limit and can thus be considered irrelevant. Manual categorisation focuses on leakages and manholes and shows that all algorithms are equally reliable at detecting the confirmed leakage. The amount of detected and relevant manhole covers differs, although this can be attributed to the fact that they do not fall under the 5 °C limit when the cold inner area of the cover is included in the anomaly. While the amount of anomalies categorised as "other" vary strongly, a more in-depth analysis shows that most of these have low absolute temperatures and clearly are not critical. Results from *KA1* evaluation paint a similar picture. While LT starts off with more than twice the others' anomaly count, no significant differences exist where relevant anomalies are concerned. In fact, the number of warm vehicles is identical across all methods, while manhole count differs only slightly.

B.6 Conclusion, limitations, and outlook

To find the most suitable algorithm for leakage detection in TIR imagery of DHSs, this paper augmented and compared three anomaly detection methods from literature using an enhanced case study from Germany. In principle, all analysed methods are capable of reliably identifying significant thermal hot-spots. This applies in particular to those caused by critical leakages with a considerable ΔT to their immediate surroundings. Differences between methods are especially evident regarding their reliability in detecting weaker anomalies and robustness in complex images containing pronounced temperature gradients not associated with leakages. SM delivers considerably more robust detection results than described in literature, though there is still a tendency towards large-scale false-positive detections. Despite enhancements and adaptations, LT continues to exhibit shortcomings in reliably detecting less conspicuous anomalies. THT was greatly improved through the proposed vignetting correction and now delivers robust detection results, highlighting the importance of appropriate preprocessing of image data. In several cases, including VC has a more significant impact than utilising another method.

Naturally, this study is subject to some limitations. Despite the diversified data, confirmed leakages are scarce which complicates generalising conclusions. A lack of published datasets, such as Friman et al. [9]'s who mention 400 confirmed leakages, means only independently acquired images could be included. This study therefore focused on the algorithms themselves, limiting the use of mechanisms to distinguish between actual leakages and false alarms. Comparisons between the discussed methods are still meaningful owing to their global application to the same data. The difficulty of human error in manual labelling is addressed as best as possible, though it remains subject to some uncertainty. A lacking willingness to share code by all except Vollmer et al. [26] prevents the exact replication, verification, and testing of some implemented methodologies described in literature. Several parameters are not specified by the original authors, leaving their definition open to interpretation and hampering reproducibility of their results. All methods were implemented to the best of our ability, as found in Ruck et al. [17].

The described findings present several opportunities for future research. Further method development and evaluation would benefit from a similar analysis on datasets containing a diverse range of confirmed leakages. Implementing deep learning to perform anomaly detection may present a viable alternative to the compared conventional methods. Such a model could be honed to the more specific task of leakage identification instead of general anomaly detection. The research can be expanded to include the classification of anomalies. Some studies mentioned in Section B.1.2 use ML to this end, though the models are not state of the art. Modern deep learning approaches could improve the reliable classification of leakages and false alarms.

Acknowledgements

The images were acquired in collaboration with the Air Bavarian GmbH and Munich's and Karlsruhe's municipal utilities companies. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

Funding

This work is supported by funding from the European Union through the AI4EOSC project (Horizon Europe) under Grant number 101058593.

CRediT author statement

Elena Vollmer: Conceptualization, Methodology, Investigation, Data Curation, Visualization, Writing — original draft, Writing — review and editing; **Julian Ruck:** Methodology, Investigation, Data Curation, Software, Formal analysis, Visualization, Writing — review and editing; **Rebekka Volk:** Investigation, Writing — review and editing, Supervision; **Frank Schultmann:** Writing — review and editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Data availability statement

The data used in this study was derived from our own experiments. The complete dataset including segmentation masks is available upon request and is also published online at `Zenodo.org`, together with the developed code [17].

Appendices

Appendix I. Grid search performance statistics

Table B.9 gives statistical insight into the performance range achieved by the algorithm variants found through the parameter grid search. Mean and standard deviations are calculated across all parameter combinations described in Section B.4.3 - 20,000 and 448 for SM and LT respectively. As the grid search was limited to those two methods, no variants for THT exist with which statistical values could be calculated. The high standard deviations demonstrates a broad performance range and highlights the importance of the grid search and finding optimal parameter constellations.

Table B.9: Statistical results across all the grid search parameter combinations. The data is formatted as "mean ± standard deviation" and given in %.

Method	IoU	F_2	R	P	DR	DR_{30}
SM	28.6 ± 7.1	48.8 ± 15.9	58.0 ± 27.6	51.9 ± 21.8	60.6 ± 29.8	63.9 ± 29.0
LT	39.4 ± 13.1	62.7 ± 13.8	74.5 ± 20.0	53.4 ± 22.0	83.5 ± 14.3	88.1 ± 13.5

Table B.10 depicts mean and standard deviations of the leakage detection algorithm variants applied to validation and test datasets, as described in Section B.5.1. These statistics are calculated over 55 variants for SM and 64 for LT, which constitute a heuristic choice of the most promising parameter constellations and a compromise between the grid search runtime and complete coverage of the parameter space. The number of tested variants differs between the two methods because each has its own, respective amount of parameters – a factor which considerably impacts runtime. Table B.11 gives an overview of the parameter values used specifically in these variants. Mean values are significantly higher and standard deviations considerably reduced compared to Table B.9 owing to the more focused parameter choice.

Table B.10: Statistical results across leakage detection algorithm variants applied to validation and test datasets. The data is formatted as "mean ± standard deviation" and given in %.

Method	**Dataset**	IoU	F_2	R	P	DR	DR_{30}
SM	validation	56.2 ± 2.9	80.5 ± 0.6	87.5 ± 2.7	61.4 ± 4.7	92.5 ± 3.4	94.7 ± 2.3
	test	52.0 ± 2.7	69.4 ± 3.0	70.5 ± 6.6	68.0 ± 9.2	79.3 ± 5.6	85.7 ± 4.5
LT	validation	51.4 ± 2.7	69.9 ± 4.9	71.9 ± 9.0	66.6 ± 10.1	80.6 ± 7.6	87.5 ± 6.7
	test	38.9 ± 2.6	50.6 ± 4.5	47.7 ± 5.6	69.7 ± 8.3	73.1 ± 7.4	77.1 ± 6.8

Table B.11: Overview of most promising parameters from in the grid search. Combinations were generated only from the "used values" to be applied to the validation and test datasets.

Method	**Parameter**	**Used Values**	**Unused Values**
SM	ΔT	4, 5, 6, 7	3
	p_{clip}	30, 40	-
	(p_{min}, p_{max})	(9999, 5), (9999, 20)	(20, 5), (40, 5), (60, 5)
	$(\Delta s_{neg}, \Delta s_{pos})$	(0, 100), (1, 99.99), (1, 99.98)	(1, 99.97), (1, 99.96)
	s_{min}	70, 80, 90, 98, 110	120, 130, 145, 160, 175
	(w, h)	(15, 15), (25, 15), (25, 25)	(35, 35)
	$\mid \ominus \mid$	true, false	-
LT	R	(30, 150), (30, 200), (40, 150), (40, 160), (50, 150), (30, 40, 60, 100, 150)	(10, 150), (20, 150)
	p_{th}	96, 97, 98, 99	10, 90, 95
	α	1.0, 1.1, 1.2, 1.3, 1.4, 1.5, 1.6, 1.7	-

Bibliography

[1] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Hauptbericht 2022 [Main report 2022]*. Report. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht (visited on 28 June 2024).

[2] Axelsson, S. (1988). "Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 26(5), pp. 686–692. DOI: 10.1109/36.7695.

[3] Berg, A., Ahlberg, J., and Felsberg, M. (2016). "Enhanced analysis of thermographic images for monitoring of district heat pipe networks". In: *Pattern Recognition Letters* 83, pp. 215–223. DOI: 10.1016/j.patrec.2016.07.002.

[4] Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen [The German Federal Ministry for Housing, Urban Development and Building] (2023). *Gesetz für die Wärmeplanung und zur Dekarbonisierung der Wärmenetze [Law for heat planning and decarbonization of heat networks]*. Germany. Enacted on 17 November 2023, effective from 1 January 2024. URL: https://www.bmwsb.bund.de/SharedDocs/gesetzgebungsverfahren/DE/kommunale-waermeplanung.html.

[5] Chicco, D. (2017). "Ten quick tips for machine learning in computational biology". In: *BioData Mining* 10(1), p. 35. DOI: 10.1186/s13040-017-0155-3.

[6] Coelho, B. N. (2019). "UAVs and Their Role in Future Cities and Industries". In: *Smart and Digital Cities: From Computational Intelligence to Applied Social Sciences*. Cham, Switzerland: Springer, pp. 275–285. DOI: 10.1007/978-3-030-12255-3_17.

[7] Cong, R., Lei, J., Fu, H., Cheng, M.-M., Lin, W., and Huang, Q. (2019). "Review of Visual Saliency Detection With Comprehensive Information". In: *IEEE Transactions on Circuits and Systems for Video Technology* 29(10), pp. 2941–2959. DOI: 10.1109/tcsvt.2018.2870832.

[8] Dempster, A. P. (1968). "A Generalization of Bayesian Inference". In: *Journal of the Royal Statistical Society, Series B (Statistical Methodology)* 30(2), pp. 205–247. DOI: 10.1007/978-3-540-44792-4_4.

[9] Friman, O., Follo, P., Ahlberg, J., and Sjokvist, S. (2014). "Methods for Large-Scale Monitoring of District Heating Systems Using Airborne Thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 52(8), pp. 5175–5182. DOI: 10.1109/TGRS.2013.2287238.

[10] Hossain, K., Villebro, F., and Forchhammer, S. (2019). "Leakage Detection in District Heating Systems Using UAV IR Images: Comparing Convolutional Neural Network and ML Classifiers". In: *27th European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain: European Association for Signal Processing (EURASIP). DOI: 10.23919/EUSIPCO45326.2019.

[11] Hossain, K., Villebro, F., and Forchhammer, S. (2020). "UAV Image Analysis for Leakage Detection in District Heating Systems using Machine Learning". In: *Pattern Recognition Letters* 140, pp. 158–164. DOI: 10.1016/j.patrec.2020.05.024.

[12] International Energy Agency (IEA) (2023). *World Energy Outlook 2023*. Technical report. Paris, France: IEA. URL: https://www.iea.org/reports/world-energy-outlook-2023 (visited on 28 June 2024).

[13] Itti, L., Koch, C., and Niebur, E. (1998). "A Model of Saliency-Based Visual Attention for Rapid Scene Analysis". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(11), pp. 1254–1259. DOI: 10.1109/34.730558.

[14] Jadon, S. (2020). "A survey of loss functions for semantic segmentation". In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. Via del Mar, Chile: IEEE, pp. 1–7. DOI: 10.1109/cibcb48159.2020.9277638.

[15] Kapur, J., Sahoo, P., and Wong, A. (1985). "A new method for gray-level picture thresholding using the entropy of the histogram". In: *Computer Vision, Graphics, and Image Processing* 29(3), pp. 273–285. DOI: 10.1016/0734-189X(85)90125-2.

[16] Ljungberg, S.-A. and Rosengren, M. (1987). "Aerial Thermography - A Tool For Detecting Heat Losses And Defective Insulation In Building Attics And District Heating Networks". In: *Thermosense IX: Thermal Infrared Sensing for Diagnostics and Control*. Vol. 780. Orlando, United States: SPIE, pp. 257–343. DOI: 10.1117/12.940525.

[17] Ruck, J., Vollmer, E., Volk, R., and Vogl, M. (2024). *Detecting District Heating Leaks in Thermal Imagery: Comparison of Anomaly Detection Methods - Source Code and Datasets*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.11085776.

[18] Shafer, G. (1976). *A Mathematical Theory of Evidence*. Princeton University Press. ISBN: 978-0691100425. DOI: 10.2307/j.ctv10vm1qb.1.

[19] Sledz, A. and Heipke, C. (2021). "Thermal Anomaly Detection Based on Saliency Analysis from Multimodal Imaging Sources". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-1-2021, pp. 55–64. DOI: 10.5194/isprs-annals-V-1-2021-55-2021.

[20] Sledz, A., Unger, J., and Heipke, C. (2020). "UAV-based Thermal Anomaly Detection for Distributed Heating Networks". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B1-2020, pp. 499–505. DOI: 10.5194/isprs-archives-XLIII-B1-2020-499-2020.

[21] SZ DJI Technology Co. Ltd. (2018). *Matrice 600 Pro*. Version 1.0. Product information. URL: https://www.dji.com/matrice600-pro (visited on 28 June 2024).

[22] SZ DJI Technology Co. Ltd. (2018). *Zenmuse XT 2: User Manual*. Version 1.0. Product information. URL: https://www.dji.com/downloads/products/zenmuse-xt2 (visited on 28 June 2024).

[23] SZ DJI Technology Co. Ltd. (2020). *Matrice 300 RTK*. Version 1.0. Product information. URL: https://enterprise.dji.com/matrice-300/specs (visited on 28 June 2024).

[24] United Nations Environment Programme (UNEP) (2024). *Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector*. Tech. rep. Nairobi, Kenya: UNEP, Global Alliance for Building and Construction (GlobalABC). DOI: 10.59117/20.500.11822/45095.

[25] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[26] Vollmer, E., Volk, R., and Vogl, M. (2023). *Automatic analysis of UAS-based thermal images to detect leakages in district heating systems: Source code and exemplary dataset*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.7851726.

[27] Xu, Y., Wang, X., Zhong, Y., and Zhang, L. (2016). "Thermal anomaly detection based on saliency computation for district heating system". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China: IEEE, pp. 681–684. DOI: 10.1109/IGARSS.2016.7729171.

[28] Yuan, W. and Hua, W. (2022). "A Case Study of Vignetting Nonuniformity in UAV-Based Uncooled Thermal Cameras". In: *Drones* 6(12), p. 394. DOI: 10.3390/drones6120394.

[29] Zack, G. W., Rogers, W. E., and Latt, S. A. (1977). "Automatic measurement of sister chromatid exchange frequency". In: *Journal of Histochemistry and Cytochemistry* 25(7), pp. 741–753. DOI: 10.1177/25.7.70454.

[30] El-Zahab, S. and Zayed, T. (2019). "Leak detection in water distribution networks: an introductory overview". In: *Smart Water* 4(1), p. 5. DOI: 10.1186/s40713-019-0017-x.

[31] Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., and Zhang, L. (2019). "Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 151, pp. 207–222. DOI: 10.1016/j.isprsjprs.2019.02.021.

C Leak Detection using Thermal Imagery: Deep learning versus Traditional Computer Vision State-of-the-Art

Abstract

As a cornerstone of climate-neutral heat supply in urban areas, district heating systems require monitoring to detect and mitigate leaks in their subterranean pipelines. Recent research has focused on an approach involving thermography, where leaks are detected as hot-spots in remote sensing imagery. To this end, various traditional computer vision algorithms have been implemented to automate anomaly detection.

This paper pursues a new approach that has so far received little attention in the context of leak detection in district heating pipelines: deep learning, specifically supervised semantic segmentation. By creating a generalisable, multi-stage training procedure to tackle the prevalent limited dataset problem, various architectures are tailored to this anomaly detection task, of which the SegFormer-B2 with Tversky loss is found to perform best. Via comprehensive quantitative, qualitative, explainable AI, and holistic evaluation, the model is assessed and compared to state-of-the-art traditional algorithmic alternatives. It is found to excel, outperforming previous intersection over union scores by almost 10 %pt and maintaining a high precision with little detriment to recall and detection rate.

Abbreviations

AI artificial intelligence

BCE Binary Cross Entropy

CAM class activation map

CNN convolutional neural network

CV computer vision

DHS district heating system

DL deep learning

DR detection rate

FCN fully convolutional network

FFN feed-forward network

FN false negative

FP false positive

GPU graphics processing unit

IoU intersection over union

LR learning rate

LT local thresholding

MiT Mix Transformer encoders

ML machine learning

MLP multi-layer perceptron

NN neural network

P precision

R recall

RGB red green blue

SM saliency mapping

SMP Segmentation-Models-PyTorch

THT triangle-histogram-thresholding

TIR thermal infrared

UAS unmanned aircraft system

VC vignetting correction

xAI explainable AI

C.1 Introduction

When it comes to providing energy to buildings, district heating systems (DHSs) offer a viable solution for urban areas and an alternative to individual, fossil-fuel-based approaches [24]. These mainly subterranean pipeline networks can supply heat from energy-generating facilities to end-energy users in an efficient and low-emissions manner – such as in Denmark, where two thirds of the population receive 89 % climate-neutral heat via DHS [2]. However, constant use over decades inevitably causes material fatigue, and thus leaks to occur. If left unchecked, these can precipitate serious damage to the system and surrounding infrastructure [13]. Considering current heat-related goals in the effort to limit anthropogenic global warming [47], a vital part of enabling sustainable cities must be to ensure the high efficiency, and thus minimal thermal losses, of these types of infrastructure.

Unfortunately, DHSs commonly either lack a form of integrated monitoring or can only provide rough leak location estimates, calling for alternative monitoring techniques [57]. To this end, a thermography-based approach has emerged, centred around Axelsson [3]'s and Ljungberg and Rosengren [30]'s finding that a heated medium leaking into pipeline surroundings will cause a localised temperature spike at the surface. This, in turn, can be identified as a hot-spot in thermal infrared (TIR) images, the acquisition of which has been greatly simplified through recent developments in unmanned aircraft system (UAS) technology. However, for the method to become financially viable for system operators, some form of automatic analysis must be performed to identify potential leaks in the tens of thousands of resulting images [13].

This highly specific branch of image analysis comes with its own set of challenges. Firstly, the nature of thermal data is fundamentally different to standard red green blue (RGB) imagery, a much more common field of research. Where the integer pixel value in each channel of an RGB will combine to a shade and hue of colour, TIRs consist of decimal temperature values that can vary greatly [53]. Given the nature of thermal sensors and UAS-based acquisition method, TIRs can suffer from various unwanted effects, such as vignetting, material-dependent measurement errors, and weather influences [50, 53]. Secondly, the task of identifying anomalies and associated existing methodology in RGBs cannot be translated directly to what is required here. Where classical outlier detection focusses on identifying data points that deviate from the majority [35], this application works at a finer spatial resolution and defines anomalies as clusters of warm pixels – also known as hot-spots – within TIR imagery. Lastly, due to the urban setting of DHSs, the number of such anomalies is greatly inflated by naturally warm elements in city environments, such as cars, manholes, street lamps, and people, which require sorting out [50]. Together with data processing and false alarm removal, anomaly detection is therefore considered one of the key steps in TIR-based DHS leak detection [52, 53].

Fuelled by the growing availability of computing resources, artificial intelligence (AI) has emerged as a fast-growing field of research with a wide range of practical applications [16]. In the case of image processing, deep learning (DL) in particular has become highly relevant owing to its versatility and performance [16]. While the use of standard RGB data is most

common, the last decade has seen an increased interest in application of DL to imagery beyond the visible light spectrum [19]. Other approaches for general pipeline inspection, such as via acoustic emission signal analysis, already implement DL to great effect [43]. Despite this, previous work on TIR-based DHSs leak detection has focused on traditional computer vision (CV) methods to identify anomalies, such as saliency mapping, local thresholding, and histogram-based methods [52]. Machine learning (ML) or, seldomly, DL has so far only been used for their classification [5, 22, 52], mainly because annotation creation is an exceptionally labour-intensive undertaking [9, 53].

This paper therefore investigates the suitability of DL, specifically semantic segmentation, for the key task of finding anomalies in TIR imagery. Taking into account all previously mentioned challenges, our contributions can be summarised as follows:

1. We prepare our specialised UAS-based TIR data for DL model training, thereby building a novel thermal anomaly segmentation dataset. Aside from specific preprocessing and input channel selection, this entails a new approach for overcoming the challenge of time-consuming annotation: automatic label generation using the best-performing traditional CV algorithm [52].
2. We propose a novel multi-stage training procedure to adapt DL to the unconventional data type and problem setting by combining the use of a large, generated dataset and small, manually labelled dataset with established adaptation techniques.
3. Through a series of ablation studies, we find the best suited DL architecture and configuration among current state-of-the-art semantic segmentation convolutional neural network (CNN) and transformer models for our real-world use case in heat-related inspection.
4. Following Vollmer et al. [52]'s form of comprehensive assessment enables us to directly compare our novel DL model variants with previously analysed traditional algorithms to determine the best approach. The evaluation is enhanced with explainable AI (xAI) for a more detailed analysis of model behaviour.
5. In line with open science principles, our UAS-based TIR DL model training dataset [39] and code[1] will be published alongside this paper to ensure reproducibility.

To this end, the paper is structured as follows: Section C.2 discusses related literature and the research gap; Section C.3 describes methodology, from data processing over model selection to implementation; Section C.4 includes a thorough model evaluation and simultaneous comparison to classical CV methods, while Section C.5 concludes the paper with an outlook.

[1] https://www.github.com/emvollmer/TASeg

C.2 Related work

After Axelsson [3] and Ljungberg and Rosengren [30]'s initial discovery, several publications discuss automatic TIR image analysis for finding DHS leakages. They generally describe a two-part problem to generate a list of meaningful suspects for network operators: 1. Extracting anomalous pixel regions and 2. Removing false alarms, meaning hot-spots not stemming from leaks [52]. While methods for both vary, the latter often includes an initial photogrammetric processing to map the images and remove areas outside the pipeline scope by masking with DHS location information [52].

In summary, the following research groups have developed methodology throughout the past decade. While each focuses on a different region – ranging from central and northern Europe to China –, all locations are characterised by a similar, generally colder climate and thus prevalence of DHSs.

1. Friman et al. [13] and Berg et al. [5] use a histogram-based method to find anomalies as the warmest percentile of pixels. They implement photogrammetric processing and experiment with feature-based ML classifiers for false alarm reduction, finding random forest to perform best [5].
2. Xu et al. [56] and Zhong et al. [59] develop a saliency mapping (SM) approach, with Sledz and Heipke [44] suggesting modifications. They also employ image georeferencing and masking.
3. Sledz et al. [45] implement a Laplacian of Gaussian blob detector, merging elliptical hot-spots to anomalous regions by temperature. In addition to image mapping and masking, they generate a digital surface model to remove false alarms.
4. Hossain et al. [21] and Hossain et al. [22] identify anomalies by applying local thresholding (LT) to various filter outputs and combining results. For false alarm removal, they implement a CNN as well as feature-based conventional classifiers and find the DL model to surpass ML alternatives, including Berg et al. [5]'s random forest.
5. Vollmer et al. [53] utilise an enhanced triangle-histogram-thresholding (THT) algorithm for hot-spot detection and remove false alarms by initial photogrammetric processing and post-extraction size, shape, and temperature evaluation.

While all describe their methods as high performing, disparate datasets and a lack of availability prevented the most suitable anomaly detection algorithm from being identified. Vollmer et al. [52] solve this problem by creating a dataset of two German cities and consistent pre- and post-processing framework. They implement, enhance, and compare the most promising algorithms, namely SM, LT, and THT, through a comprehensive quantitative, qualitative, and holistic evaluation. THT is found to be the most reliable with novel measures like vignetting correction (VC) included in pre-processing. In contrast to most related work, they publicly share both code and datasets [38]. [52]

As is clear from the given overview, traditional algorithms have so far dominated the field of anomaly detection for finding DHS leaks via TIR imagery. This can likely be attributed to the novelty of the domain and required annotation effort.[2] Following Vollmer et al. [52]'s insights, we are able to address multiple gaps in literature in this paper. We propose a solution to the annotation quandary, present an optimised DL model for the anomaly detection problem, and are able to compare the best AI variants to the existing classical algorithms by performing the same holistic evaluation.

C.3 Methodology

C.3.1 Data preparation

To allow for comparable results, Vollmer et al. [52]'s UAS-based TIR datasets [38] form the basis of this study. The given data consist of almost 3.000 images from 7 UAS flights of the two German cities Munich and Karlsruhe. Specific acquisition guidelines were adhered to to ensure that useable imagery was captured for the task at hand. Flights were carried out in the colder seasons and at night for minimal thermal reflectance and a maximum delta between DHS flow temperatures and the environment [53]. This ensures that leaks can be clearly distinguished as thermal anomalies from their surroundings [53]. Furthermore, flights are only performed in dry weather conditions, as rain and snowfall greatly diminish TIR image quality [53]. Due to the nature of thermal sensors, TIR resolution is already considerably lower than that of standard RGBs, meaning the adherence to these conditions is essential to obtaining viable data.

Pre-processing is focused on counteracting unwanted effects in TIR data and focusing the analysis on areas of interest. Therefore, it mainly encompasses VC to mitigate radial distortion, the extraction of temperature arrays, clipping data to reduce measurement errors, and georeferencing to remove areas outside the DHS pipeline scope [52, 53]. This provides every image T with a full corrected temperature array T_u (unmasked) and one reduced to the relevant areas T_m (masked) [52]. The combination of stringent acquisition guidelines and image preprocessing mean that the temperature distributions across datasets are comparatively similar (see Appendix C.I). With regards to DL model training, the focus will therefore lie on honing a model to the given data as opposed to generalisability.

Originally developed to handle standard RGB imagery, DL models commonly expect three-channel inputs [50]. If one wishes to use inputs of differing channel counts, an additional convolutional layer can be included to map these to the expected three [50]. An ablation study was conducted to identify the most suitable combination of input channels, which is summarised in Appendix C.II. The best model performance is achieved by using all available data and stacking to match the required dimensions: (T_m, T_m, T_u). Doubling the masked temperature array helps focus the model on the task at hand, while including

[2] For their AI classification model training, Hossain et al. [22] report annotating 243,082 images by hand.

the unmasked T_u provides additional context information, in particular to areas close to the masking border. Given the TIR resolution of 512×640, this creates data inputs of dimension $(512, 640, 3)$. Data normalisation is based on the image channels' arithmetic means and standard deviations.

Of the acquired images, 290 were manually labelled at the pixel-level using a custom labelling tool. This annotated subset was the only data previously used by Vollmer et al. [52] for method development. TIRs were divided into train, validation, and test splits via random assignment and heuristic greedy algorithm to remove overlapping images in different splits.

In contrast, this study uses both the small, labelled subset and previously unannotated images for DL model development. Instead of performing laborious annotations by hand, the best traditional CV algorithm from Vollmer et al. [52] - namely THT - is instead used to this end. The method identifies a fitting threshold per image based on the assumption that pixels of interest will reside in the upper tail, and thus warmer end, of the TIR-based histogram [53]. In its general form, the algorithm draws a triangle hypotenuse from the histogram's peak to its outer right edge [53]. Orthogonal distances between hypotenuse and each bin are calculated iteratively to find the longest, which in turn defines a threshold as the corresponding bin's temperature value [53]. Specific adaptations help tailor the algorithm to the task at hand, including a nuanced peak selection to ensure the warmest among all local maxima is used and the placement of a pixel percentage limitation on the chosen threshold to prevent overestimation [53].

Applying the THT method to a TIR produces a binary labelling mask with each pixel defining the corresponding image's as anomalous or not. These outputs match those created by manual annotation and can therefore be used together for DL model training. The divide into splits is performed according to the afore-mentioned procedure from Vollmer et al. [52]. Table C.1 provides an overview of Vollmer et al. [52]'s manually annotated and the newly generated sets, both of which are used for model training. As images from the same dataset and thus UAS flight cannot be viewed as completely independent [52], one dataset – specifically *MU2* – is excluded from all but the test set [52]. This allows for an unbiased assessment during model evaluation.

Table C.1: Overview of the automatically generated and manually annotated datasets. Data based on Ruck et al. [38].

	Generated		**Manual**		
	Train	Val	Train	Val	Test
# images	2142	404	172	52	45
MU1	355	155	38	23	
MU2					41
MU6	691	71	34	7	2
MU15	168	7	13		
MU16	294		4	10	2
KA1	162	117	41	12	
KA2	472	54	42		

C.3.2 Model development

C.3.2.1 Neural network architectures

Of the many possible DL approaches worth considering for anomaly detection, this study implements supervised binary semantic segmentation. The choice reflects the definition of anomalies introduced in Section C.1 as warm regions, or hot-spots, within a TIR image. Such pixel-wise granularity is necessary not only to allow for a comparison with the traditional CV algorithms, but in particular to enable post-processing steps for false alarm mitigation, such as anomaly shape analysis [53]. While image-level classification or even object detection would make for simpler problem formulations, these approaches yield only coarse outputs – either a single label per image or anomaly bounding boxes – making them unsuitable here. Consequently, this task might more precisely be described as anomaly segmentation [35], as it diverges somewhat from the classical, predominantly classification-based field of outlier detection [35]. However, to remain consistent in our comparison with traditional CV methods [53], we will refer to it using the umbrella term anomaly detection.

A comparatively new research field, semantic image segmentation builds upon Long et al. [31]'s fully convolutional network (FCN), an architecture capable of pixel-wise classifications. Most modern segmentation CNNs follow one of two common design patterns: encoder-decoder structures that enable precise boundary delineation and pyramid pooling modules that capture multi-scale contextual information [7]. Numerous architecture adaptations have led to significant performance improvements and established models like the U-Net [37] (encoder-decoder with skip connections), PSPNet [58] (pyramid pooling), and DeepLabV3+ [8] (a hybrid of spatial pooling and decoder) as benchmarks in the field. Very recently, however, transformer-based models have been shown to outperform these architectures in different semantic segmentation tasks [29].

Transformers are a class of neural networks (NNs) characterised by highly parallelised processing and the use of self-attention mechanisms that allow global correlations in the input data to be captured [48]. The models generally consist of a set of serially connected encoders and decoders, which convert the input data into a set of latent vectors and then generate output data from said vectors. Such architectures, while highly effective across various applications, require a comparatively high amount of resources for training and, thus, specialised hardware. For this reason, variants such as Xie et al. [55]'s lightweight SegFormer have recently been introduced, which consist of far fewer parameters and have been shown to outperform common architectures such as the Swin-Transformer while requiring minimal hardware specs [29].

In light of these currently competing architecture variants, we perform experiments to assess the suitability of the transformer versus conventional semantic segmentation CNNs for the specific task at hand. To this end, the SegFormer is compared with the three mentioned, common architectures – U-Net [37], PSPNet [58], and DeepLabV3+ [8] – each representing a different NN type. The results of these experiments are presented in Section C.3.2.4, after discussing the details of the training procedure and configuration.

C.3.2.2 Training procedure

A multi-stage training procedure was developed to adapt DL architectures to the given circumstances. These are challenging owing to the unusual image type and lack of sufficiently large, labelled dataset. The latter is particularly problematic in higher-resolution data, such as UAS-based imagery, as increased granularity produces more specific, and therefore less generalisable, samples [40]. The developed methodology, as visualised in Fig. C.1, is model independent, easily transferable, and can be used across architectures to improve performance when facing similarly challenging use cases.

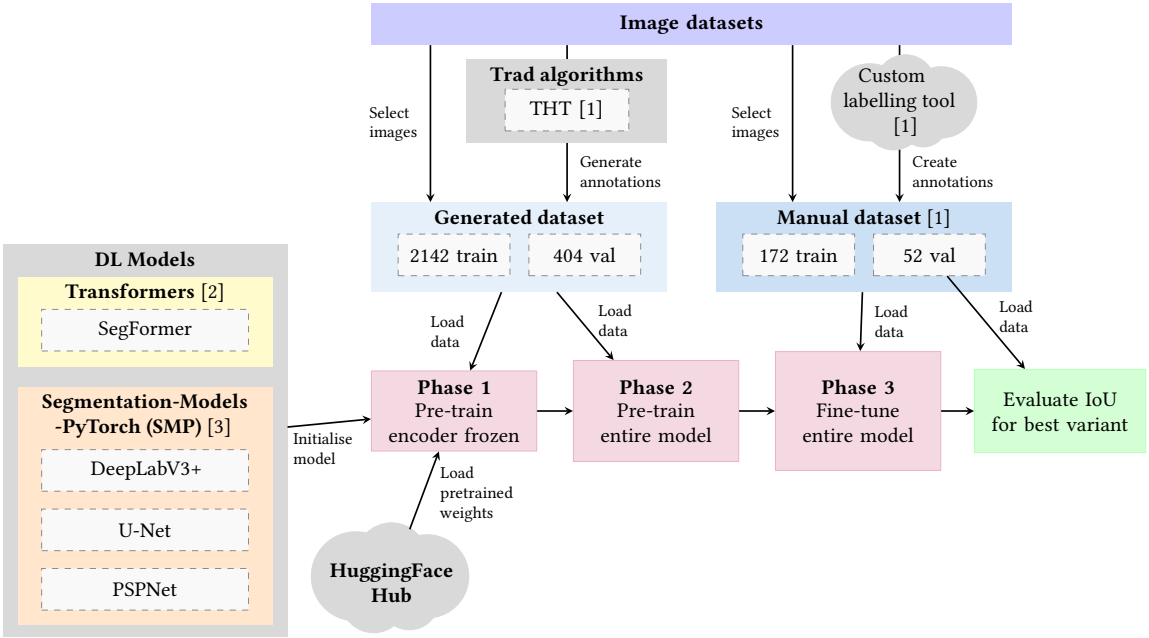


Figure C.1: Visualisation of the developed multi-stage DL model training procedure. Reference [1] refers to Vollmer et al. [52], [2] to Wolf et al. [54], and [3] to Iakubovskii [23].

When applied to the task at hand, it enables a step-wise adaptation from a known domain – multi-class semantic segmentation of RGB imagery – to the target one – binary segmentation of imbalanced, UAS-based TIR data. This begins by initialising with RGB-learned weights, followed by leveraging a large, automatically labelled TIR dataset, and concludes with the application of manually annotated images. The first adaptation step thus shifts the focus from RGB to the infrared spectrum, thermal-specific patterns, and general characteristics of the target domain. In a second adaptation, the model is fine-tuned to a small, high quality dataset to learn precise class boundaries and correct any previously induced biases.

Tackling the common issue of limited annotated datasets begins at model initialisation by implementing transfer learning: Instead of starting with random weight values, the model is loaded with pretrained ones from training on public databases. While most available weights are based on RGB datasets, studies such as Li et al. [28] show that adopting these to TIRs still significantly improves the performance of semantic segmentation. In this

study, we therefore use the fully and densely labelled semantic segmentation database ADE20K [60] at a resolution of 512×512 pixels, which is popular in benchmarking due to a high scene diversity, large number of classes, and detailed annotation granularity.

To make use of the limited amount of annotated data to its fullest extent, training itself is divided into three phases. As visualised in Fig. C.1, the number of epochs increases with each phase to reflect their growing importance in refining model performance.[3] The first two phases exploit binarised outputs from the conventional THT approach as segmentation masks, thus addressing one of the key challenges in UAS-based semantic segmentation: labour-intensive annotation [9]. In phase 1, all layers of the encoder are frozen for the first rounds of training. This prevents large gradients early on, which can cause the encoder to lose its previously learnt ability to extract meaningful features while also minimising resource requirements [16]. Training continues in phase 2 without frozen weights to better adapt the entire model to the imbalanced binary TIR dataset, and allow it to learn problem-specific features. In the last phase, the model is fine-tuned on the manually annotated data. This final and longest training with a small learning rate helps the model master the specifically desired anomaly segmentation behaviour. The variant that achieves the highest intersection over union (IoU) score on the validation split is selected at the end of training. Appendix C.III breaks down the impact of each of the described training phases as well as the multi-step procedure as a whole and highlights the advantages of using a generated dataset alongside a high-quality manually labelled one.

C.3.2.3 Training configuration

In addition to the afore-described procedure, the following hyperparameters are chosen for model training. Ablation studies help identify some of the most suitable choices for this use case.

Loss function As is common in real-world semantic segmentation, this study's datasets are characterised by a significant class imbalance [9, 33]. With binary problems such as this one, the majority of pixels are assigned to the background, which causes the other class – anomalies – to be under-represented in both instance and pixel counts [33]. Highly skewed data are problematic for all manner of DL, including semantic segmentation, as they bias a model towards the majority class [27]. The selection of a suitable loss function helps counteract this unwanted effect by emphasising the importance of the minority class during training [26, 27].

Table C.2 summarises an ablation study using four such functions. Performance is measured based on the common semantic segmentation metrics IoU, recall (R), precision (P), and F_β

[3] For the exact values, see Appendix C.III and Section C.3.2.4.

score[4]. With ground truth annotations y and model predictions $\hat{y}$, the following is given for false positives (FPs) and false negatives (FNs): $FP = \hat{y} \cdot (1 - y)$ and $FN = (1 - \hat{y}) \cdot y$.

The four tested loss functions are Binary Cross Entropy (BCE), Jaccard, Dice, and Tversky. While BCE uses similarity between y and $\hat{y}$ at pixel level [26], Jaccard is derived from IoU with some adjustments to ensure differentiability. Specifically, the intersection operator is replaced by multiplication and the union by summation or subtraction, while adding ϵ prevents a division by zero [11]. Dice loss is derived in analogous fashion from the dice coefficient [46]. Lastly, Tversky generalises dice loss for refined control over weighting of FP (via α) and FN (via β) [41]. With $\beta > \alpha$ and $\alpha + \beta = 1$, we penalise FN more than FP and increase R, as is desired for leak detection.

Table C.2: Ablation study for loss function selection.

Name	Function	Performance			
		IoU	F_2	R	P
BCE	$L_{BCE}(y, \hat{y}) = -(y \log(\hat{y}) + (1 - y) \log(1 - \hat{y}))$	68.5	80.9	80.7	81.8
Jaccard	$L_J(y, \hat{y}) = 1 - \frac{(y \cdot \hat{y}) + \epsilon}{(y + \hat{y} - y \cdot \hat{y}) + \epsilon}$	67.6	79.6	79.0	82.5
Dice	$L_D(y, \hat{y}) = 1 - \frac{(2y \cdot \hat{y}) + \epsilon}{(y + \hat{y}) + \epsilon}$	**69.1**	80.7	80.0	**83.6**
Tversky ($\alpha = 0.3, \beta = 0.7$)	$L_T(y, \hat{y}) = 1 - \frac{(y \cdot \hat{y}) + \epsilon}{(y \cdot \hat{y}) + \alpha \cdot FN + \beta \cdot FP + \epsilon}$	68.1	**81.5**	**81.9**	80.1

Overall, the performance scores show only minor differences between the tested loss functions. However, during training with a frozen encoder, BCE was found not to converge in IoU with simultaneous convergence of P towards 1 and R to 0. This indicates that BCE does not sufficiently penalise a FN classification of foreground pixels under certain conditions due to the strong data imbalance, making the function unsuitable for this study. Both Dice and Tversky losses showed the most promising results and were selected for final model training.

Learning rate scheduler and optimiser To enable optimal model convergence, the learning rate (LR) is decreased in the course of model training.[5] Two schedulers are tested: 1. a `PolynomialLR` with exponent 1.0 [55], which reduces the LR in a linear fashion, and 2. a `ReduceLROnPlateau`, which lowers it when a select variable - i.e. validation loss - does not decrease for a certain number of training steps. Owing to better performance, `PolynomialLR` is selected. Analogous to Xie et al. [55], an AdamW optimiser [32] is used to adjust the model weights according to the scheduler-defined LR.

Data augmentation To avoid overfitting on account of the comparatively small splits, data augmentation is implemented by randomly modifying the training set to increase

[4] Here F_2, as R takes precedence over P in leak detection [52].

[5] This prevents individual training steps from inciting drastic changes which impede the finding of local optima [16].

image amounts. In the context of image processing, common techniques include mirroring, enlargement, section rotations, or a combination thereof [16]. For this case study, the applied transformations comprise vertical or horizontal mirroring, elastic distortion, or so-called `ShiftScaleRotate`, whereby random image rotation is combined with either section enlargement or reduction and a random horizontal or vertical shift. Prior to these, the entire temperature array is altered through the addition of a value selected at random from a uniform distribution over the interval $[-2, 2]$. This helps increase robustness against temperature fluctuations and focus the model on relative differences rather than absolute values.

C.3.2.4 Final model selection

SegFormer architectures are available with various Mix Transformer encoders (MiT) ranging from small (B0) to large (B5), out of which B0 to B4 are tested. As per Section C.3.2.1 and analogous to Xie et al. [55], results for the conventional semantic segmentation CNN DeepLabV3+ [8] with a ResNet101 encoder pre-trained on ImageNet [10] is included. Both U-Net [37] and PSPNet [58] are trained in similar fashion for comparative purposes. To reduce the imbalance between fore- and background classes, all annotation masks with less than 40 anomalous pixels were excluded from the training datasets.

All models were trained with a batch size of 16 for 15, 35, and 60 epochs, respectively, in the three training phases of Section C.3.2.2. In the case of the SegFormer-B4, the batch size was halved to 8 to accommodate the required graphics memory and the epoch count increased to 95 for the fine tuning phase to ensure complete convergence. In all cases, the images of the training dataset were artificially duplicated to increase dataset size for phase three.[6]

Table C.3 shows the performance achieved by the different model variants on the validation dataset. Among the CNNs, the DeepLabV3+ most often achieves the best results, closely followed by U-Net, while PSPNet lags behind with up to 18.8 %pt difference in metrics across and 15.2 %pt between loss functions. Generally, however, the SegFormer architectures outperform all conventional models, especially when comparing results for each loss function. Among the transformer variants, the midrange encoders B2 and B3 show the most promise.

These results differ from Xie et al. [55] in two significant ways. Firstly, the fact that SegFormers B2 and B3 achieve the highest IoUs contradicts literature such as Xie et al. [55], where IoU scores increase continually with architecture size. While the reason for this is unknown, it seems plausible that overfitting effects may still occur due to the small dataset size. Secondly, DeepLabV3+ variants not only yield lower IoUs than all SegFormers, but habitually score less across all other metrics when comparing results for each loss function. As Xie et al. [55] report that their similarly configured DeepLabV3+ scores a

[6] This is equivalent to increasing the number of epochs by a factor of two and determining the evaluation metrics for the validation dataset in every second epoch.

Table C.3: Comparison of model variants on the validation split. Results are colour-coded from white (low) to green (high).

Model	Encoder	Loss	IoU	F_2	R	P
SegFormer	MiT-B0	Dice	66.9	79.5	79.0	81.4
SegFormer	MiT-B0	Tversky	65.2	82.2	84.5	74.0
SegFormer	MiT-B1	Dice	66.6	79.6	79.4	80.4
SegFormer	MiT-B1	Tversky	65.2	81.2	82.8	75.4
SegFormer	MiT-B2	Dice	69.5	79.8	78.4	85.9
SegFormer	MiT-B2	Tversky	70.2	84.5	85.9	79.4
SegFormer	MiT-B3	Dice	71.5	82.4	81.8	85.0
SegFormer	MiT-B3	Tversky	69.1	83.2	84.2	79.4
SegFormer	MiT-B4	Dice	64.8	78.0	77.6	79.8
SegFormer	MiT-B4	Tversky	67.4	81.2	81.7	79.3
U-Net	ResNet101	Dice	62.9	75.1	73.7	81.0
U-Net	ResNet101	Tversky	64.9	79.6	80.2	77.4
PSPNet	ResNet101	Dice	54.9	72.3	73.3	68.7
PSPNet	ResNet101	Tversky	53.4	75.0	79.1	62.2
DeepLabV3+	ResNet101	Dice	64.7	77.4	76.6	80.6
DeepLabV3+	ResNet101	Tversky	64.5	81.0	82.9	74.4

considerably higher mean IoU than the SegFormer-B0 variant, it can be assumed that the ability of transformers to capture global correlations is even more advantageous for the use case investigated here. The SegFormer is therefore confirmed as the architecture of choice.

A comparison of loss function impact across all variants shows the results generally follow the pattern identified in Table C.2: Dice loss increases IoU and P, while Tversky maximises F_2 and R. Generally, however, utilising Tversky L_T allows for a significant increase in the latter metrics with only minor losses to IoU. After IoU, R - and, consequently, F_2 - take precedence over P for leakage detection [52]. The SegFormer-B2 with Tversky loss is therefore selected as the winning AI model for this study, as it achieves maximum R and F_2 with the second highest IoU and an average P score.

Fig. C.2 shows the general structure of the SegFormer architecture adapted to this study's anomaly detection problem and training procedure. An input image of size $H \times W$ is divided into 4×4 pixel patches, each of which is converted into a linear embedding via 2D convolutional layer and fed into the first of four transformer blocks that make up the encoder. The transformers extract features hierarchically at up to $\frac{1}{32}$ resolution of the original image. The extracted features are passed to a multi-layer perceptron (MLP) decoder, which predicts a binary segmentation mask and returns probability values by applying a sigmoid function. The prediction is expanded to match the original resolution via upsampling.

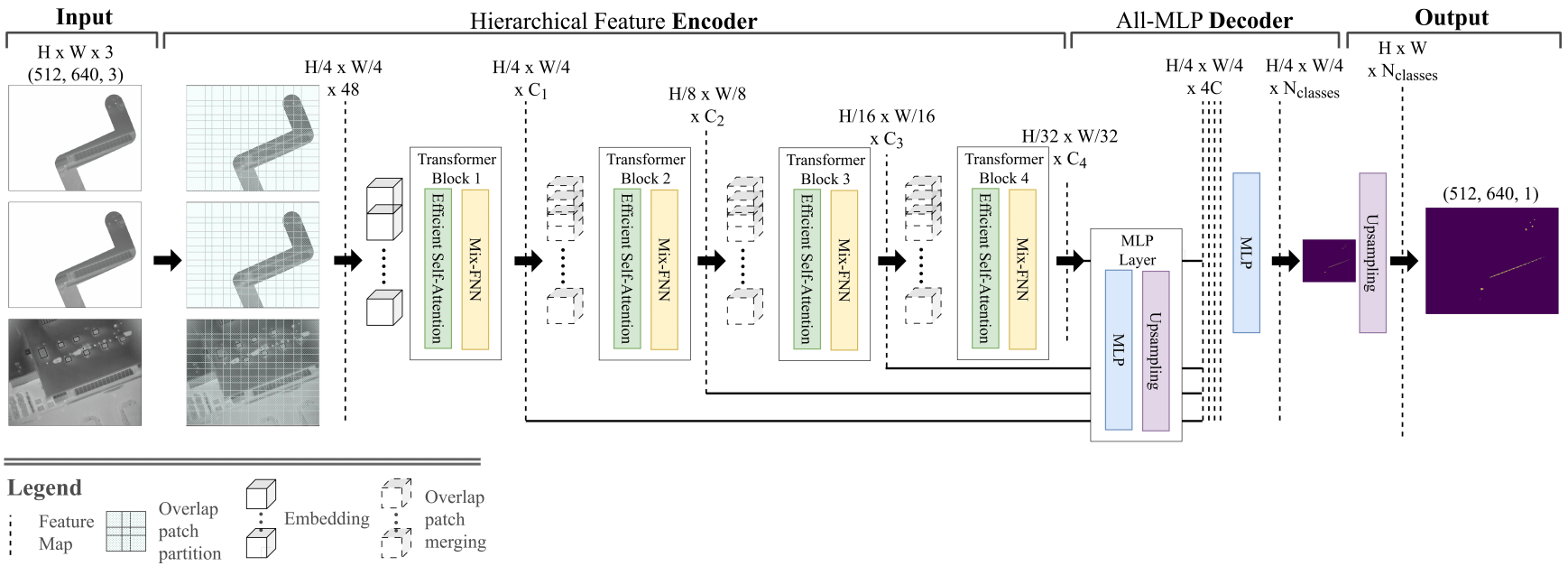


Figure C.2: SegFormer architecture adapted to the TIR anomaly detection problem, with every transformer block containing a feed-forward network (FFN) and the decoder including MLPs layers. Image based on Xie et al. [55].

C.3.3 Implementation

All DL models are implemented via the 'PyTorch' [36] and 'PyTorch-Lightning' [12] libraries. To ensure access to a wide range of model architectures, the 'Segmentation-Models-PyTorch (SMP)' [23] and 'Hugging-Face-Transformers' [54] toolboxes were used. The use of these libraries enables a higher degree of flexibility compared to frameworks that abstract more from the details of the underlying implementation. For data augmentation, Buslaev et al. [6]'s 'albumentation' library is used.

The code was implemented on the bwUniCluster2.0, a high-performance computing cluster operated by the Federal State of Baden-Wuerttemberg in Germany for university use. A single NVIDIA A100 [34] graphics processing unit (GPU) with 50 GB of graphics memory was used for all model trainings and pipeline runs. Given these hardware specifications, the winning SegFormer variant from Section C.3.2.4 took approximately 72 min to train. Energy requirements, measured via the 'perun' package [18], amounted to 0.258 kWh and 0.108 kgCO2e. This is about 1.3 times the requirement for U-Net (0.205 kWh) and 1.4 times the requirement for DeepLabV3+ (0.191 kWh) training.[7]

C.4 Evaluation and comparison

The SegFormer model described in Section C.3.2.4 is evaluated in a quantitative, qualitative, and holistic manner. This not only ensures a comprehensive assessment of AI performance for leak detection, but also enables a comparison with state-of-the-art CV algorithms from Vollmer et al. [52].

C.4.1 Quantitative evaluation

A wide range of semantic segmentation metrics are evaluated to cover all aspects of a thorough and quantitative assessment. These include recall (R), precision (P), intersection over union (IoU), F_2 score, detection rate (DR), and detection rate 30 (DR_{30}) to match those utilised by Vollmer et al. [52]. The last two are custom metrics, defining the proportion of anomalies identified out of all and those larger than 30 pixels respectively [52].

Table C.4 shows the results for the SegFormer applied to the validation and test splits of the manual dataset, coined evaluation dataset in [52]. Both a default threshold of 0.5 (*Th@0.5*) and a lower threshold of 0.1 (*Th@0.1*) are used for the analysis. These values are selected heuristically and based on well-performing ones from comparable implementations in literature. Given the inherent requirement for conservative handling to avoid sorting out true leak candidates [52], the binarisation should strive to minimise FNs. This means

[7] For reference, Gowda et al. [17] report an energy consumption of 79.5 kWh for training a DeepLabV3 on four NVIDIA V100 GPUs, illustrating how energy consumption can vary across setups and highlighting this study's comparatively resource-efficient configuration.

the selection must tend towards mid- and low-range values.[8] A similar observation is made by Alkan and Karasaka [1] in their study of different thresholds for binary semantic segmentation of remote sensing imagery. They find the best performance is achieved by 0.5, followed by 0.12, which guides the threshold selection in this study.

Table C.4: Quantitative results of the SegFormer compared to traditional CV algorithms from Vollmer et al. [52], evaluated on the manually annotated validation and test sets (see Table C.1). Results are colour-coded from grey (low) over white (mid) to green (high).

Method	Configuration	Validation						Test					
		IoU	F_2	R	P	DR	DR_{30}	IoU	F_2	R	P	DR	DR_{30}
SegFormer	Th@0.5	70.7	84.6	85.9	80.0	94.3	96.8	61.3	73.3	71.6	81.1	79.8	86.3
	Th@0.1	69.6	85.9	88.7	76.4	95.2	97.8	61.6	75.2	74.6	78.1	83.2	89.2
THT	with VC	59.8	77.5	79.5	70.7	88.6	88.2	47.8	72.8	79.5	54.5	79.8	85.3
	without VC	54.0	65.3	62.4	80.1	78.1	79.6	37.0	50.3	48.1	61.6	37.8	42.2
SM	MaxIoU	60.3	80.0	83.4	68.5	88.6	92.5	55.0	67.4	65.2	77.7	72.3	80.4
	MaxIoU@85	57.1	81.2	88.1	61.8	94.3	94.6	53.3	67.6	66.4	73.0	75.6	82.4
	MaxIoU@90	52.5	80.3	90.2	55.6	93.3	95.7	46.0	71.8	79.2	52.3	85.7	92.2
LT	MaxIoU	51.6	62.7	59.6	79.4	71.4	79.6	35.2	43.4	39.0	78.1	63.0	67.6

To allow for a direct comparison, Table C.4 also includes results of the best performing variants among the classical algorithms THT, SM, and LT [52]. This highlights the DL model's aptitude at producing the desired segmentation behaviour. The Segformer model scores by far the highest IoUs on both validation and test sets, beating the previous best by around 9.7 %pt and 6.5 %pt respectively. It particularly excels at achieving a high P without detriment to R and DR, something the traditional methods struggle with. On the test split, for instance, the SegFormer *Th@0.1* surpasses THT with VC's P score by 23.7 %pt with a comparatively small R loss of 4.9 %pt whilst even achieving an increase in DR and DR_{30} of 3.4 %pt and 3.9 %pt respectively. In general, both SegFormer configurations almost consistently outperform the classical CV state-of-the-art. Setting the threshold lower generally produces slightly better results, though it considerably increases the number of identified anomalies, including false alarms.

C.4.2 Qualitative evaluation

A qualitative evaluation, shown in Fig. C.3, is performed on the same exemplary images used by Vollmer et al. [52]. This provides means to comparatively assess the SegFormer's ability of handling common scenarios in leakage detection. The table includes results from the THT algorithm as the winning method in previous work [52].

Overall, this qualitative comparison shows the SegFormer model to deliver more robust results. In imagery containing both an exceptionally conspicuous leak as well as smaller anomalies (Fig. C.3.1), the model identifies all hot-spots without classifying as large a pixel

[8] While high thresholds may ensure low FP rates, this comes at the detriment of FN.

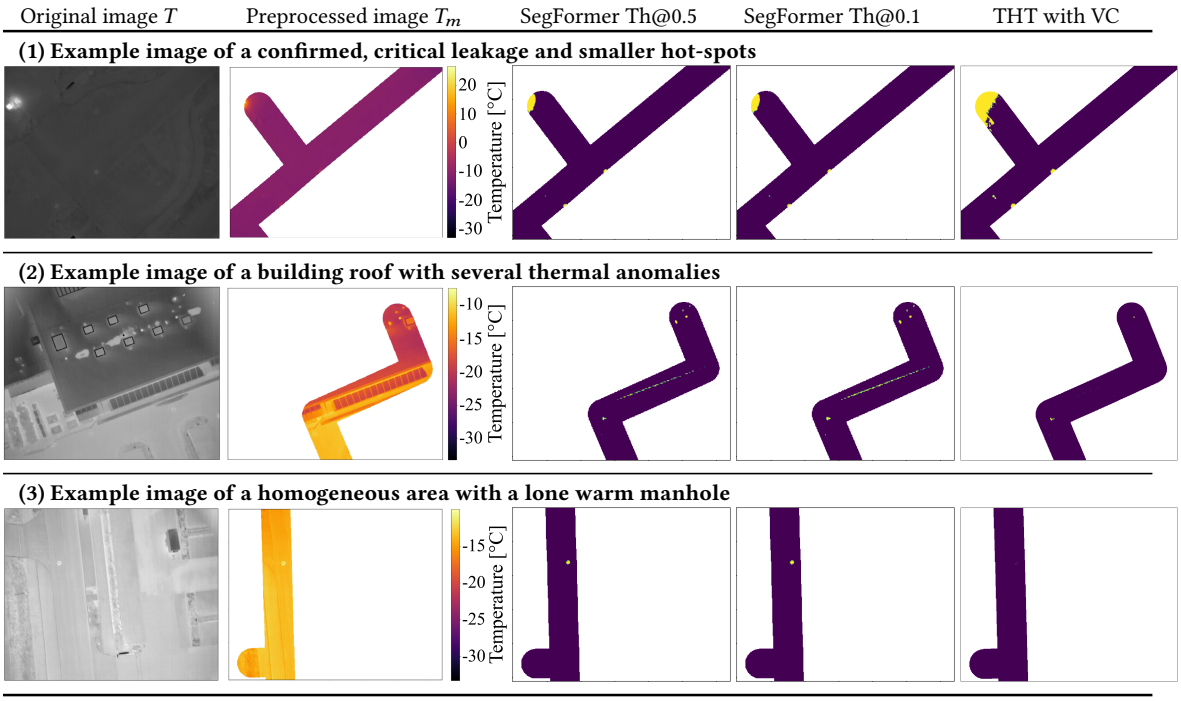


Figure C.3: Segmentation masks predicted by the SegFormer for three example scenarios, with [52]'s winning THT for comparison.

area as THT. In scenes where a uniform threshold does not allow for sufficiently accurate differentiation and THT struggles (Figs. C.3.2 and C.3.3), the SegFormer model is capable of discerning relevant anomalies that were previously missed. In addition, the generated segmentation masks are precise and comprehensible for a human observer.

C.4.3 Model explanation

Further insights into the DL model can be obtained through a comparatively new branch of research: xAI. Motivated by the inherent black-box nature of AI models, this field aims to provide explanations of why models behave in a certain way and what guides the decision-making behind their predictions [20]. Of the large variety of existing xAI methods, explanations are most often created through visualisation techniques, commonly class activation map (CAM)-based methods [25]. One of the most influential of these, Grad-CAM, visualises important regions in an image for a specific class by analysing gradients in the last convolutional layer, enabling it to be model-agnostic [42]. Although most techniques for explaining image-based AI models focus on classification tasks [15], this method was adapted to semantic segmentation via implementations such as the Seg-Grad-CAM [49].

While both Sections C.4.1 and C.4.2 have highlighted the SegFormer's[9] ability to excel at anomaly detection, xAI can help check if the model behaves as intended. To this end, Gildenblat and contributors [14]'s PyTorch Grad-CAM toolbox is adapted to work with a segmentation model using non-integer TIRs. Post-processing is applied to remove unwanted artefacts in the explanations.[10] Fig. C.4 shows two exemplary image inputs, associated annotation masks, resulting SegFormer predictions, and xAI explanations. These last are visualised as heat maps, with model interest increasing from blue to red.

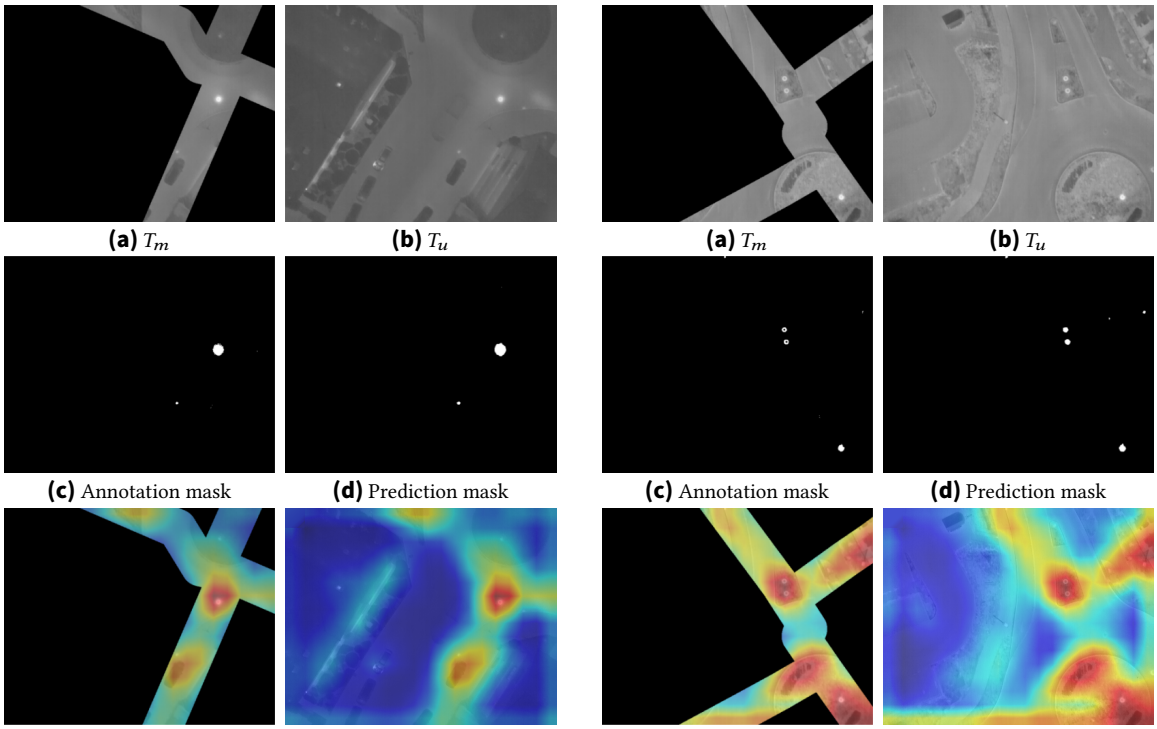


Figure C.4: Seg-Grad-CAM explanations for exemplary TIR input images.

The first image C.4.1 features an example in which prediction and annotation masks are equivalent. The segmentation Grad-CAM outputs in C.4.1.e) and f) showcase how the choice of three-channel input constellations, consisting of duplicated C.4.1.a) and single C.4.1.b), helps to focus the model on the relevant image areas above and around DHS pipelines. While the model generally highlights all warmer regions, only those within the mask are attributed with high importance, gaining the associated pixels a place in the prediction output. The input channel selection is therefore confirmed as having the intended effect on model behaviour.

[9] To exemplify, the 0.5 threshold is utilised to balance anomaly amounts while ensuring high quantitive scores.

[10] The explanations include an excess mask-location-dependent highlight in the upper left corner due to the definition of masked pixels as negative values instead of 'None'.

The second example C.4.2 exhibits an output C.4.2.d) that differs from the given ground truth C.4.2.c). Specifically, more pixels are predicted to be anomalous than are defined as such in the annotation mask. The Grad-CAM explanation reveals the reason for this: The model attributes just as much attention to the street lamp at the upper right edge and manholes in the image centre as it does the manhole in the bottom right corner. In contrast, the annotation mask only includes the warmer pixels along the centre covers' edges and excludes the street lamp, as its temperature is lower in comparison. As the explanations in both examples show, the model's focus areas always extend beyond the anomalies themselves, indicating the model takes anomalies' local surroundings into account for its decision-making process. This allows for local maxima to be identified regardless of their absolute temperature and explains the model's nuanced segmentation behaviour. However, said model trait also highlights the necessity for a post-detection anomaly categorisation so that such false alarms may be sorted out [50].

C.4.4 Evaluation of the analysis pipeline

A final evaluation of the model integrated into the image analysis pipeline helps assess the SegFormer's holistic aptitude for leakage detection. Aside from masking the inferred results, this includes a post-processing classification of all identified anomalies according to the temperature difference to their surroundings: uncritical ($\Delta T < 5°C$), moderate ($5°C \leq \Delta T < 10°C$), pronounced ($10°C \leq \Delta T < 15°C$), and critical ($15°C \leq \Delta T$) [52]. As in Vollmer et al. [52], the datasets *MU2* and *KA1* are analysed. All anomalies classified as – at minimum – moderate were checked and categorised manually. To ensure an unbiased evaluation for the *MU2* dataset, no images from *MU2* were included in the training or validation splits of the generated dataset.

Table C.5 summarises the SegFormer results and compares them to the traditional THT method. The SegFormer identifies more anomalies across both datasets, showing it operates more conservatively, and thus favourably, for leakage detection. At the same time, the average anomaly area is generally considerably smaller than that of THT, confirming Section C.4.2's observation of the SegFormer's capability to draw more nuanced contours around anomalies.

Regarding *MU2*, the large, critical leakage is detected equally reliably by the DL model as its traditional algorithmic counterpart. The number of relevant anomalies identified by the SegFormer is considerably higher, again demonstrating a more conservative approach. A manual categorisation shows that these mostly pertain to the category "other", where an in-depth analysis reveals their main source as hot-spots on building rooftops (caused by, e.g., chimneys) with a commonly low absolute temperature. The difference in detected manhole covers can be ascribed to some no longer falling below the 5 °C threshold when the cold inner area is defined as part of the anomaly. For the *KA1* dataset, the overall picture is more homogeneous. While specific assignment to moderate and pronounced categories

Table C.5: Leakage detection pipeline evaluation results for the datasets *MU1* and *KA1*.

		MU2		***KA1***	
		SegFormer Th@0.5	THT with VC	SegFormer Th@0.5	THT with VC
# of anomalies		1112	709	1038	647
average anomaly area		134.8	195.4	193.0	202.7
Classified by ΔT	uncritical	942	561	962	567
	moderate	134	105	74	62
	pronounced	11	18	6	18
	critical	25	25	0	0
# of relevant anomalies		170	148	76	80
Classified by type (manually)	leakage	19	20	0	0
	manhole	61	82	52	51
	car	90	46	10	10
	other			14	19

differs,[11] the more significant number of relevant anomalies and type classifications is very similar. In particular, the number of anomalies caused by warm vehicles and manholes is almost or exactly identical, highlighting a comparable performance between AI and classical CV methodologies.

As a final analysis, Table C.6 compares both methods in terms of required resources, specifically total and individual step durations. These are derived from runs using the hardware described in Section C.3.3 and multiprocessing with batch sizes of 16.[12] Total values should be seen as reference points, as deviations can still occur between pipeline runs.[13]

As expected, the main focus of the comparison lies on the anomaly detection step, where the methodology deviates. This is highlighted by the equal times per image or anomaly for Steps 1, 3, and 4. Although one might expect the considerably less complex THT method to outperform the DL model, Table C.6 paints a different picture. Inference using the SegFormer-B2 model is almost as fast as the implementation of the traditional CV algorithm. This corroborates the high performance and efficiency reported by Xie et al. [55] for their lightweight transformer architecture. However, the here achieved 4.8 FPS for anomaly detection is nowhere near their recorded 24.5 FPS,[14] revealing room for improvement and the possibility of surpassing THT run times with code optimisations.

[11] This may be attributed to the fact that the SegFormer generally defines anomaly boundaries closer around hot-spots, yielding slightly warmer surroundings and thus somewhat smaller temperature differences.

[12] It should be noted that the pipeline was not designed with a time constraint in mind and that runs depend greatly on available hardware and options for parallelisation.

[13] For example, the standard deviation for the dataset statistics calculation step across 9 consecutive runs is 10.4 s, though this drops to 2.1 s when excluding the initial run.

[14] This value is given for the SegFormer-B2, for single-scale inference and a batch size of 16 on the ADE20K dataset [55].

Table C.6: Pipeline run times exemplified on *KA1*, with 496 images and anomaly amounts listed in Table C.5.

Method	Calculation	Pipeline step durations [s]				Total duration [s]
		Dataset statistics	Anomaly detection	Anomaly extraction	Anomaly classification	
THT (with VC)	total	22.31	95.13	40.40	7.10	165.57
	per image / anomaly	0.05	0.19	0.06	0.01	0.33 / 0.26
SegFormer (Th@0.5)	total	22.47	103.65	60.98	9.15	197.17
	per image / anomaly	0.05	0.21	0.06	0.01	0.40 / 0.19

C.5 Conclusion

This study greatly advances UAS and TIR-based leak detection by tackling the critical step of finding thermal anomalies via a DL semantic segmentation model. It is among the first to apply DL to this energy-related use case while providing extensive insights into the utilised methodology. A novel, multi-stage training procedure enabled the development of a high-performing SegFormer model, overcoming the one of the biggest challenges in UAS-based semantic segmentation: limited annotated data. This procedure is easily adaptable to other use cases and may therefore function as a guide for similar implementations of domain shift. Compared to traditional state-of-the-art CV algorithms, the DL model is found to offer a high degree of flexibility and aptitude for achieving the desired segmentation behaviour. Both quantitative and qualitative evaluations show that the SegFormer considerably improves upon results from the best performing classical equivalent, a conclusion supported by the holistic assessment. Smaller anomalies are reliably found, the generated segmentation masks are precise, and the number of detected, yet implausible anomalies is significantly lower compared to the traditional algorithms from Vollmer et al. [52]. The xAI analysis sheds further light on model characteristics, showcasing how it focuses on local maxima by considering anomalies' immediate surroundings and how the choice of combined masked and unmasked inputs has the desired effect of attributing more importance to areas around the DHS. Given all these afore-mentioned characteristics, this study's SegFormer model is found to surpass existing traditional CV methods, thereby establishing a new state-of-the-art in anomaly detection for TIR-based DHS leak detection.

Naturally, this study is subject to some limitations. Though diverse, the datasets used in this study are comparatively small and include only two German cities. As no other research group from Section C.2 has made their data publicly available, the model's data foundation could not be enhanced with imagery from other regions. While the necessity for DHSs and, in turn, these forms of monitoring approaches is greatest in colder countries similar to Germany, it is unclear how well the model will generalise across varying DHSs, sensor types, and flight heights. Additionally, adjustments to the model (such as the inclusion of this kind of new data) require retraining, which presupposes the availability of appropriate computing resources. The corresponding hardware is often cost- and energy-intensive and may not be generally available.

The improvements achieved by the DL model for this vital step in automatic TIR-based DHS leak detection highlight various opportunities for future studies. The inclusion of datasets from various different regions can help test and train the model's generalisability and ensure robustness towards more diverse urban landscapes. Coupled with its integration into an active learning loop [40], model performance could be further improved with a comparably low effort. The ultimate goal would be the creation of an expansive dataset that includes multitudes of confirmed leaks, which would allow the model to be tailored to the more exclusive task of leak detection instead of thermal anomaly detection and subsequent false alarm removal. As this is contingent upon a wide-scale sharing of (annotated) data, an interim approach could be the combination of given datasets with publicly available ones, such as Vollmer et al. [51], as a first step towards implementations that eliminate the need for downstream classification.

On a broader scale, enhancing the analysis with a temporal assessment and automatic comparison between TIR data may help transform the method into a regular monitoring approach. Given the ever-growing interest in UAS-based urban monitoring, existing multi-sensor applications could be expanded to include the required image acquisition [4]. Coupled with code optimisation to enable faster run times and potentially real-time implementation[55], this study's SegFormer-based automatic TIR analysis for DHS leak detection could become an integral part of UAS inspections of our future smart cities.

Acknowledgments

The datasets were acquired in collaboration with the Air Bavarian GmbH and Munich's and Karlsruhe's municipal utilities companies. The authors acknowledge support by the state of Baden-Wuerttemberg through bwHPC. This work is supported by funding from the European Union through the AI4EOSC project (Horizon Europe) under Grant number 101058593.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT author statement

Elena Vollmer: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. **Julian Ruck**: Conceptualisation, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – review & editing. **Rebekka**

Volk: Funding acquisition, Writing – review & editing, Supervision. **Frank Schultmann**: Writing – review & editing, Supervision.

Data availability statement

All data, code, and configurations used in this work will be made available with this publication via Zenodo [39] and GitHub (`https://www.github.com/emvollmer/TASeg`).

Appendices

Appendix I. Temperature distributions

This appendix compares the temperature distributions between the different datasets as well as splits used for model training. While there are some fluctuations between the individual distributions shown in Fig. C.5, Fig. C.6 highlights how the majority of the data is similarly positioned between −15 °C to −5 °C.

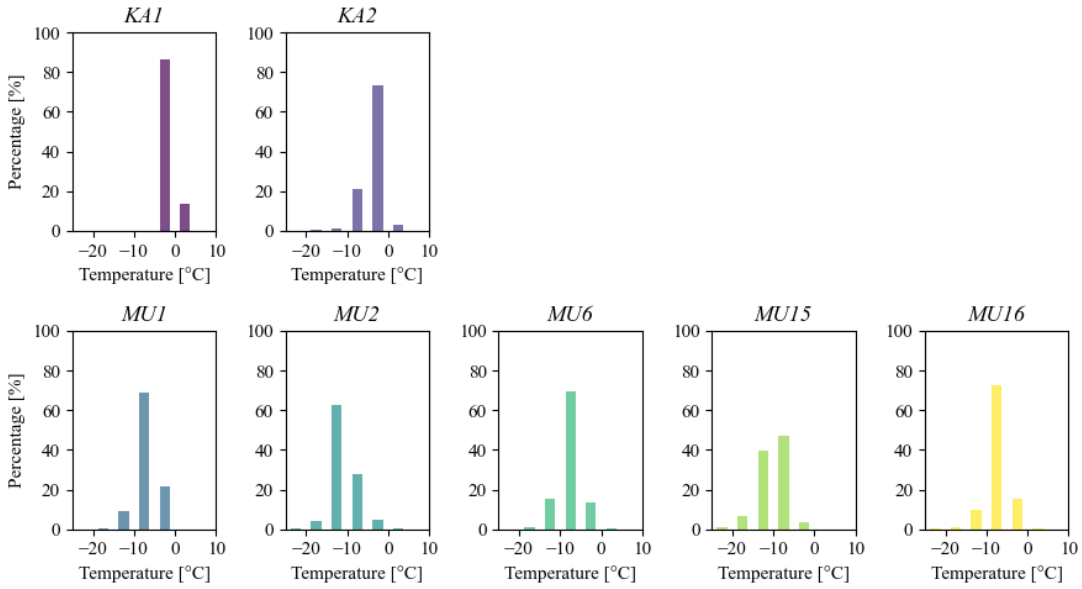


Figure C.5: Histograms of temperature distributions per individual dataset.

Once combined into train, validation, and test splits for model training, the distributions become very similar, as highlighted in Fig. C.7.

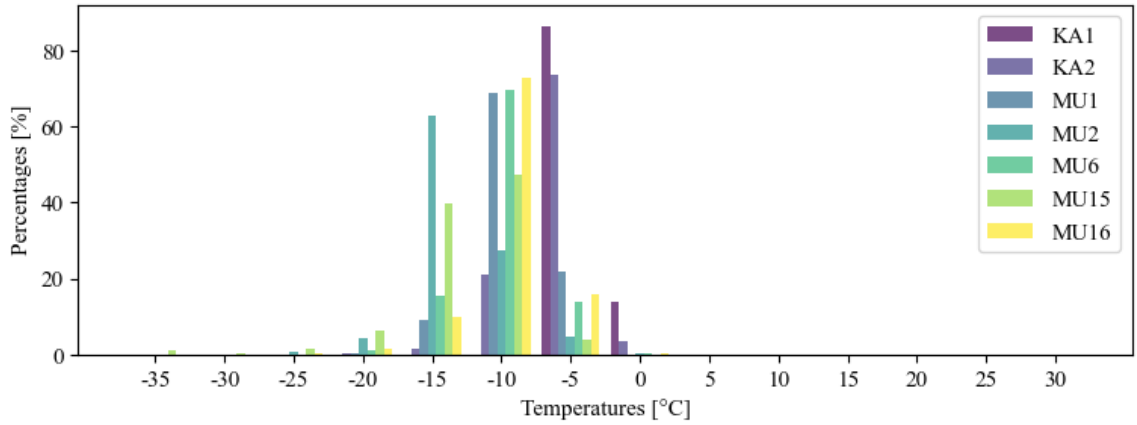


Figure C.6: Histograms of temperature distributions per individual dataset, grouped in one plot.

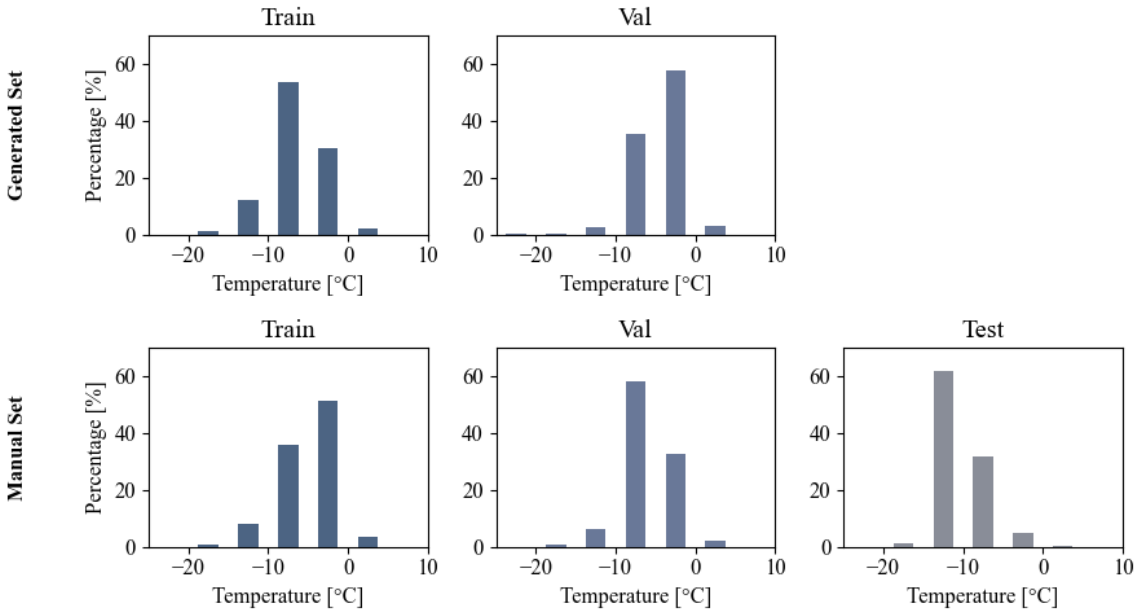


Figure C.7: Histograms of temperature distributions in train, validation, and test splits for model training.

Appendix II. Ablation study of input channel configurations

This appendix compares the impact of channel definitions on model performance, specifically the contributions of T_m and T_u. Although the focus lies on the masked images and thus the areas above and around the DHS, the entire unmasked image may provide additional useful context information. For instance, the area leftover after masking may only constitute a tiny portion of the original image (e.g. a piece of road at the image edge), which can be easily misclassified without context information is the given surface temperature is simply higher due to higher ambient temperatures or material. To assess the validity of this assumption, various combinations are compared via the SegFormer-B2 model with Tversky loss.

While conventional DL architectures for image analysis expect three channel inputs owing to that being the standard format of RGB imagery, models can be adapted to accept other channel counts by including an initial convolutional layer [50]. Through this method, various configurations can be tested to identify the best-suited inputs. These encompass: 1. single-channel inputs (T_u) as a reference, 2. single-channel inputs (T_m) as a baseline for masked data on its own, 3. two-channel inputs (T_m, T_u) to assess the impact of including context information, 4. three-channel inputs (T_m, T_m, T_u) to assess the performance when combining and weighting masked and unmasked data.

As the results in Table C.7 show, the additional inclusion of context information through T_u improves the performance of the model with respect to all metrics. Doubling the T_m layer and thereby emphasising the masked data more strongly, results in a further performance increase throughout. For this reason, the inputs used in this study are three-channel (T_m, T_m, T_u).

Table C.7: Comparison of model variants with on the validation split. Results are colour-coded from white (low) to green (high).

Model	Encoder	Loss	Channels	IoU	F_2	R	P
SegFormer	MiT-B2	Tversky	(T_m, T_m, T_u)	70.2	84.5	85.9	79.4
SegFormer	MiT-B2	Tversky	(T_m, T_u)	69.4	83.8	85.1	79.0
SegFormer	MiT-B2	Tversky	(T_m)	67.4	82.2	83.3	78.1
SegFormer	MiT-B2	Tversky	(T_u)	43.9	66.9	71.6	53.1

Appendix III. Impact of the multi-phase training procedure

Different experiments were conducted to assess the impact of various aspects of the developed mutli-phase training procedure, based on an exemplary Segformer configuration. Table C.8 breaks down the performance for each training phase. The results clearly highlight how each consecutive step is able to improve all evaluated metrics.

Table C.9 investigates the influence of the generated dataset on performance. For this, the standard training procedure with 15, 35, and 60 epochs per respective phase is compared to only using the manual dataset, both for the standard duration of phase 3 and the total

Table C.8: Performance comparison after each phase, evaluated on the validation split. Results are colour-coded from white (low) to green (high).

Model	Encoder	Loss	Phase	Epochs	Dataset	Encoder	IoU	F_2	R	P
SegFormer	MiT-B2	Tversky	1	15	generated	frozen	40.4	62.6	66.5	50.7
SegFormer	MiT-B2	Tversky	2	35	generated	unfrozen	62.7	80.7	83.4	72.0
SegFormer	MiT-B2	Tversky	3	60	manual	unfrozen	70.2	84.5	85.9	79.4

epoch count. The results bring to light several influencing factors of both datasets and procedure. Firstly, the model requires more time to focus on the new target domain when using only the manual dataset. The model trained for 60 epochs solely on that set is defined by an extremely high R but low IoU, indicating classical characteristics of early training such as over-segmentation and poor boundary definition. In comparison, a model trained for only 50 epochs on the generated set (see Table C.8) is already better adapted to the UAS-based TIRs.

After 110 epochs, the manual dataset-based model has a similarly refined focus and outperforms phase 2 results. Overall, however, the multi-step procedure, which combines the use of both datasets, still surpasses utilising only the high-quality manual one on all counts.

Table C.9: Performance comparison with and without generated dataset, evaluated on the validation split. Results are colour-coded from white (low) to green (high).

Model	Encoder	Loss	Phase(s)	Epochs	IoU	F_2	R	P
SegFormer	MiT-B2	Tversky	3	60	49.1	79.0	91.1	51.6
SegFormer	MiT-B2	Tversky	3	110	68.1	83.8	85.8	76.8
SegFormer	MiT-B2	Tversky	1, 2, and 3	15, 35, and 60	70.2	84.5	85.9	79.4

Bibliography

[1] Alkan, D. and Karasaka, L. (2023). "Segmentation of Landsat-8 Images for Burned Area Detection with Deep Learning". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLVIII-M-1-2023, pp. 455–461. DOI: 10.5194/isprs-archives-XLVIII-M-1-2023-455-2023.

[2] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Hauptbericht 2022 [Main report 2022]*. Report. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht (visited on 19 Dec. 2024).

[3] Axelsson, S. (1988). "Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 26(5), pp. 686–692. DOI: 10.1109/36.7695.

[4] Bayomi, N. and Fernandez, J. E. (2023). "Eyes in the Sky: Drones Applications in the Built Environment under Climate Change Challenges". In: *Drones* 7(10), p. 637. DOI: 10.3390/drones7100637.

[5] Berg, A., Ahlberg, J., and Felsberg, M. (2016). "Enhanced analysis of thermographic images for monitoring of district heat pipe networks". In: *Pattern Recognition Letters* 83, pp. 215–223. DOI: 10.1016/j.patrec.2016.07.002.

[6] Buslaev, A., Iglovikov, V. I., Khvedchenya, E., Parinov, A., Druzhinin, M., and Kalinin, A. A. (2020). "Albumentations: Fast and Flexible Image Augmentations". In: *Information* 11(2). DOI: 10.3390/info11020125.

[7] Chen, L.-C., Papandreou, G., Schroff, F., and Adam, H. (2017). *Rethinking Atrous Convolution for Semantic Image Segmentation*. DOI: 10.48550/ARXIV.1706.05587. arXiv: 1706.05587.

[8] Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., and Adam, H. (2018). *Encoder-Decoder with Atrous Separable Convolution for Semantic Image Segmentation*. DOI: 10.48550/arXiv.1802.02611.

[9] Cheng, J., Deng, C., Su, Y., An, Z., and Wang, Q. (2024). "Methods and datasets on semantic segmentation for Unmanned Aerial Vehicle remote sensing images: A review". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 211, pp. 1–34. DOI: 10.1016/j.isprsjprs.2024.03.012.

[10] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "ImageNet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 248–255. DOI: 10.1109/cvpr.2009.5206848.

[11] Duque-Arias, D., Velasco-Forero, S., Deschaud, J.-E., Goulette, F., Serna, A., Decencière, E., and Marcotegui, B. (2021). "On Power Jaccard Losses for Semantic Segmentation". In: *International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISAPP)*. Vol. 5. SCITEPRESS - Science and Technology Publications, pp. 561–568. DOI: 10.5220/0010304005610568.

[12] Falcon, W. and The PyTorch Lightning team (2019). *PyTorch Lightning*. Zenodo. Version 1.8.6. DOI: 10.5281/zenodo.7469930.

[13] Friman, O., Follo, P., Ahlberg, J., and Sjokvist, S. (2014). "Methods for Large-Scale Monitoring of District Heating Systems Using Airborne Thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 52(8), pp. 5175–5182. DOI: 10.1109/TGRS.2013.2287238.

[14] Gildenblat, J. and contributors (2021). *PyTorch library for CAM methods*. URL: https://github.com/jacobgil/pytorch-grad-cam.

[15] Gipiškis, R., Tsai, C.-W., and Kurasova, O. (2024). "Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey". In: *ICT Express* 10(6), pp. 1331–1354. DOI: 10.1016/j.icte.2024.09.008.

[16] Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Adaptive computation and machine learning. Cambridge, United States: MIT Press. ISBN: 978-0-262-03561-3. URL: https://www.deeplearningbook.org/.

[17] Gowda, S. N., Hao, X., Li, G., Gowda, S. N., Jin, X., and Sevilla-Lara, L. (2024). *Watt For What: Rethinking Deep Learning's Energy-Performance Relationship*. DOI: 10.48550/arXiv.2310.06522. arXiv: 2310.06522.

[18] Gutiérrez Hermosillo Muriedas, J. P., Flügel, K., Debus, C., Obermaier, H., Streit, A., and Götz, M. (2023). "perun: Benchmarking Energy Consumption of High-Performance Computing Applications". In: *Euro-Par 2023: Parallel Processing*. Ed. by Cano, J., Dikaiakos, M. D., Papadopoulos, G. A., Pericàs, M., and Sakellariou, R. Cham, Switzerland: Springer Nature Switzerland, pp. 17–31.

[19] He, Y., Deng, B., Wang, H., and Cheng, L. (2021). "Infrared machine vision and infrared thermography with deep learning: A review". In: *Infrared Physics & Technology* 116, p. 103754. DOI: 10.1016/j.infrared.2021.103754.

[20] Holzinger, A., Saranti, A., Molnar, C., Biecek, P., and Samek, W. (2022). "Explainable AI Methods - A Brief Overview". In: *xxAI - Beyond Explainable AI: International Workshop*. Ed. by Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., and Samek, W. Cham, Switzerland: Springer, pp. 13–38. DOI: 10.1007/978-3-031-04083-2_2.

[21] Hossain, K., Villebro, F., and Forchhammer, S. (2019). "Leakage Detection in District Heating Systems Using UAV IR Images: Comparing Convolutional Neural Network and ML Classifiers". In: *27th European Signal Processing Conference (EUSIPCO)*. A Coruña, Spain: European Association for Signal Processing (EURASIP). DOI: 10.23919/EUSIPCO45326.2019.

[22] Hossain, K., Villebro, F., and Forchhammer, S. (2020). "UAV Image Analysis for Leakage Detection in District Heating Systems using Machine Learning". In: *Pattern Recognition Letters* 140, pp. 158–164. DOI: 10.1016/j.patrec.2020.05.024.

[23] Iakubovskii, P. (2019). *Segmentation Models Pytorch*. URL: https://github.com/qubvel/segmentation_models.pytorch (visited on 20 Aug. 2024).

[24] International Energy Agency (IEA) (2023). *World Energy Outlook 2023*. Technical report. Paris, France: IEA. URL: https://www.iea.org/reports/world-energy-outlook-2023 (visited on 19 Dec. 2024).

[25] Islam, M. R., Ahmed, M. U., Barua, S., and Begum, S. (2022). "A Systematic Review of Explainable Artificial Intelligence in Terms of Different Application Domains and Tasks". In: *Applied Sciences* 12(3), p. 1353. DOI: 10.3390/app12031353.

[26] Jadon, S. (2020). "A survey of loss functions for semantic segmentation". In: *IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. IEEE, pp. 1–7. DOI: 10.1109/cibcb48159.2020.9277638.

[27] Johnson, J. M. and Khoshgoftaar, T. M. (2019). "Survey on deep learning with class imbalance". In: *Journal of Big Data* 6(1), p. 27. DOI: 10.1186/s40537-019-0192-5.

[28] Li, C., Xia, W., Yan, Y., Luo, B., and Tang, J. (2021). "Segmenting Objects in Day and Night: Edge-Conditioned CNN for Thermal Image Semantic Segmentation". In: *IEEE Trans. Neural Netw. Learn. Syst.* 32(7), pp. 3069–3082. DOI: 10.1109/TNNLS.2020.3009373.

[29] Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., and Guo, B. (2021). "Swin Transformer: Hierarchical Vision Transformer using Shifted Windows". In: *IEEE International Conference on Computer Vision (ICCV)*. Montreal, Canada (virtual): IEEE, pp. 9992–10002. DOI: 10.1109/iccv48922.2021.00986.

[30] Ljungberg, S.-A. and Rosengren, M. (1988). "Aerial and Mobile Thermography to Assess Damages and Energy Losses from Buildings and District Heating Networks - Operational Advantages and Limitations". In: *XVIth ISPRS Congress, Technical Commission VII: Interpretation of Photographic and Remote Sensing Data*. Ed. by Murai, S. Vol. XXVII, Part B7. Kyoto, Japan: ISPRS, pp. 348–359. URL: https://www.isprs.org/proceedings/XXVII/congress/part7/348_XXVII-part7.pdf.

[31] Long, J., Shelhamer, E., and Darrell, T. (2015). "Fully convolutional networks for semantic segmentation". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Boston, Unites States: IEEE, pp. 3431–3440. DOI: 10.1109/cvpr.2015.7298965.

[32] Loshchilov, I. and Hutter, F. (2017). *Decoupled Weight Decay Regularization*. DOI: 10.48550/ARXIV.1711.05101. arXiv: 1711.05101.

[33] Nogueira, K., Faita-Pinheiro, M. M., Marques Ramos, A. P., Gonçalves, W. N., Junior, J. M., and Dos Santos, J. A. (2024). "Prototypical Contrastive Network for Imbalanced Aerial Image Segmentation". In: *IEEE Winter Conference on Applications of Computer Vision (WACV)*. Waikoloa, United States: IEEE, pp. 8351–8361. DOI: 10.1109/WACV57701.2024.00818.

[34] NVIDIA Corporation (2022). *A100 Tensor Core GPU*. accessed 19 December 2024. URL: https://www.nvidia.com/en-us/data-center/a100/ (visited on 2 Feb. 2024).

[35] Pang, G., Shen, C., Cao, L., and Hengel, A. V. D. (2022). "Deep Learning for Anomaly Detection: A Review". In: *ACM Computing Surveys* 54(2), pp. 1–38. DOI: 10.1145/3439950.

[36] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Curran Associates, Inc., pp. 8024–8035. DOI: 10.48550/arXiv.1912.01703.

[37] Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Ed. by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. Cham, Switzerland: Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[38] Ruck, J., Vollmer, E., Volk, R., and Vogl, M. (2024). *Detecting District Heating Leaks in Thermal Imagery: Comparison of Anomaly Detection Methods - Source Code and Datasets*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.11085776.

[39] Ruck, J., Vollmer, E., Volk, R., and Vogl, M. (2025). *Thermal Anomaly Segmentation Dataset - Thermal UAS-based Images from Germany with Annotations for Semantic Segmentation Model Training*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.14287864.

[40] Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., and Ryo, M. (2023). "Ten deep learning techniques to address small data problems with remote sensing". In: *Int. J. Appl. Earth Obs. Geoinf.* 125, p. 103569. DOI: 10.1016/j.jag.2023.103569.

[41] Salehi, S. S. M., Erdogmus, D., and Gholipour, A. (2017). "Tversky Loss Function for Image Segmentation Using 3D Fully Convolutional Deep Networks". In: *Mach. Learn. Med. Imaging*. Springer International Publishing, pp. 379–387. DOI: 10.1007/978-3-319-67389-9_44.

[42] Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., and Batra, D. (2020). "Grad-CAM: Visual Explanations from Deep Networks via Gradient-based Localization". In: *International Journal of Computer Vision* 128(2), pp. 336–359. DOI: 10.1007/s11263-019-01228-7.

[43] Siddique, M. F., Ahmad, Z., and and, J.-M. K. (2023). "Pipeline leak diagnosis based on leak-augmented scalograms and deep learning". In: *Engineering Applications of Computational Fluid Mechanics* 17(1), p. 2225577. DOI: 10.1080/19942060.2023.2225577.

[44] Sledz, A. and Heipke, C. (2021). "Thermal Anomaly Detection Based on Saliency Analysis from Multimodal Imaging Sources". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-1-2021, pp. 55–64. DOI: 10.5194/isprs-annals-V-1-2021-55-2021.

[45] Sledz, A., Unger, J., and Heipke, C. (2020). "UAV-based Thermal Anomaly Detection for Distributed Heating Networks". In: *ISPRS - International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XLIII-B1-2020, pp. 499–505. DOI: 10.5194/isprs-archives-XLIII-B1-2020-499-2020.

[46] Sudre, C. H., Li, W., Vercauteren, T., Ourselin, S., and Cardoso, M. J. (2017). "Generalised Dice Overlap as a Deep Learning Loss Function for Highly Unbalanced Segmentations". In: *Deep Learn. Med. Image Anal. Multimodal Learn. Clin. Decis. Support*. Springer International Publishing, pp. 240–248. DOI: 10.1007/978-3-319-67558-9_28.

[47] United Nations Environment Programme (UNEP) (2024). *Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector*. Tech. rep. Nairobi, Kenya: UNEP, Global Alliance for Building and Construction (GlobalABC). DOI: 10.59117/20.500.11822/45095.

[48] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). "Attention Is All You Need". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Ed. by Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R. Vol. 30. Long Beach, United States: Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[49] Vinogradova, K., Dibrov, A., and Myers, G. (2020). "Towards Interpretable Semantic Segmentation via Gradient-weighted Class Activation Mapping". In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34(10), pp. 13943–13944. DOI: 10.1609/aaai.v34i10.7244.

[50] Vollmer, E., Benz, M., Kahn, J., Klug, L., Volk, R., Schultmann, F., and Götz, M. (2025). "Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18, pp. 6206–6216. DOI: 10.1109/JSTARS.2025.3537330.

[51] Vollmer, E., König, S., Horstmann, V., Klug, L., Kahn, J., Volk, R., and Vogl, M. (2025). *Thermal Urban Feature Segmentation - Multispectral (RGB + Thermal) UAS-based images from Germany with annotations*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.10814413.

[52] Vollmer, E., Ruck, J., Volk, R., and Schultmann, F. (2024). "Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods". In: *Automation in Construction* 168, p. 105709. DOI: 10.1016/j.autcon.2024.105709.

[53] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[54] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. von, Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T. L., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. M. (2020). "Transformers: State-of-the-Art Natural Language Processing". In: *Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, pp. 38–45. DOI: 10.48550/arXiv.1910.03771.

[55] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J. M., and Luo, P. (2021). "SegFormer: Simple and Efficient Design for Semantic Segmentation with Transformers". In: *Advances in Neural Information Processing Systems (NeurIPS)*. Vol. 34. Vancouver, Canada (virtual): Curran Associates, Inc., pp. 12077–12090. URL: 10.48550/arXiv.2105.15203.

[56] Xu, Y., Wang, X., Zhong, Y., and Zhang, L. (2016). "Thermal anomaly detection based on saliency computation for district heating system". In: *IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. Beijing, China: IEEE, pp. 681–684. DOI: 10.1109/IGARSS.2016.7729171.

[57] El-Zahab, S. and Zayed, T. (2019). "Leak detection in water distribution networks: an introductory overview". In: *Smart Water* 4(1), p. 5. DOI: 10.1186/s40713-019-0017-x.

[58] Zhao, H., Shi, J., Qi, X., Wang, X., and Jia, J. (2017). "Pyramid Scene Parsing Network". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. Honolulu, Unites States: IEEE, pp. 6230–6239. DOI: 10.1109/cvpr.2017.660.

[59] Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., and Zhang, L. (2019). "Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions". In: *ISPRS Journal of Photogrammetry and Remote Sensing* 151, pp. 207–222. DOI: 10.1016/j.isprsjprs.2019.02.021.

[60] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. (2019). "Semantic Understanding of Scenes Through the ADE20K Dataset". In: *Int. J. Comput. Vis.* 127, pp. 302–321. DOI: 10.1007/s11263-018-1140-0.

D Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering

Abstract

Deep learning (DL) is one of the key tools for analyzing images beyond the visible light spectrum, such as thermal data, for energy-related inspection and fault detection. However, publications using multispectral data focus on developing specialized models to handle quality issues without considering the imagery itself. This article investigates how feature engineering (FE), the process of adapting raw data to serve as DL training data, can impact performance when transferring prevalent model architectures to combined red, green, blue (RGB) thermal imagery. The popular U-Net is utilized for the common task of multiclass semantic segmentation in remote sensing. A comprehensive ablation study is performed on a novel, uncrewed aircraft system-based dataset from two German cities to detect thermal urban features. Common performance metrics, training, and energy consumption statistics are compared to find the most suitable combination of platform-specific and general enhancing FE while identifying the impact of resolution, channel count, RGB, and color information.

The study reveals FE to significantly influence predictive performance, where the choice of ablation parameters are found to have a 7 %pt to 10 %pt impact. Computational resource utilization depends on image size, following a logarithmic growth curve. Importantly, the study demonstrates that in-depth FE of thermal imagery can replace the need for additional RGB data.

Abbreviations

AI artificial intelligence

CE cross entropy

CLAHE contrast limited adaptive histogram equalization

DL deep learning

DHS district heating system

FE feature engineering

GE general enhancement

GPU graphics processing unit

IoU intersection over union

LWIR long-wavelength infrared

MFNet Multispectral Fusion Network

ML machine learning

P precision

PS platform-specific

PSTNet Penn Subterranean Thermal Network

RGB red, green, blue

RGBT red, green, blue, thermal

RS remote sensing

SD standard deviation

TIR thermal infrared

UA uncrewed aircraft

UAS uncrewed aircraft system

UM unsharp masking

VC vignetting correction

D.1 Introduction

While political and social efforts strive to limit anthropogenic global warming, the building sector remains one of the greatest contributors to climate change [45]. It is responsible for approximately 30 % of the global energy demand, primarily due to operational requirements such as heating [45]. The German government has therefore recently passed a law to establish a concept for country-wide heat supply, whereby municipalities are charged with creating comprehensive plans for climate-neutral heating in their communes [7]. Similar legislation already in effect in Scandinavian countries such as Denmark has led to the implementation of centralized technologies – specifically district heating systems (DHSs) – which currently provide two thirds of the population with 89 % climate-neutral heat [3]. However, the efficient operation of such systems still presents a significant challenge. In

Germany, for example, network losses were estimated at over 10 % in 2022 [3]. Keeping in mind current and future heating-related aims, a key part of enabling sustainable cities must lie in ensuring a high level of efficiency and minimal thermal losses in all involved infrastructure.

A versatile and holistic monitoring approach can be achieved by combining modern technologies: Uncrewed aircraft systems (UASs) with multispectral sensors gather image data to be efficiently analyzed via computer vision methods like artificial intelligence (AI) [5]. Where images are concerned, deep learning (DL) has proven to be extremely effective for classification, detection, and segmentation tasks in numerous domains. The last decade has seen an increase in the application of DL to imagery beyond the visible light – 380 nm to 700 nm wavelength – spectrum [18, 25, 39]. Thermal infrared (TIR) data, for instance, captures emitted electromagnetic waves corresponding to surrounding temperatures and thus can provide context information on heat sources [15, 25]. With regard to cities, its use ranges from autonomous driving over crowd counting to the maintenance of energy-related systems, including defect detection in DHS, solar technologies, or transmission lines, and building inspections [15, 25, 48, 53]. These problems are increasingly being addressed with complex, high-level semantic segmentation models as they enable pixel-wise classification and more nuanced contextual perception [25]. Merging standard red, green, blue (RGB) with TIR imagery is said to enrich scene understanding, help supply missing information under complex illumination, and greatly increase segmentation accuracy [25, 39, 53].

However, transferring DL to novel data types comes with a set of challenges. While images are generally subject to noise and acquisition-dependent artifacts [29], TIRs are particularly susceptible due to the involved sensor technology. They suffer from substantially lower resolution and undesirable effects – most prominently blurring and non-uniformity [18, 51, 52]. Although data quality is known to have a great impact on DL performance [29], studies focus on model adaptation instead of manual feature engineering (FE) - the process of adapting raw data for its use in DL model training - to tackle the issue.

This article, therefore, presents a thorough investigation on the effects of FE on remote sensing (RS) red, green, blue, thermal (RGBT)-based DL. The challenging and increasingly popular computer vision task of semantic segmentation is examined within the framed context of city infrastructure monitoring - specifically for identifying thermal, meaning heat-related, urban features and anomalies. Our contributions are as follows:

1. We identify key factors impacting quality in UAS-based RGBTs as either platform-specific (notably vignetting) or general (specifically contrast and blurring) and find suitable algorithms for mitigation.
2. For increased impact, we adapt a prevalent DL model to a new, real-world case study in heat-related inspection, thus creating a novel RGBT dataset.
3. An extensive ablation study is conducted to analyze numerous data-related aspects – including filters, channel constellations, and image sizes – with particular emphasis on the impact of information loss. Our comprehensive evaluation compares not only

a broad range of performance metrics, but also resource statistics to assess our AI in terms of sustainability [10].

4. We provide best-practice conclusions for multispectral remote sensing (RS) image acquisition and analysis for future sustainable cities.

Additionally, following open science principles, our novel RGBT dataset [47] and code[1] are published alongside this article, thus ensuring reproducibility.

Fig. D.1 visualizes the developed data processing pipelines, including registration to create an RGBT dataset forming the basis of the FE ablation study.

The rest of this article is organized as follows. After covering related work in Section D.2, all elements in Fig. D.1 are introduced in Section D.3 and detailed in Section D.4. Results are presented in D.5 and discussed in Section D.6, and finally, Section D.7 concludes this article.

D.2 Related work

A review in Neupane et al. [34] of semantic segmentation in RS found that 89% of papers compare different models, while 93% perform architecture-based ablation studies. Similarly in multispectral data, Kütük and Algan [25], Song et al. [39], and Zhang et al. [53] list various models developed for RGBT data – such as the Penn Subterranean Thermal Network (PSTNet) [38] – but omit FE.

Many studies, especially in RS, disregard FE owing to their use of benchmark datasets [34]. Publications that inspect data cleaning in RGB imagery find that it improves DL performance. Undesirable illumination can be mitigated with contrast-enhancing algorithms such as contrast limited adaptive histogram equalization (CLAHE) to increase segmentation accuracy [49]. In satellite imagery, atmospheric artifacts can be reduced by increasing contrast, i.e., through unsharp mask and median filtering to enhance classification accuracy and decrease computational complexity [36].

Fewer studies look into FE in TIR-based DL models. In spite of the technology's numerous uses, He et al. [18] find that current implementations are more akin to laboratory than industrial applications. This may allow for less artifact-ridden data and explain the model-driven focus. TIR quality improvement is only described as increasing resolution by using specifically developed DL architectures [18]. While Chaverot et al. [8] observe that DL commonly replaces image processing, they find TIR quality enhancement, i.e., via deblurring to significantly improve object detection. Data cleaning of TIRs for DL is described by Herrmann et al. [19], who apply various filters to mimic RGB appearance, and find that these improve the performance of RGB-pretrained DL.

[1] https://github.com/emvollmer/TUFSeg

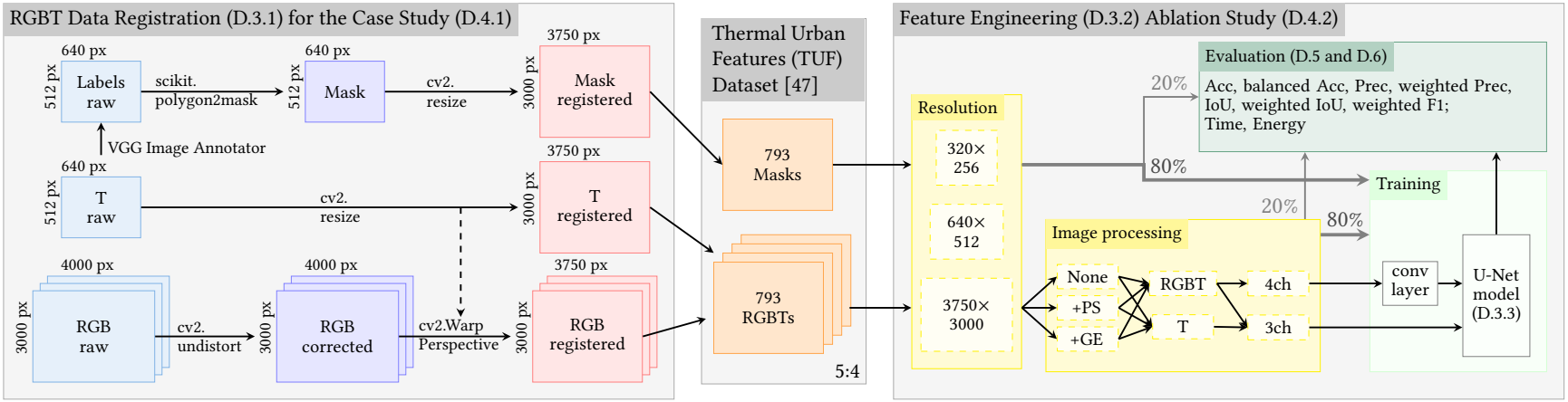


Figure D.1: Overview of the study and developed data processing pipelines: From RGB and TIR image registration for RGBT dataset creation to exhaustive FE ablation study. The numbers in brackets refer to manuscript sections containing further details on the step in question.

Though FE improves RGB- and TIR-based DL performance, there exists – to the knowledge of the authors – no study exploring the effects on combined RGBT-based semantic segmentation. This article, therefore, utilizes the well-known U-Net model for an ablation study focused on the central task of multiclass thermal urban feature segmentation. By identifying common and anomalous heat sources in urban environments, DL can help detect DHS leakages conservatively by removing false alarms [48].

D.3 Methods

D.3.1 RGBT data registration

RGBT imagery is made up of four channels: three color and one temperature-dependent one. Recording these data requires both a visible light and infrared sensor. The most common and cheapest thermal sensors are uncooled microbolometers, which capture LWIR radiation emitted by all objects of common temperature on Earth ($-83°C$ to $727°C$). The resulting outputs are in grayscale, with lighter pixels denoting higher temperatures. [15, 18, 51]

Typically, two separate sensors are used for acquisition [39], placed side by side to match the fields of view [38]. Hybrid cameras can facilitate the process by incorporating both technologies [17, 39]. RS acquisition, especially by UAS, is simplified through merged dual camera and gimbal systems [20]. In all cases, the raw images require alignment to compensate differing fields of view, resolution, and aspect ratios [17, 32]. TIRs (around 640×512 or less [18]) generally have a considerably lower resolution than RGBs (around 1920×1080 [39] or up to 4000×3000 [32]).

Aligning RGBs and TIRs typically consists of two steps: distortion correction and image registration [20, 38]. Depending on the used camera, images typically suffer from two types of distortions: radial (barrel or pincushion effects) or tangential (lack of lens alignment with the image plane) [33]. To correct these, a camera's intrinsic and extrinsic parameters need to be estimated. Most simply, this is achieved by recording reference images of a standardized pattern, such as checkerboard of black and white squares, with which calibration functions from computer vision libraries such as OpenCV [6] can estimate a camera's intrinsic matrix and distortion coefficients. When applied, these will correct sensor-specific distortion in the given data. [2, 20, 31, 38]

After distortion correction, the two sensor outputs are aligned on pixel-level to compensate any offset between camera views (existing even for dual cameras). A simple registration approach estimates the homography matrix that describes the cameras' relationship with coordinate pairs from matching key points. Both TIR and RGB are manually sampled to identify a list of corresponding pixel locations for matrix estimation. Applying this transformation warps the images to the same resolution. [20, 31]

D.3.2 Feature engineering

Raw imagery can be affected by noise, lack of contrast, problematic lighting conditions, blur, and other artifacts. How severely these are expressed depends on the utilized technology and acquisition conditions. For instance, capturing images by popular UAS [18, 39] elicits pronounced non-uniformity in TIRs [51]. Nighttime recordings are most useful for TIR information due to low thermal reflectance, but RGBs suffer from reduced luminance.

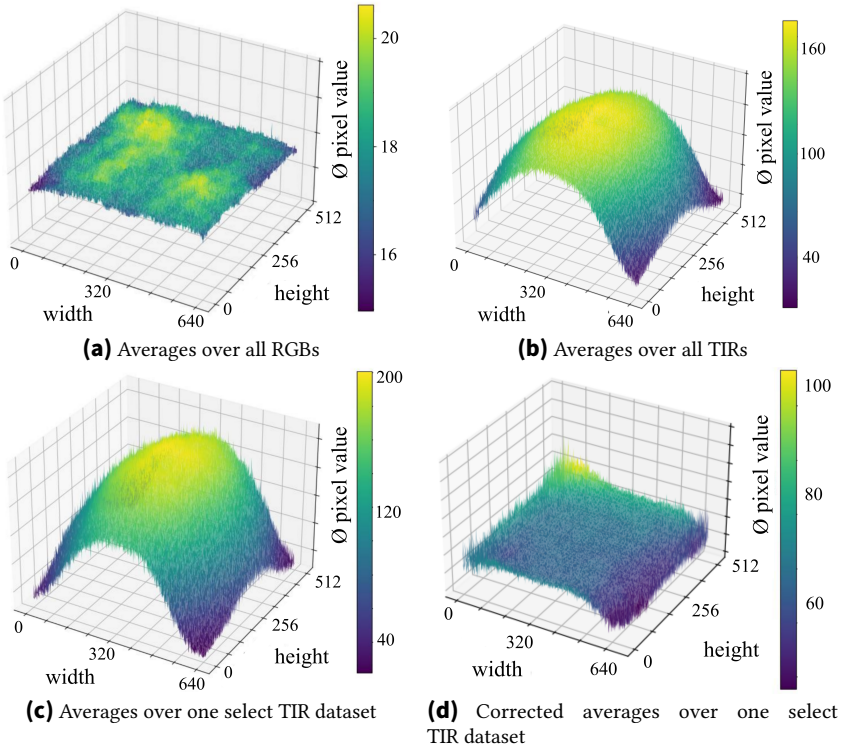
Observations from previous studies and existing conditions help focus on two key aspects: platform-specific and general image enhancement FE. Platform-specific speaks to the compensation of effects induced by the utilized RS acquisition method – in this case UAS. General enhancement mitigates quality issues unrelated to the form of acquisition and instead pertaining to the sensors themselves. These are centered around contrast and blurring.

D.3.2.1 Platform-specific feature engineering

UAS image acquisition has several advantages, including high flexibility, easy operability, and high spatial resolution through low flight heights [23]. Uncrewed aircrafts (UAs) are utilized with dual RGB and TIR sensors for crop [23], building [32], power line [9] monitoring, and many more. However, the prevalent conditions during flights can particularly affect the thermal camera, as its sensor, lens, and housing temperatures are affected by heat from the gimbal motor and coolness from propellers' slipstream [51]. The resulting non-uniformity manifests as a cooling of image corners and edges [51]. Although microbolometers have an integrated correction to compensate for fixed pattern noise, they cannot adequately offset this so-called "vignetting" or "halo" effect [51].

Fig. D.2 visualizes the average pixel values of all thermal versus visual images. Ideally, the averages are close to equal for a uniform distribution, as is true for RGBs [see Fig. D.2a]. The TIR channel, however, displays an extreme radial deviation [see Fig. D.2b]. This effect manifests differently depending on individual flight conditions [see Fig. D.2c].

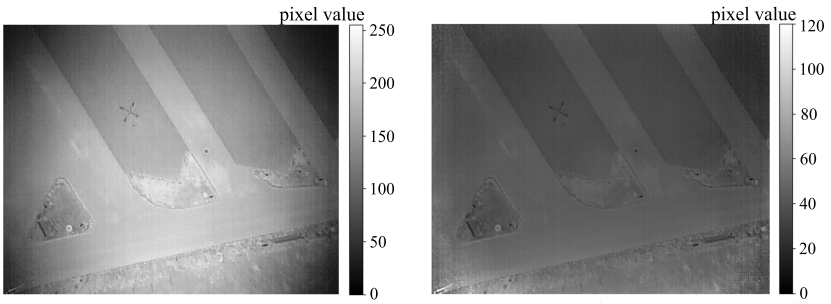
Various vignetting correction (VC) approaches exist, but they do not always negate the entire effect and often require a reference image [24, 51]. Therefore, a method utilizing radial polynomial functions as described in Bal and Palus [4] is implemented. Pixels are grouped based on radial distance to the image center and bin averages used to model a vignetting function. The function is approximated for each image, so that magnitude variations between images or datasets do not pose a problem. Here, 100 bins and a tenth degree polynomial function are found to be optimal. Fig. D.3 exemplifies how the algorithm corrects substantial vignetting in a TIR from Fig. D.2c's dataset. Fig. D.2d shows the corrected pixel distribution for that specific dataset.



(a) Averages over all RGBs **(b)** Averages over all TIRs

(c) Averages over one select TIR dataset **(d)** Corrected averages over one select TIR dataset

Figure D.2: Location-dependent pixel distributions, colorized according to pixel values from blue (low) to yellow (high).



(a) TIR image **(b)** VC TIR

Figure D.3: Visualization of the VC algorithm.

D.3.2.2 General enhancement

Image enhancement refers to the improvement of quality, specifically contrast and blur, regardless of the acquisition platform.

Contrast enhancement Undesirable lighting conditions can greatly degrade image quality. Phenomena like brightness, shadows, over-, or underexposure induce noise and obscure object features, especially outdoors. Contrast enhancement restores images to compensate such effects and increase quality. Illumination is particularly problematic in RGBs when images are acquired at night. [49]

Several algorithms can perform this task, most prominently the histogram-based CLAHE [49] and Retinex [35] methods. Both are designed to compensate for spatially varying non-uniformity resulting from varying brightness (i.e. enhance local contrast), but can also suppress noise in RGBs and TIRs [40, 49, 52]. [27]

Histogram equalization enhances image contrast by stretching existing pixel values across all possible values. Though effective, this technique can introduce noise and reduce information content in images with high contrast. An improved variant is CLAHE, a method that first divides the image into small areas before applying histogram equalization to limit excessive contrast and increase noise robustness in low-contrast areas. [49, 50]

Retinex imitates the behavior of photoreceptors and ganglion cells in the human eye (retina) and processing structures of the mind (cortex). Images are interpreted as a multiplication of illumination (overall scene lighting) and reflectance (intrinsic scene properties). By separating the two components and applying local logarithmic luminance compression and spectral brightening, varied illumination can be compensated and quality improved. [35, 40]

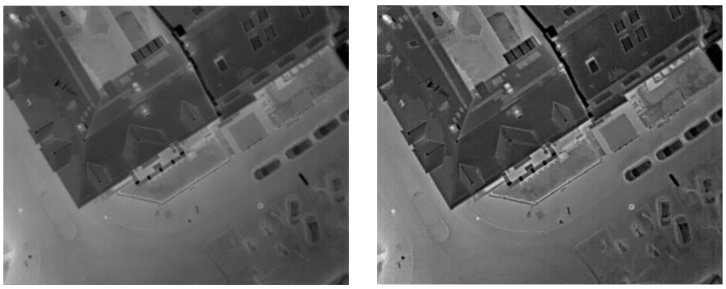
While Retinex can be applied directly to three-channel data, CLAHE only works with single channels. It is common practice to convert RGBs into $L \times A \times B$ color space (where L represents luminance and A and B color distributions) and enhance the L channel [54]. Fig. D.4 compares both algorithms. Retinex is capable of extracting significantly more information from darker image regions while preserving the structure of illuminated ones. It is therefore implemented for contrast increase here. As the VC from Section D.3.2.1 can noticeably darken TIRs, the algorithm is also applied to increase their luminance.

(a) RGB image **(b)** CLAHE RGB **(c)** Retinex RGB

Figure D.4: Visualization of the contrast enhancing algorithms CLAHE (D.4b) and Retinex (D.4c).

Deblurring TIRs generally suffer from blurring, which complicates the detection of smaller or farther objects. This can be mitigated with the unsharp masking (UM) algorithm, as visualized in Fig. D.5. The technique involves blurring an image I with a noise-reducing filter (typically Gaussian) G and concatenating the output with I according to D.1. λ is a positive integer defining the effect's strength with a default of 1, used here to prevent additional, unwanted noise artifacts. [8, 11]

$$I_{corr} = I + \lambda(I - G(I)) \tag{D.1}$$



(a) TIR image **(b)** UM TIR

Figure D.5: Visualization of the UM algorithm.

UM is also commonly applied to RGBs. In satellite imagery, it has been shown to increase DL classification accuracy by reducing artifacts and enhancing edges [36] and therefore is used on our remotely sensed RGBs as well.

D.3.3 Model architecture

Despite the ongoing development of specialized models (see Section D.2), the U-Net architecture [37] is still the most prevalent in RS [28]. The model enhances a classical fully convolutional network by including skip connections between encoder and decoder [37]. This improves segmentation when working with limited training data, as is common for real-world implementations [37]. The model excels in various fields, is among the most popular for urban feature segmentation [30, 34, 44], and known to proficiently analyze multispectral satellite data [22]. Its continued relevance makes the U-Net a good choice, as it allows for more generalisable and thus significant results.

Architecture-related details are based on [34]'s survey of urban feature segmentation in RS images. The most common backbone (also for RGBT data [39]) is the ResNet, which is why the ResNet-152 is selected as an encoder [34]. Chiefly, cross entropy (CE) functions are used to determine loss [34]. An adapted variant – focal CE – targets difficult-to-learn instances and helps manage the common issue of class imbalance [13, 26]. As this works well for training a U-Net on multispectral satellite data [13], the sigmoid focal CE function [26] is implemented here.

Transfer learning can balance limited datasets [1]. Initializing a model with weights from other trainings helps compensate for a lack of data with knowledge from a related task [1]. While public weights are based on RGBs, ablation studies such as [32] have shown DL performance to significantly improve when used with RGBT data. Therefore, the U-Net encoder backbone is loaded with weights trained on the popular ImageNet dataset [12], as done in Adiba et al. [1]. Because of this, the model only accepts inputs with the same channel count. Two options exist to adapt to RGBT images: concatenate the data into three channels or map the fourth channel to the given three with a preceding convolutional layer. Both are compared here.

D.4 Experiment

The overall data handling procedure developed for this case study experiment is visualized in Fig. D.1. The following subsections address relevant aspects of both pipelines.

D.4.1 Case study description and data registration

The case study comprises images from various suburban regions around Munich (Germany), captured from 8 p.m. to 6 a.m. in December 2019 with temperatures ranging from −5 °C to 2 °C. Additional data from urban areas in a second German city, Karlsruhe, help diversify the study. These recordings were acquired in January and March 2022 with temperatures of 0 °C to 3 °C.

Acquisition took place via Matrice 600 Pro [41] and 300 RTK [43] UA. The flight was almost entirely automated to follow a lawnmower pattern at 60 m altitude in nadir. RGBT imagery was captured simultaneously with DJI's Zenmuse XT2 [42], a combined gimbal and camera incorporating a 4k RGB and thermal sensor by FLIR. The RGBs have a size of 4000×3000 pixels, the TIRs 640×512.

Of 8,452 combined images, 793 are selected for annotation – 700 from Munich and 93 from Karlsruhe. Owing to an 88% overlap, only every ninth image depicts an entirely new scene and is therefore worth considering for annotation. Of these, trained experts select the most suitable by avoiding duplicate areas and motion blur due to turns during UAS flight. Within the TIRs, we identify nine classes of common thermal urban features and label a total of 8,010 polygons using Visual Geometry Group's VGG Image Annotator [14]. The annotation masks required for semantic segmentation are generated from these polygons by assigning each pixel a number per its defined class and all unlabelled pixels to the background. The class distribution in Appendix D.I highlights an annotation and pixel imbalance common to these types of tasks.

The raw images are processed according to Section D.3.1, as shown in the left part of Fig. D.1. This case study's RGBs suffer from a distinct fisheye distortion, which is corrected using 129 key points before image registration with Python's OpenCV [6] and scikit [46]. A result of annotating the TIRs is an aspect ratio of $5 : 4$. The data are scaled to 3000×3750 to match RGB resolution.

D.4.2 Ablation study

Table D.1 shows how the ablation study investigates manual FE for DL. Aside from image filters, we examine the effect of information loss – specifically color when reducing channel count and content when varying resolution – to identify attributes influencing model performance.

Table D.1: Overview of parameter combinations and channel input definitions.
Legend: proc = processing, ch = channel, ret = retinex

Parameters			Model channel inputs			
proc	**ch**	**data**	**ch1**	**ch2**	**ch3**	**ch4**
none	3	T	T	T	T	-
+PS	3	T	vc T	vc T	vc T	-
+GE	3	T	vc T	ret T	um T	-
none	3	RGBT	gray RGB	T	T	-
+PS	3	RGBT	gray RGB	vc T	vc T	-
+GE	3	RGBT	ret RGB	ret T	um T	-
none	4	RGBT	R	G	B	T
+PS	4	RGBT	R	G	B	vc T
+GE	4	RGBT	ret RGB	ret T	um RGB	um T

Based on Section D.3.2, three key FE options are defined: none, platform-specific (PS) processing – meaning vignetting removal –, and general enhancement (GE) – meaning additional Retinex and unsharp mask algorithms. These are applied in parallel (individually per channel) as this has been shown to yield better results in TIR-based DL than consecutive preprocessing [19].

As discussed in Section D.3.3, both the concatenation into three-channel and model adaptation to four-channel inputs are tested. RGBs are converted into single-channel greyscales, meaning the relevance of color can be investigated. With a thermography-centered objective, we can additionally investigate the sole use of TIR inputs as a baseline. This alleviates higher processing costs of combining RGBs and TIRs and potential registration discrepancies [25].

In all Table D.1 configurations, the input data is scaled to high (3750×3000), mid (640×512), and low (320×256) resolution, as larger files require more significant hardware capacities. We can thereby investigate what impact image size has on RGBT or TIR model performance and whether an inverted U-shaped relationship exists here. To allow for a simple, statistical analysis [32], each configuration is trained with four seeds (see Appendix D.II for more details).

The resulting 108 models are evaluated using a wide range of semantic segmentation metrics: overall accuracy, balanced accuracy, precision (P), weighted P, weighted F1-score, mean intersection over union (IoU), and weighted mean IoU. These are chosen based on the most common metrics in RS urban feature detection [34] and RGBT segmentation [39]. The balanced and weighted variants consider class imbalance by calculating class-wise scores and determining (weighted) averages. These metrics, alongside their standard equivalents, are particularly helpful for a comprehensive estimation of model performance here.

Measured resource metrics include time used for FE and model training as well as energy consumption in accordance with sustainable AI principles [10]. The energy used is calculated with Perun [16], which outputs both kWh and $kgCO_2eq$.

D.5 Results

Table D.2 summarizes the ablation study results, divided into performance and resource related metrics. To compensate fluctuations due to seed initialization, these are presented as “mean ± standard deviation (SD)", calculated across the four selected seeds (see Appendix D.II).

D.6 Discussion

D.6.1 Performance

D.6.1.1 Quantitative evaluation

A characteristic trend of overall high- and low- scoring metrics confirms the presence of a strong class imbalance. A generally high mean accuracy (88 % to 96 %) and low balanced accuracy (48 % to 59 %) signifies that dominant classes are predicted much more accurately than underrepresented ones. Like balanced accuracy, precision considers each class equally and lies in a comparable range. With a weighting factor defined by the number of true class instances, weighted precision balances class size discrepancies and scores very high (92 % to 96 %). The averaged mean IoU scores are lowest (39 % to 48 %), while the class-weighted variant reaches 83 % to 92 %. Even higher scores are obtained by the weighted F1 metric. In total, the choice of parameter values accounts for a 7 %pt to 10 %pt difference across the various performance metrics.

SDs are only 3.7 % on average, but a 2.74 % median and 10 % to 18 % peak values signify outliers – most prominently the three-channel TIR mid-resolution none, RGBT mid-sized PS, and high-resolution GE. These stem from individual seeds curtailing performance – here numbers 1,000, 1,234,567, and 1,000 respectively. While subsequent analyzes use mean values, these outliers can influence general conclusions.

The simplest data processing without filters using three-channel RGBT high resolution images scores highest for the lower metrics, specifically balanced accuracy, mean IoU and second for precision. The higher performing metrics (accuracy, weighted precision, weighted IoU, and weighted F1) almost all peak for PS FE with the highest resolution four-channel inputs. Despite this, the overall best results are found for GE three-channel mid sized TIR data, which scores best in both precision-related metrics and second in all others. This is an interesting observation, as it indicates that supplementary RGBs are not necessarily as necessary when the TIRs are feature engineered. Table D.2 reveals a clear pattern for near to all metrics regarding the sole use of thermal imagery: PS and GE FE improve upon performance (up to 5 %pt averaged). To allow further study, filters and channels are analyzed separately from image sizes.

ch	data	proc	res	Performance Metrics							Resource Metrics			
				Accuracy	balanced Acc	Precision	weighted P	IoU	weighted IoU	weighted F1	t_{data} [min]	t_{train} [min]	energy [kWh]	energy [kgCO_2e]
3	T	none	high	0.915 ± 0.067	0.556 ± 0.043	0.570 ± 0.033	0.946 ± 0.017	0.453 ± 0.056	0.870 ± 0.081	0.918 ± 0.054	1.310 ± 0.088	15.12 ± 1.154	0.180 ± 0.016	0.076 ± 0.007
			mid	0.892 ± 0.137	0.566 ± 0.041	0.562 ± 0.097	0.948 ± 0.032	0.449 ± 0.108	0.848 ± 0.163	0.899 ± 0.117	0.458 ± 0.299	14.46 ± 0.084	0.167 ± 0.004	0.070 ± 0.002
			low	0.904 ± 0.095	0.514 ± 0.079	0.542 ± 0.062	0.943 ± 0.031	0.417 ± 0.086	0.860 ± 0.116	0.909 ± 0.084	0.062 ± 0.045	4.720 ± 0.020	0.051 ± 0.000	0.021 ± 0.000
		+PS	high	0.933 ± 0.042	0.561 ± 0.025	0.574 ± 0.045	0.954 ± 0.012	0.460 ± 0.043	0.895 ± 0.050	0.936 ± 0.034	2.982 ± 0.491	15.13 ± 0.895	0.192 ± 0.007	0.080 ± 0.003
			mid	0.945 ± 0.013	0.507 ± 0.063	0.573 ± 0.052	0.950 ± 0.013	0.435 ± 0.053	0.904 ± 0.023	0.941 ± 0.017	2.423 ± 0.288	14.50 ± 0.104	0.180 ± 0.003	0.076 ± 0.001
			low	0.951 ± 0.004	0.531 ± 0.024	0.572 ± 0.006	0.956 ± 0.003	0.455 ± 0.015	0.916 ± 0.006	0.950 ± 0.004	0.418 ± 0.032	4.715 ± 0.017	0.053 ± 0.001	0.022 ± 0.001
		+GE	high	0.949 ± 0.009	0.538 ± 0.037	0.573 ± 0.056	0.954 ± 0.006	0.457 ± 0.034	0.910 ± 0.015	0.944 ± 0.011	5.208 ± 0.456	14.54 ± 0.121	0.202 ± 0.011	0.084 ± 0.004
			mid	0.954 ± 0.001	0.575 ± 0.004	0.584 ± 0.019	0.960 ± 0.002	0.477 ± 0.010	0.921 ± 0.002	0.953 ± 0.002	4.792 ± 0.270	14.37 ± 0.120	0.193 ± 0.003	0.080 ± 0.001
			low	0.946 ± 0.017	0.542 ± 0.042	0.531 ± 0.045	0.955 ± 0.008	0.435 ± 0.035	0.909 ± 0.022	0.945 ± 0.014	1.245 ± 0.048	4.718 ± 0.039	0.059 ± 0.001	0.024 ± 0.001
3	RGBT	none	high	0.948 ± 0.013	0.585 ± 0.031	0.580 ± 0.036	0.955 ± 0.008	0.478 ± 0.029	0.911 ± 0.018	0.947 ± 0.011	1.632 ± 0.555	14.54 ± 0.147	0.172 ± 0.005	0.072 ± 0.002
			mid	0.949 ± 0.004	0.563 ± 0.031	0.580 ± 0.025	0.956 ± 0.005	0.464 ± 0.025	0.913 ± 0.009	0.948 ± 0.007	0.510 ± 0.262	14.50 ± 0.057	0.168 ± 0.003	0.070 ± 0.001
			low	0.943 ± 0.011	0.537 ± 0.032	0.565 ± 0.021	0.952 ± 0.007	0.444 ± 0.034	0.904 ± 0.017	0.941 ± 0.011	0.062 ± 0.019	4.690 ± 0.043	0.051 ± 0.001	0.022 ± 0.001
		+PS	high	0.947 ± 0.007	0.522 ± 0.040	0.545 ± 0.051	0.955 ± 0.005	0.434 ± 0.041	0.911 ± 0.010	0.946 ± 0.006	2.940 ± 0.423	15.05 ± 1.193	0.186 ± 0.019	0.078 ± 0.008
			mid	0.916 ± 0.073	0.515 ± 0.076	0.526 ± 0.111	0.924 ± 0.068	0.407 ± 0.110	0.861 ± 0.113	0.902 ± 0.096	2.420 ± 0.313	14.40 ± 0.079	0.178 ± 0.005	0.074 ± 0.002
			low	0.948 ± 0.010	0.509 ± 0.041	0.552 ± 0.016	0.952 ± 0.008	0.434 ± 0.020	0.909 ± 0.016	0.945 ± 0.010	0.390 ± 0.000	4.748 ± 0.096	0.054 ± 0.002	0.022 ± 0.001
		+GE	high	0.878 ± 0.158	0.516 ± 0.094	0.546 ± 0.075	0.944 ± 0.035	0.417 ± 0.109	0.834 ± 0.183	0.888 ± 0.135	5.300 ± 0.402	14.42 ± 0.295	0.200 ± 0.013	0.084 ± 0.005
			mid	0.951 ± 0.008	0.501 ± 0.017	0.543 ± 0.049	0.955 ± 0.006	0.430 ± 0.026	0.915 ± 0.012	0.948 ± 0.009	4.730 ± 0.306	14.55 ± 0.034	0.196 ± 0.005	0.082 ± 0.002
			low	0.950 ± 0.006	0.505 ± 0.036	0.527 ± 0.031	0.955 ± 0.004	0.429 ± 0.018	0.914 ± 0.008	0.948 ± 0.006	1.235 ± 0.031	4.775 ± 0.147	0.060 ± 0.002	0.025 ± 0.001
4	RGBT	none	high	0.937 ± 0.030	0.574 ± 0.035	0.563 ± 0.065	0.949 ± 0.019	0.458 ± 0.066	0.896 ± 0.043	0.937 ± 0.027	1.955 ± 0.590	14.81 ± 0.356	0.180 ± 0.008	0.075 ± 0.003
			mid	0.931 ± 0.042	0.555 ± 0.073	0.563 ± 0.062	0.953 ± 0.015	0.458 ± 0.071	0.892 ± 0.055	0.934 ± 0.036	0.600 ± 0.232	15.31 ± 1.361	0.179 ± 0.026	0.075 ± 0.011
			low	0.953 ± 0.005	0.498 ± 0.022	0.557 ± 0.022	0.956 ± 0.004	0.439 ± 0.018	0.918 ± 0.009	0.950 ± 0.006	0.335 ± 0.359	4.930 ± 0.315	0.057 ± 0.010	0.024 ± 0.004
		+PS	high	0.956 ± 0.004	0.534 ± 0.018	0.576 ± 0.006	0.960 ± 0.003	0.452 ± 0.010	0.923 ± 0.006	0.954 ± 0.003	3.192 ± 0.599	14.77 ± 0.316	0.187 ± 0.007	0.078 ± 0.003
			mid	0.915 ± 0.054	0.513 ± 0.048	0.525 ± 0.059	0.939 ± 0.020	0.402 ± 0.062	0.868 ± 0.067	0.917 ± 0.044	2.462 ± 0.318	14.56 ± 0.122	0.180 ± 0.004	0.075 ± 0.002
			low	0.919 ± 0.070	0.482 ± 0.047	0.527 ± 0.081	0.943 ± 0.028	0.395 ± 0.084	0.878 ± 0.082	0.921 ± 0.060	0.432 ± 0.078	4.762 ± 0.046	0.054 ± 0.001	0.022 ± 0.001
		+GE	high	0.936 ± 0.036	0.558 ± 0.025	0.556 ± 0.047	0.951 ± 0.017	0.449 ± 0.051	0.897 ± 0.048	0.938 ± 0.031	5.580 ± 0.632	15.06 ± 0.988	0.209 ± 0.009	0.087 ± 0.004
			mid	0.952 ± 0.005	0.563 ± 0.047	0.551 ± 0.026	0.958 ± 0.002	0.460 ± 0.021	0.919 ± 0.006	0.951 ± 0.004	5.020 ± 0.256	14.57 ± 0.074	0.199 ± 0.006	0.083 ± 0.003
			low	0.946 ± 0.010	0.504 ± 0.027	0.539 ± 0.046	0.952 ± 0.006	0.425 ± 0.030	0.908 ± 0.014	0.944 ± 0.009	1.268 ± 0.028	4.765 ± 0.026	0.060 ± 0.001	0.025 ± 0.000
minimum mean value				0.878	0.482	0.525	0.924	0.395	0.834	0.888	0.063	4.690	0.051	0.021
maximum mean value				0.956	0.585	0.584	0.960	0.478	0.923	0.954	5.580	15.31	0.209	0.087
range of mean values				0.078	0.102	0.059	0.036	0.083	0.089	0.065	5.518	10.62	0.158	0.066

Table D.2: Ablation study results. Increased opacity denotes more favorable values (higher for performance, lower for resources). All are formatted as "mean ± standard deviation (SD)", calculated across the four seed initializations.
Legend: ch = channel, proc = processing, res = resolution.

Fig. D.6 compares averages over all image sizes. Overall, raw data yields the best results for lower metrics, while GE (followed by PS) scores highest in the higher metrics. As the latter measure representative performance despite class imbalance, this indicates that FE helps overall performance, but does not aid with underrepresented classes.

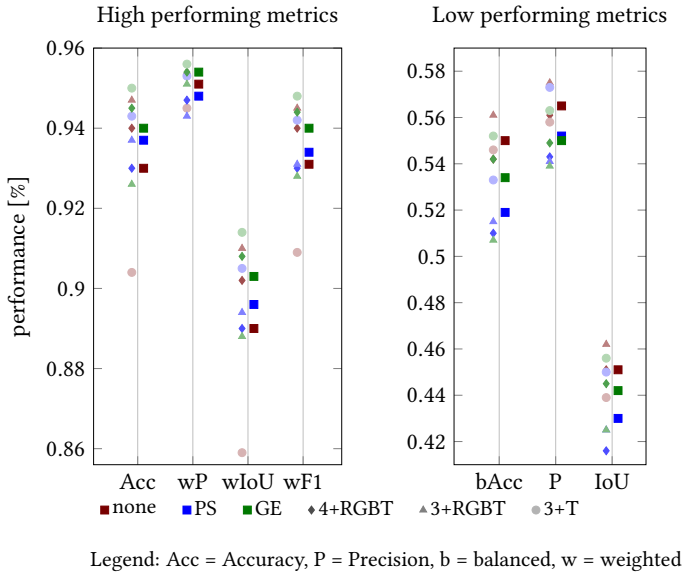


Figure D.6: Performance metrics for filter and channel input combinations, averaged over all image sizes.

Preferable channel combinations also vary. For the raw data, 3+RGBT inputs are best, closely followed by 4+RGBT. Using only TIRs yields conspicuously low results (up to 5 %pt decrease), indicating that RGBs should be included when utilising raw data. Contrarily, the closeness of 3+ and 4+RGBT scores shows that the loss of color caused by the greyscale transformation has little impact.

For PS and GE, the best results are achieved with 3+T inputs, meaning the sole use of TIRs is beneficial. This could imply that RGB FE is less helpful than initially hypothesized, but individual score comparisons contradict this. The mentioned outliers may be causing the discrepancy. Compared to raw 3+T inputs, both PS processing and added RGB information individually improve results, yet their combination in PS 3+RGBT performs worse than either. This may indicate an issue with data registration.

An evaluation of how image size influences performance finds the highest resolution (3750×3000) to provide the best results most often, followed closely by mid-sized images. Only in few instances does the smallest resolution yield the best models. Interestingly, for GE, mid resolution always produces the highest scores. Analysing averages for each image size shows that highest resolution images are particularly helpful in increasing performance of the low performing metrics by up to 4 %pt. For higher scoring metrics, total averages vary only slightly (around or less than 1 %pt), meaning differences have a smaller impact.

D.6.1.2 Qualitative evaluation

The qualitative comparison in Fig. D.7 visualizes the impact of FE and RGB information loss and its influence on model behavior. The generated predictions allow conclusions to be drawn about how variants trained on differently processed data react to representative scenarios. In line with Fig. D.6, a representative model is selected for low (raw 3+RGBT) and high (GE 3+T) performing metrics each - specifically 640x512 resolution and seed number 42, since these are closest to their variants' average performance.

Fig. D.7 clearly underlines the previous differentiation between high and low performing metrics. Common classes (i.e. buildings and cars) are overlooked far less often than the less prevalent ones (i.e. street lamps) because of annotated instance and pixel amounts (see Appendix D.I).

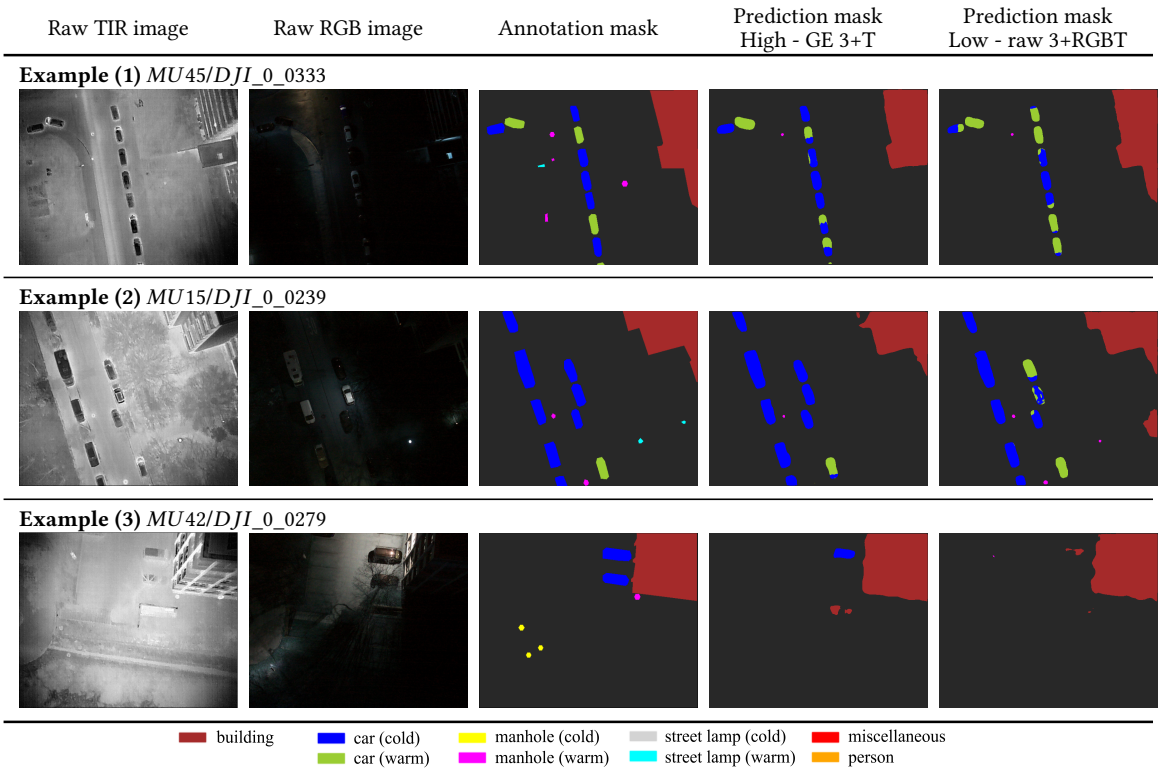


Figure D.7: Qualitative comparison of high (3+T) versus low (3+RGBT) performing metrics winners, exemplified via the model variant of resolution 640 × 512 and seed 42.

A closer look at example D.7.(1) shows the 3+T model trained on FE data to be capable of more nuanced distinctions between common objects, such as warm and cold cars. It specifically associates only the warm pixels with said class, actually surpassing the annotation mask in detail. The 3+RGBT model trained on raw data displays a tendency to over-classify, in this instance warm cars. This might be explained with the fact that the combined FE implemented in this study is able to compensate thermal halos like those around the cold vehicles in the raw TIR image. In contrast, this may also be what allows the raw 3+RGBT variant to be somewhat more sensitive to highly underrepresented classes,

as shown in example (2). Though not entirely correct, the raw 3+RGBT model's prediction is closer to the annotation mask in terms of warm manholes and street lamps, illustrating the reason for it scoring highest among the low performing metrics.

While related work generally sets the expectation that including additional RGB information will bolster segmentation performance [25], example D.7.(3) contradicts this assumption. Although the RGB visually aids human observers in identifying the cars in the top image half, the results show the 3+RGBT model utilising the raw imagery as incapable of doing so. The other variant, relying solely on FE thermal imagery, correctly classifies the non-obscured car. This, again, supports an important conclusion derived from the quantitative analysis, specifically that carefully selected FE can compensate or even outperform a lack of additional RGB data.

D.6.2 Resource utilization

As expected, the consumption of both time and energy for model training is mostly dependent on image resolution. Fig. D.8 exemplifies the correlation using averages for each image size. Interestingly, the relationship seems to follow a logarithmic growth curve. Although high resolution images are 6× the size of mid ones, the difference in resource requirements is negligible – especially compared to small images half their size. Significant time and energy savings can only be achieved with small images or, conversely, higher resolutions are not problematic in either regard after a certain saturation threshold has been met.

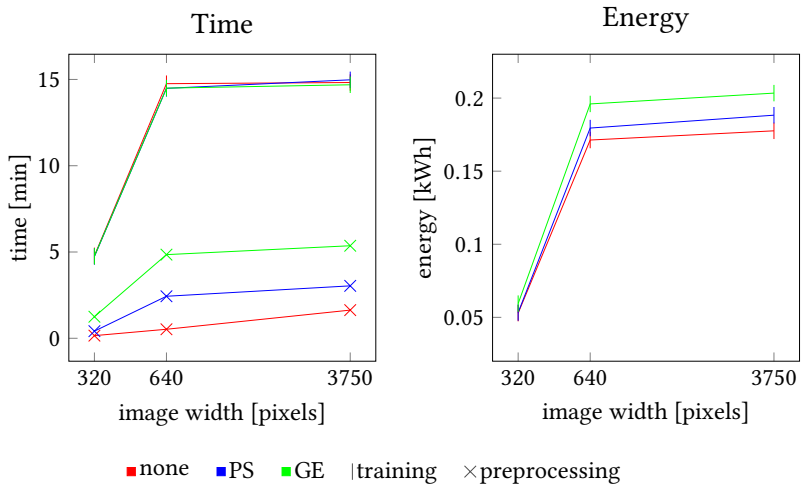


Figure D.8: Resource resolution relationship.

The impact of FE on resource utilization, as highlighted in Fig. D.8, is secondary in comparison to resolution. FE is negligible with respect to the training duration, as Fig. D.8 shows that the training times are only seconds apart. Where energy consumption is concerned, training with FE instead of raw imagery requires somewhat more kWh. These numbers increase only slightly for higher resolutions, from ca. 12 % to 14 %. However,

these are one-time costs, as the model requires training only once before it can be utilized. In terms of data preprocessing, GE unsurprisingly takes the longest, followed by PS. This also means that more time must be anticipated for inference, although it only amounts to 0.4 s (GE) versus 0.12 s (raw) per image for the highest resolution.

D.7 Conclusion

This study analyzed the influence of FE – specifically PS (vignetting correction) and GE (contrast increase and deblurring) – on a novel RGBT dataset for the task of multiclass thermal urban feature segmentation. We find such manual FE to account for a 7 %pt to 10 %pt difference in performance and can thus discredit the common assumption that DL will directly infer engineered features itself. While supplementary RGB information is beneficial when working with raw imagery, the overall best results are achieved by applying in-depth GE FE and using only TIR data. This has a significant implication for thermography-focused implementations as it means a less expensive acquisition with simple thermal sensors can counteract a lack of RGB information if FE is performed. In the context of managing more sustainable cities, this may have wide-reaching implications for future designs of smart city monitoring applications, where the problem of leakage detection in pervasive DHS technology can be combined with other, solely thermography-based applications such as building and solar panel inspections by cheaper means.

Regarding image size, high resolution data not only yields the best results most often, but also particularly improves those metrics performing less well due to class imbalance. Owing to the analysis of time and energy consumption, we now know that this costs surprising little additional resources owing to their logarithmic relationship. In contrast to findings of previous work, high quality TIR sensors are crucial in collecting data that will improve performance of economical, heat-related DL models.

This study is subject to some limitations. Only four seeds were used for initialization, lessening the statistical significance of the calculated SDs [32]. Data annotation and registration are subject to human error. In future, FE studies – especially regarding RGBs and data assimilation – can provide further insights. Additional experiments can help quantify the contribution of each implemented filter to DL model performance using standard, spectral, and multispectral data. This includes analyses using explainable AI to help further understand and characterize the models trained on differently preprocessed data. Implementing the presented FE in combination with other models will help to assess the general applicability of the derived conclusions.

Acknowledgments

The datasets were acquired in collaboration with the Air Bavarian GmbH and Munich's and Karlsruhe's municipal utilities companies. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

Declarations

This work is funded by the European Union through the AI4EOSC project (Horizon Europe) under Grant number 101058593.

CRediT author statement

Elena Vollmer: Conceptualization, Methodology, Data Curation, Investigation, Software Development, Formal Analysis, Visualization, Writing — original draft preparation, Writing — review and editing. **Mishal Benz:** Conceptualization, Methodology, Writing — review and editing. **James Kahn:** Data Curation, Software Development. **Leon Klug:** Methodology, Software Development, Visualization. **Rebekka Volk:** Writing — review and editing, Supervision. **Frank Schultmann:** Writing — review and editing, Supervision. **Markus Götz:** Writing — review and editing, Supervision. All authors have read and agreed to the published version of the manuscript.

Data availability statement

The utilized, novel RGBT dataset will be made available with this publication via Zenodo [47]. The code is available at `https://github.com/emvollmer/TUFSeg`.

Appendices

Appendix I

An overview of class distributions is given in Table D.3.

Table D.3: Overview of class distributions.

Class	Polygon count	Pixel count
background	-	205,357,517
building	1,404	48,111,260
car (cold)	2,532	3,804,713
car (warm)	1,036	1,993,912
manhole (cold)	520	92,415
manhole (warm)	1,379	244,538
miscellaneous	81	50,762
person	275	38,901
street lamp (cold)	100	22,822
street lamp (warm)	683	133,400

Appendix II

Due to case study size, the data are simply split into 80 % training and 20 % test sets. While the split is randomized, three conditions are ensured: 1. all classes are represented in both sets, 2. both sets contain images from both cities, and 3. the annotation distribution is close to 80-20.

Each configuration is trained with four seeds, arbitrarily chosen to initialize model weights unspecified by transfer learning. These are: 42, 1000, 1 234 567, and 10 110 110. All variants are trained on the bwUniCluster2.0 high-performance computing system using a NVIDIA A100-PCI GPU for 35 epochs and a batch size of 8. We use Python `3.8` with OpenCV `4.6.0.66` and scikit-image `0.19.3` for processing, [21]'s segmentation_models `1.0.1` (tensorflow `2.10.0`) for training, tensorflow-addons `0.20.0` for loss definition, and scikit-learn `1.3.2` for evaluation.

Bibliography

[1] Adiba, A., Hajji, H., and Maatouk, M. (2019). "Transfer learning and U-Net for buildings segmentation". In: *New Challenges in Data Sciences: Acts of the Second Conference of the Moroccan Classification Society*. Kenitra, Morocco: Association for Computing Machinery, pp. 1–6. DOI: 10.1145/3314074.3314088.

[2] Adrian, K. and Gary, B. (2016). *Learning OpenCV 3: Computer Vision in C++ with the OpenCV Library*. 1st ed. O'Reilly Media. ISBN: 9781491937990.

[3] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Hauptbericht 2022 [Main report 2022]*. Report. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht (visited on 28 June 2024).

[4] Bal, A. and Palus, H. (2023). "Image Vignetting Correction Using a Deformable Radial Polynomial Model". In: *Sensors* 23(3), p. 1157. DOI: 10.3390/s23031157.

[5] Bayomi, N. and Fernandez, J. E. (2023). "Eyes in the Sky: Drones Applications in the Built Environment under Climate Change Challenges". In: *Drones* 7(10), p. 637. DOI: 10.3390/drones7100637.

[6] Bradski, G. (2000). "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* 4(11), pp. 120–125.

[7] Bundesministerium für Wohnen, Stadtentwicklung und Bauwesen [The German Federal Ministry for Housing, Urban Development and Building] (2023). *Gesetz für die Wärmeplanung und zur Dekarbonisierung der Wärmenetze [Law for heat planning and decarbonization of heat networks]*. Germany. Enacted on 17 November 2023, effective from 1 January 2024. URL: https://www.bmwsb.bund.de/SharedDocs/gesetzgebungsverfahren/DE/kommunale-waermeplanung.html.

[8] Chaverot, M., Carré, M., Jourlin, M., Bensrhair, A., and Grisel, R. (2023). "Improvement of small objects detection in thermal images". In: *Integrated Computer-Aided Engineering* 30(4), pp. 311–325. DOI: 10.3233/ICA-230715.

[9] Choi, H., Yun, J. P., Kim, B. J., Jang, H., and Kim, S. W. (2022). "Attention-Based Multimodal Image Feature Fusion Module for Transmission Line Detection". In: *IEEE Transactions on Industrial Informatics* 18(11), pp. 7686–7695. DOI: 10.1109/TII.2022.3147833.

[10] Debus, C., Piraud, M., Streit, A., and Götz, M. (2023). "Reporting electricity consumption is essential for sustainable AI". In: *Nature Machine Intelligence* 5, pp. 1176–1178. DOI: 10.1038/s42256-023-00750-1.

[11] Deng, G. (2011). "A Generalized Unsharp Masking Algorithm". In: *IEEE Transactions on Image Processing* 20(5), pp. 1249–1261. DOI: 10.1109/TIP.2010.2092441.

[12] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). "Imagenet: A large-scale hierarchical image database". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, pp. 248–255. DOI: 10.1109/CVPR.2009.5206848.

[13] Dong, R., Pan, X., and Li, F. (2019). "DenseU-Net-Based Semantic Segmentation of Small Objects in Urban Remote Sensing Images". In: *IEEE Access* 7, pp. 65347–65356. DOI: 10.1109/ACCESS.2019.2917952.

[14] Dutta, A. and Zisserman, A. (2019). "The VIA Annotation Software for Images, Audio and Video". In: *27th ACM International Conference on Multimedia*. Nice, France: Association for Computing Machinery, pp. 2276–2279. DOI: 10.1145/3343031.3350535.

[15] Gade, R. and Moeslund, T. B. (2014). "Thermal Cameras and Applications: A Survey". In: *Machine Vision and Applications* 25(1), pp. 245–262. DOI: 10.1007/s00138-013-0570-5.

[16] Gutiérrez Hermosillo Muriedas, J. P., Flügel, K., Debus, C., Obermaier, H., Streit, A., and Götz, M. (2023). "perun: Benchmarking Energy Consumption of High-Performance Computing Applications". In: *Euro-Par 2023: Parallel Processing*. Cham: Springer Nature Switzerland, pp. 17–31. DOI: 10.1007/978-3-031-39698-4_2.

[17] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., and Harada, T. (2017). "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes". In: *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. DOI: 10.1109/IROS.2017.8206396.

[18] He, Y., Deng, B., Wang, H., and Cheng, L. (2021). "Infrared machine vision and infrared thermography with deep learning: A review". In: *Infrared Physics & Technology* 116, p. 103754. DOI: 10.1016/j.infrared.2021.103754.

[19] Herrmann, C., Ruf, M., and Beyerer, J. (2018). "CNN-based thermal infrared person detection by domain adaptation". In: *Autonomous Systems: Sensors, Vehicles, Security, and the Internet of Everything*. Ed. by Dudzik, M. C. and Ricklin, J. C. Vol. 10643. Orlando, United States: SPIE, p. 8. DOI: 10.1117/12.2304400.

[20] Hou, Y., Volk, R., Chen, M., and Soibelman, L. (2021). "Fusing tie points' RGB and thermal information for mapping large areas based on aerial images: A study of fusion performance under different flight configurations and experimental conditions". In: *Automation in Construction* 124, p. 103554. DOI: 10.1016/j.autcon.2021.103554.

[21] Iakubovskii, P. (2019). *Segmentation Models*. https://github.com/qubvel/segmentation_models. Accessed 10 July 2024.

[22] Iglovikov, V., Mushinskiy, S., and Osin, V. (2017). *Satellite Imagery Feature Detection using Deep Convolutional Neural Network: A Kaggle Competition*. DOI: 10.48550/arXiv.1706.06169. arXiv: 1706.06169.

[23] Jiang, J., Zheng, H., Ji, X., Cheng, T., Tian, Y., Zhu, Y., Cao, W., Ehsani, R., and Yao, X. (2019). "Analysis and Evaluation of the Image Preprocessing Process of a Six-Band Multispectral Camera Mounted on an Unmanned Aerial Vehicle for Winter Wheat Monitoring". In: *Sensors* 19(3). DOI: 10.3390/s19030747.

[24] Kordecki, A., Palus, H., and Bal, A. (2016). "Practical vignetting correction method for digital camera with measurement of surface luminance distribution". In: *Signal, Image and Video Processing* 10, pp. 1417–1424. DOI: 10.1007/s11760-016-0941-2.

[25] Kütük, Z. and Algan, G. (2022). "Semantic Segmentation for Thermal Images: A Comparative Survey". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 286–295. DOI: 10.1109/CVPRW56347.2022.00043.

[26] Lin, T.-Y., Goyal, P., Girshick, R., He, K., and Dollár, P. (2018). "Focal Loss for Dense Object Detection". In: *IEEE International Conference on Computer Vision (ICCV)*. IEEE, pp. 2999–3007. DOI: 10.1109/ICCV.2017.324.

[27] Liu, J., Wang, X., Chen, M., Liu, S., Shao, Z., Zhou, X., and Liu, P. (2014). "Illumination and Contrast Balancing for Remote Sensing Images". In: *Remote Sensing* 6(2), pp. 1102–1123. DOI: 10.3390/rs6021102.

[28] Lv, J., Shen, Q., Lv, M., Li, Y., Shi, L., and Zhang, P. (2023). "Deep learning-based semantic segmentation of remote sensing images: a review". In: *Frontiers in Ecology and Evolution* 11. DOI: 10.3389/fevo.2023.1201125.

[29] Maharana, K., Mondal, S., and Nemade, B. (2022). "A review: Data pre-processing and data augmentation techniques". In: *Global Transitions Proceedings*. International Conference on Intelligent Engineering Approach(ICIEA-2022) 3(1), pp. 91–99. DOI: 10.1016/j.gltp.2022.04.020.

[30] Majidizadeh, A., Hasani, H., and Jafari, M. (2023). "Semantic Segmentation of UAV Images Based on U-Net in Urban Area". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* X-4/W1-2022, pp. 451–457. DOI: 10.5194/isprs-annals-X-4-W1-2022-451-2023.

[31] Mayer, Z., Kahn, J., Götz, M., Hou, Y., Beiersdörfer, T., Blumenröhr, N., Volk, R., Streit, A., and Schultmann, F. (2023). "Thermal Bridges on Building Rooftops". In: *Sci Data* 10(1), p. 268. DOI: 10.1038/s41597-023-02140-z.

[32] Mayer, Z., Kahn, J., Hou, Y., Götz, M., Volk, R., and Schultmann, F. (2023). "Deep learning approaches to building rooftop thermal bridge detection from aerial images". In: *Automation in Construction* 146, p. 104690. DOI: 10.1016/j.autcon.2022.104690.

[33] Mehta, D., Bagubali, A., Joseph, A. N., Kumar, V., Karar, V., and Poddar, S. (2017). "Radial distortion estimation using analytical technique". In: *2nd IEEE International Conference on Recent Trends in Electronics, Information & Communication Technology (RTEICT)*. IEEE, pp. 1587–1591. DOI: 10.1109/RTEICT.2017.8256866.

[34] Neupane, B., Horanont, T., and Aryal, J. (2021). "Deep Learning-Based Semantic Segmentation of Urban Features in Satellite Images: A Review and Meta-Analysis". In: *Remote Sensing* 13(4), p. 808. DOI: 10.3390/rs13040808.

[35] Parihar, A. S. and Singh, K. (2018). "A study on Retinex based method for image enhancement". In: *2nd International Conference on Inventive Systems and Control (ICISC)*. Coimbatore, India: IEEE, pp. 619–624. DOI: 10.1109/ICISC.2018.8398874.

[36] Robinson, Y. H., Vimal, S., Khari, M., Hernández, F. C. L., and Crespo, R. G. (2020). "Tree-based convolutional neural networks for object classification in segmented satellite images". In: *The International Journal of High Performance Computing Applications* 0(0), p. 1094342020945026. DOI: 10.1177/1094342020945026.

[37] Ronneberger, O., Fischer, P., and Brox, T. (2015). "U-Net: Convolutional Networks for Biomedical Image Segmentation". In: *Medical Image Computing and Computer-Assisted Intervention (MICCAI)*. Ed. by Navab, N., Hornegger, J., Wells, W. M., and Frangi, A. F. Cham, Switzerland: Springer, pp. 234–241. DOI: 10.1007/978-3-319-24574-4_28.

[38] Shivakumar, S. S., Rodrigues, N., Zhou, A., Miller, I. D., Kumar, V., and Taylor, C. J. (2020). "PST900: RGB-Thermal Calibration, Dataset and Segmentation Network". In: *2020 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 9441–9447. DOI: 10.1109/ICRA40945.2020.9196831.

[39] Song, K., Zhao, Y., Huang, L., Yan, Y., and Meng, Q. (2023). "RGB-T image analysis technology and application: A survey". In: *Engineering Applications of Artificial Intelligence* 120, p. 105919. DOI: 10.1016/j.engappai.2023.105919.

[40] Strat, S., Benoit, A., and Lambert, P. (2014). "Retina enhanced bag of words descriptors for video classification". In: *22nd European Signal Processing Conference (EUSIPCO)*. IEEE, pp. 1307–1311.

[41] SZ DJI Technology Co. Ltd. (2018). *Matrice 600 Pro*. Version 1.0. Product information. Accessed 28 June 2024. URL: https://www.dji.com/matrice600-pro.

[42] SZ DJI Technology Co. Ltd. (2018). *Zenmuse XT 2: User Manual*. Version 1.0. Product information. Accessed 28 June 2024. URL: https://www.dji.com/downloads/products/zenmuse-xt2.

[43] SZ DJI Technology Co. Ltd. (2020). *Matrice 300 RTK*. Version 1.0. Product information. Accessed 28 June 2024. URL: https://enterprise.dji.com/matrice-300/specs.

[44] Ulku, I., Barmpoutis, P., Stathaki, T., and Akagunduz, E. (2020). "Comparison of single channel indices for U-Net based segmentation of vegetation in satellite images". In: *12th International Conference on Machine Vision (ICMV)*. Vol. 11433. SPIE, p. 1143319. DOI: 10.1117/12.2556374.

[45] United Nations Environment Programme (UNEP) (2024). *Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector*. Tech. rep. Nairobi, Kenya: UNEP, Global Alliance for Building and Construction (GlobalABC). DOI: 10.59117/20.500.11822/45095.

[46] Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., Gouillart, E., and Yu, T. (2014). "scikit-image: image processing in Python". In: *PeerJ* 2, e453. DOI: 10.7717/peerj.453.

[47] Vollmer, E., König, S., Horstmann, V., Klug, L., Kahn, J., Volk, R., and Vogl, M. (2025). *Thermal Urban Feature Segmentation - Multispectral (RGB + Thermal) UAS-based images from Germany with annotations*. Zenodo. Version 1.0.0. DOI: 10.5281/zenodo.10814413.

[48] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[49] Wang, T.-s., Kim, G. T., Kim, M., and Jang, J. (2023). "Contrast Enhancement-Based Preprocessing Process to Improve Deep Learning Object Task Performance and Results". In: *Applied Sciences* 13(19), p. 10760. DOI: 10.3390/app131910760.

[50] Yoshimi, Y., Mine, Y., Ito, S., Takeda, S., Okazaki, S., Nakamoto, T., Nagasaki, T., Kakimoto, N., Murayama, T., and Tanimoto, K. (2023). "Image preprocessing with contrast-limited adaptive histogram equalization improves the segmentation performance of deep learning for the articular disk of the temporomandibular joint on magnetic resonance images". In: *Oral Surgery, Oral Medicine, Oral Pathology and Oral Radiology*. DOI: 10.1016/j.oooo.2023.01.016.

[51] Yuan, W. and Hua, W. (2022). "A Case Study of Vignetting Nonuniformity in UAV-Based Uncooled Thermal Cameras". In: *Drones* 6(12), p. 394. DOI: 10.3390/drones6120394.

[52] Zeng, X., Xu, J., and Gao, X. (2020). "A Potential Method for the Nonuniformity Correction and Noise Removal of Infrared Thermal Image". In: *Acta Physica Polonica A* 137, pp. 1055–1060. DOI: 10.12693/APhysPolA.137.1055.

[53] Zhang, Y., Sidibé, D., Morel, O., and Mériaudeau, F. (2021). "Deep multimodal fusion for semantic image segmentation: A survey". In: *Image and Vision Computing* 105, p. 104042. DOI: 10.1016/j.imavis.2020.104042.

[54] Zhou, M., Jin, K., Wang, S., Ye, J., and Qian, D. (2018). "Color Retinal Image Enhancement Based on Luminosity and Contrast Adjustment". In: *IEEE Transactions on Biomedical Engineering* 65(3), pp. 521–527. DOI: 10.1109/TBME.2017.2700627.

E Assessing the Economic Viability of Thermography-based Leak Detection for District Heating Systems

Abstract

District heating systems (DHSs) play a central role in Europe's sustainable heat transition, yet pipeline leaks threaten their efficiency and reliability. Thermography-based leak detection (TLD), particularly via remote sensing, has emerged as a promising non-invasive localisation approach, however its cost-effectiveness remains unexplored.

This study presents the first economic assessment of TLD for DHS pipeline leak detection based on newly collected empirical data centred around German-speaking countries. To this end, leak growth and associated ongoing and repair costs are modelled and a break-even analysis is performed to compare TLD to baseline operational scenarios. The results show that TLD can recover its costs within months after deployment, even under conservative assumptions, and consistently outperforms passive leak response strategies. Based on these findings, the study offers recommendations for integrating TLD into network monitoring strategies and contributes to the broader evaluation of innovative and cost-efficient leak detection methods for DHS operators.

Abbreviations

BEA break-even analysis

BEP break-even point

DHS district heating system

GSD ground sampling distance

HDPE high-density polyethylene

ILDS integrated leak detection system

L large

LD leak detection

M medium

PIP pre-insulated pipe

PIRP pre-insulated rigid pipe

PUR polyurethane rigid foam

S small

SA sensitivity analysis

TIR thermal infrared

TLD thermography-based leak detection

UAS unmanned aircraft system

XL very large

E.1 Introduction

E.1.1 Context

In light of the ongoing energy crisis, securing sustainable and resilient heat supply systems has become a central concern for policymakers and urban planners in Europe. Among the most promising technologies in this regard are district heating systems (DHSs), which provide thermal energy to buildings through centralised heat production and distribution networks. These systems, of which there are more than 19,000 in the EU alone [8], have evolved to offer efficient, sustainable, and cost-effective alternatives to decentralised, fossil-fuel-based heating [20]. In Denmark, where two-thirds of the population receive 89 % climate-neutral heat via DHSs [2], such networks are already forming the backbone of a low-carbon heating strategy.

However, maintaining DHS efficiency and reliability is vital, especially as the components age and deteriorate over time. In Germany, network losses have consistently ranged between 10 % to 14 % over the past two decades [2]. A key contributor to these losses are leaks in the underground transport pipelines, the most failure-prone component of DHSs [36, 39]. Left undetected, such leaks not only reduce efficiency but also risk escalating into severe infrastructural damage and high repair costs [9, 52]. With around 600,000 km of pipelines in operation worldwide [36] – a third of which lie within the EU [8, 19] – this issue presents a major challenge.

To address it, various leak detection (LD) approaches have been developed, including conventional tools like integrated leak detection systems (ILDSs) or novel techniques such as correlators, tracer gas, or thermography. Thermal infrared (TIR) imaging, especially when remotely sensed, enables the non-invasive, flexible localisation of heat anomalies that may indicate underground pipeline leaks [9, 45]. However, despite algorithmic and

technological advances in this field, the critical question of economic viability remains unanswered.

E.1.2 Related work

Thermography is mentioned as a promising and robust option for LD in DHS pipelines in various reviews, notably Zhou et al. [53] and Latif et al. [24]. Several studies have explored thermography-based leak detection (TLD) as a technical solution. Foundational research by Ljungberg and Rosengren [25] and Axelsson [5] demonstrated that underground leaks in DHSs lead to measurable hot-spots on the surface. More recent efforts have advanced the algorithmic analysis of TIR data, integrating machine learning and computer vision techniques to improve detection accuracy and reduce false positives to present operators with a viable list of leak candidates [6, 9, 15, 18, 40, 45, 46].

However, the economic dimension of this approach has received little attention. Existing work typically focusses on technical feasibility or data processing methods, with no assessment of operational costs or financial break-even points (BEPs). This may be attributed to the novelty of TLD, from which follows a lack of data with which to quantify the economy of all process steps and involved aspects – not just regarding the method itself, but also the economic impact of leaks on DHSs. So far, only Tuikka [42] provide an overview of LD methods used by Finnish DHS operators from a practical perspective to shed light on cost-effectiveness and reliability. However, they so do in a strictly qualitative manner and without a specific focus on TLD.

E.1.3 Objectives and contributions

This study addresses the research gap by providing the first economic assessment of TLD for pipeline leak detection in DHS. Focusing on the real-world use case of German-speaking countries, this work makes several novel contributions:

- We introduce and evaluate newly collected empirical data from the most extensively developed area of DHSs in Europe [8].
- Exemplary DHS pipeline leak growth and associated ongoing and repair costs are modelled for the first time, offering insight into long-term impacts of leaks.
- The practical cost for TLD is estimated, with which a first-of-its-kind break-even analysis (BEA) is performed. With it, key economic thresholds for TLD viability are established.
- Given the findings of both the empirical study and BEA, recommendations are provided for use of TLD and LD in general to help DHS operators in their decision making.

By bridging the gap between technical potential and economic feasibility, this work contributes to the informed adoption of innovative TLD techniques that can improve the sustainability and reliability of DHSs operation. To this end, the paper is organised as follows: After Section E.2 provides a methodological context, Section E.3 evaluates the empirical study. The analysis is described in Section E.4, while recommendations are derived in Section E.5. The study concludes with Section E.6.

E.2 Foundations and methodological context

E.2.1 District heating

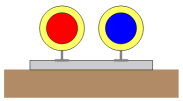
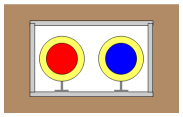
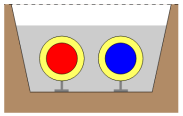
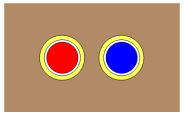
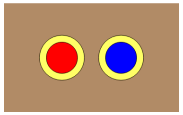
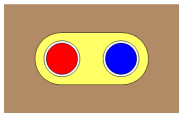
DHS development can be categorised into four successive generations. The first commercial systems emerged in the late 19th century using steam as a heat carrier, with temperatures exceeding 200 °C. The second generation in the 1920s shifted to pressurised hot water as a more efficient and less hazardous option, while the third reduced distribution temperatures further and introduced diverse heat sources in response to the oil crises of the 1970s. A shift to the fourth generation is currently underway, mainly defined by the integration of renewables and minimisation of losses. [28, 48]

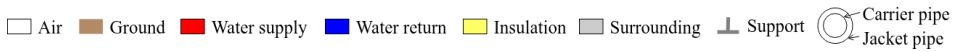
The technical improvement of DHSs throughout the decades is marked by the development of materials that allowed for cost-effective construction. This is particularly true for the pipelines themselves, which constitute the lion's share of network installation costs [31]. Table E.1 presents the different types of pipelines used in DHSs, classified by installation method. While early generation pipes were placed within concrete ducts, the advent of ground-installed pre-insulated pipes (PIPs) helped overcome the cost barrier and enabled the transition to the third generation [51]. Initial implementations with various casing materials, including fibre or asbestos cement, gave way to plastic, specifically PUR as an insulation material and high-density polyethylene (HDPE) for casings. Combined with steel carriers, composite pre-insulated rigid pipes (PIRPs) have become the most widely installed pipeline systems today [14].

Nowadays, DHSs are built around five primary heat sources: combined heat and power, industrial excess heat, waste-to-energy, geothermal, and biomass [35]. The hot water medium is commonly chemically treated[1] to counteract internal corrosion. Pumps circulate the transport medium to users through pipelines and the heat is transferred to the connected buildings either directly or through exchangers [51]. Most commonly, the cooled medium returns through a parallel network, making it a two-pipe (supply and return) system [31]. Having often been installed several decades ago, most systems today are third generation and require modernisation [37].

[1] Treatments remove solids, oxygen, salts, and stabilise pH and hardness [4, 51].

Table E.1: Installation methods and pipeline types (based on Heating [14], Heipke and Tödter [15], Roscher [38], Winkens [49], and Woods [51]).

Installation		Gen	Visualisation	Advantages	Disadvantages	Usage	Carrier & insulation	Casing	Surroundings
Above ground		2, 3		- economical installation, control and repair	- design requirements - exposure to weather and vandalism	Early pipeline type. Mainly implemented for industrial and commercial sites.	Steel / copper with thermal insulation	Metal / plastic	Concrete / steel supports or suspensions
Underground duct-based		1, 2		- robust - long life time	- expensive installation and repair - large construction areas	Earliest pipeline type. Nowadays only installed when ductless is not possible or for combined solutions with other pipes.	Metals with thermal insulation	Metal	Concrete or brick
Underground ductless	Pour-in-place	2		- low installation cost compared to ducts - robust when no installation errors	- more expensive installation than PIP	Intermediary pipeline type. Nowadays outdated in favour of PIP.	Steel	-	- powders with water-resistant additives - bituminous / cement-bonded block insulation
	PIP – Glide	3		- very high temperature and pressure resistance - lower heat losses due to vacuum	- more expensive than rigid systems	Special application for higher stress conditions (i.e. river culverts, groundwater areas)	Steel with vacuum and mineral fibre	Steel	Ground
	PIP – Rigid	3		- high temperature and pressure resistance - low installation cost and time	- resistance not as high as for glide systems	Nowadays most prevalent pipeline type.	Steel with PUR *Dated: ductile cast iron*	Plastic (HDPE) *Dated: fibre / asbestos cement*	Ground
	PIP – Flexible	3		- self-compensating - ideal for smaller pipe diameters	- plastic version has limited temperature and pressure resistance	Nowadays most prevalent type for smaller pipe diameters	Cross-linked polyethylene / polybutene or steel / copper with PUR	Plastic (HDPE)	Ground
	PIP – Twin	3		- lower installation cost (smaller trench width) - lower heat loss	- precise routing required, complex routing is costly - only for smaller pipe diameters	Special applications (i.e. house to house connect-ions). Alternative implementation to all PIP single-tube variants.	Steel / PEX / PB with PUR	Plastic (HDPE)	Ground



With more than 34,000 km of trench length, Germany not only has the largest share of DHS pipelines in Europe but also records the highest growth [8]. Historically speaking, DHSs have been prevalent there since the 1920s as the first in all of Europe [48]. This means that their networks consist of a variety of generations and all types of installation methods shown in Table E.1 are represented. Here, too, PIRPs are the most common pipelines [22] and this trend towards newer generations is reflected in the distribution of heat carriers given today (see Figure E.1) [2]. All these aspects make Germany an interesting use case for the study of leaks and their effect on DHS operations.

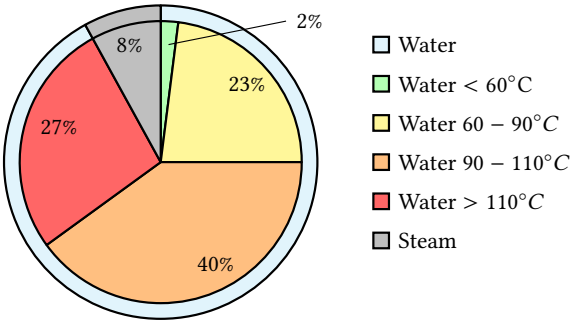


Figure E.1: Distribution of carrier characteristics in German DHSs in 2021 [2].

E.2.2 The need for leak detection

As providers of essential urban heating, DHSs form part of critical infrastructure [30]. Failures can cause severe disruptions and safety risks, as highlighted by outages in Lithuania (2006) [55], Harbin (2020) [26], and even deadly incidents such as in Zhengzhou (2021) [30].

Over time, DHSs face degradation owing to their extreme operating conditions, which include high temperatures, pressure, and humidity [30]. Of all components, pipelines have been found to be the most vulnerable to damage [16, 27]. Four years of data from Heilongjiang province, China, show that 56 % of faults originate in pipelines – over 2.3 and 5.5 times more than those linked to valves and compensators, respectively [39]. These failures are attributed to both internal and external causes, including corrosion, installation errors, or environmental exposure [16, 36, 43]. Among these, external corrosion is most commonly identified as the leading cause [33, 55][2].

If not caused by accidental external damage, pipe failures are usually preceded by micro-damage [50]. The presence of leaks is therefore indicated by the increasing need for water input, so-called makeup water [43]. Even small leaks can cause severe financial losses due to the requirement to chemically and thermally treat additional water and reduction in efficiency [50, 53]. Early detection is therefore crucial for reducing operational costs and avoiding major disruptions later on. Given the uneven wear across systems [33],

[2] In Yekaterinburg, prolonged water exposure caused 80 % of pipe ruptures [33].

regular, data-driven assessments are essential for prioritising maintenance [50]. To this end, effective LD tools are key to ensuring safe, efficient, and sustainable operations.

E.2.3 Methods for leak detection

A wide range of methods and tools exist for LD in DHS pipelines. They differ greatly in terms of methodology, technology readiness level, commercial availability, and practical application. LD systems can be dynamic, i.e. the technology involved consists of mobile devices, or static, when sensors are fixed in one place [24]. While various publications detail the many methods in existence [3, 13, 14, 22, 24, 51, 52, 53], the following subsections describe select methods of interest, in particular those of relevance to German DHSs.

E.2.3.1 Operational changes

To confirm the existence of a leak, DHS operators commonly look out for initial indicators from static, in-built sensors. These are focused around changes in operational procedure, such as consistently increasing water loss and pressure drops [3, 13, 51]. Leaks can thereby be determined for the network as a whole or specific subsections via bypass pipes around isolation valves [3]. However, as no precise localisation can be performed using such indicators, it is prudent to evaluate the operational changes as best as possible before proceeding to more exact LD methods [3].

E.2.3.2 Visual and mechanical-technological methods

Expert visual inspections of DHSs pipelines can help assess the general condition and thus identify early-stage defects or material degradation. This is particularly relevant for older or complex sections, such as duct-based, above-ground, and basement pipes [3]. Methods range from simple visual checks of accessible pipes, over environmental indicators (snow melts above PIPs [13] or steam rising from nearby manholes), to advanced techniques.

More complex tools include dynamic camera-equipped inspection crawlers as well as ultrasonic wall-thickness measurement devices and moisture sensors [3]. Inspection machines like the telescopic crawler or Crawler-Eye enable a moving, visual assessment of non-walkable duct-based pipes via onboard cameras [3]. They allow for a systematic documentation and localisation of faults, but are limited by cable drag, physical obstructions in the inspection path, and data storage requirements for the image data [24]. Ultrasonic wall-thickness measurement, for instance via impact echo, can be applied to in-service pipes if at least one side is directly accessible [24], providing accurate information on material degradation or corrosion – although it requires proper calibration and trained personnel [3]. Moisture meters offer a simple, non-destructive means to verify suspected leaks in PIRP after a pipe has been laid bare [3].

E.2.3.3 Integrated systems

Newer pipeline types, specifically PIRPs, allow for the installation of ILDS. These consist of a pair of wires within the PUR-based thermal insulation layer with which moisture penetration can be measured and thus leaks detected [14, 51]. Moisture ingress alters the resistance between the wires and pipes, enabling precise localisation and the possibility for complete pipeline monitoring [3].

ILDS commonly rely on one of two measurement techniques. The resistance comparison method operates on the principle of an unloaded voltage divider: moisture-induced changes in resistance alter the voltage applied to the monitoring wire. This delta is used to detect and localise faults by comparing resistance values along the pipeline. While effective, the method can only provide averaged results when multiple faults occur, meaning the first leak must be repaired before the next can be located. In contrast, the impulse runtime method sends high-frequency pulses along the monitoring wire. These are partially reflected at points of altered resistance – such as moisture ingress or wire breaks – and the time taken for the reflections to return is used to calculate the location. This technique offers high spatial resolution and can accurately identify and locate multiple faults, although higher moisture levels are necessary for localisation than for the resistance method. [14, 21]

The utilised technique depends on the type of implemented ILDS, of which there are several variants. The most common is the Nordic system [13], which is also the German standard [14, 22]. Two bare or tinned copper wires are connected in monitoring loops via soldered joints and junction sleeves [21]. The low wiring material cost is a key reason for its popularity [14]. LD is commonly achieved with the impulse runtime method [21]. The second most common method is the Brandes system, which features a NiCr sensor wire with perforated insulation and a return copper wire connected in continuous monitoring loops [14]. It relies on the resistance comparison method [21]. Other less popular variants include HDW and ISOTRONIC, which use different types of twisted copper wires and the impulse runtime method [21].

Regardless of wiring system and measurement technique, all ILDSs can be controlled in one of three ways: central, decentral, or manual [14]. Respectively, this means that either all monitoring circuit information is collected centrally, on-site monitoring devices need to be scanned regularly, or portable equipment is required for inspections at scheduled intervals [14].

E.2.3.4 Thermography

A non-invasive, pipeline-agnostic method for detecting leaks is TLDs [3]. When subsurface leaks occur, the heat of the medium diffuses through the surrounding soil, causing an increase of surface temperature [9]. These hot-spots can be identified in TIR images captured by thermographic cameras. Though comparatively recent, TLD is being increasingly adopted for DHSs leak detection [3, 24, 36]. Existing empirical studies such as Tuikka [42], for instance, already report the active implementation of TLD by Finnish DHS companies.

One such operator noted a decrease in the required reliance on external notifications by the public from over 50 % to around 10 % to 15 % owing to this technology [42], thereby showcasing the usefulness of the method in practice.

TLD can be conducted from the ground using handheld or vehicle-mounted cameras [3], or from the air via unmanned aircraft systems (UASs), helicopters, or fixed-wing aircrafts [13, 54]. Aerial surveys enable rapid, large-scale coverage of urban DHSs without the extensive resources and effort required for manned inspections [13, 53]. UAS-based acquisition is particularly advantageous owing to its flexibility, comparatively low cost, and ability to capture high resolution images with a large ground sampling distance (GSD) [46]. General disadvantages of TLD include its dependency on complementary software or expensive expert viewing to interpret the large amounts of resulting data, as well as its sensitivity to external conditions [24, 36, 46]. The latter means specific acquisition conditions must be met to collect data of sufficient quality (see Appendix E.I) [46], a fact that is amplified by the lower resolution of TIRs compared to standard colour images [44].

E.2.3.5 Other methods

Further methods implemented for LD in DHS pipelines include, but are not limited to, acoustic and tracer-based techniques.

Vibrations induced by leaks in water pipes generate acoustic signals – typically hissing noises – that can be detected using various techniques [24]. One of these is correlation analysis, a non-destructive method that uses synchronised acoustic sensors to calculate leak locations based on the time delay between signals [3, 24]. It requires active pipeline use, a confined search area, and at least two experts to operate the measuring system [3]. Small leaks, changing operating conditions, different pipe materials, and in particular background noise can greatly impede performance [24].

Tracer-based methods involve adding a detectable substance to the heated medium, which escapes through leak points and can then be identified above ground. Helium, a innocuous and inert gas, can be dissolved into the DHS water and diffuses out through leaks to the surface, where it can be detected by sensitive measurement instruments [3, 13, 52]. Leaks can be effectively detected across all pipe types this way, though the search space needs to be limited to restrict the gas input and soil and weather conditions need consideration [13]. Similarly, the fluorescent and biologically safe dye Uranin can be used to colour the DHS's water to make leaks visible. This is particularly useful for sections where a leak may connect DHS and drinking water systems, although dye removal can take weeks or months [3, 13]. Both methods are effective without disrupting active operation but require careful preparation, regulatory approval, and post-detection flushing or neutralisation [13].

E.2.4 Motivation for methodological focus on TLD

Of the various LD techniques, this study's focus lies on TLD for several reasons. First and foremost, even without the latest advances in automatic image analysis, it is recommended

as a viable method by various handbooks [13, 14, 43] as well as the German Consortium on District Heating [3]. Compared to more proactive static methods, such as ILDS, TLD is not just limited to PIRPs and not wholly reliant upon initial correct installation, but entirely pipeline-agnostic[3] [3]. Where most visual and mechanical-technological methods require direct pipeline contact, the thermographic approach is entirely non-invasive and – via remote sensing – able to reach manually inaccessible areas. Given recent developments in UAS technology and automatic image analysis, the manual and organisational effort associated with TLD is reduced, particularly compared to acoustic and tracer-based methods. While this study's focus lies solely on leak detection and localisation, German Working Committee on District Heating (AGFW) [3] also highlight TLD's versatility for condition monitoring – and even quality assurance during installation for some pipeline types. Given this combination of aspects and the fact that TLD is of yet lacking any form of economic assessment, this scalable and flexible method is selected as the focus of our study.

E.3 Empirical study

A central part of this study lay in conducting a series of interviews to gain insight into DHS management as well as the impact and handling of leaks in practice. To this end, 35 network operators from DHSs of various sizes and locations were contacted, of which nine – eight from Germany and one from Switzerland – agreed to participate. The interviews were conducted between January and July 2025 and consisted of a list of survey questions which can be found in Appendix E.II. While the responses varied greatly in detail as well as quantity, this chapter summarises the supplied information as best as possible. Where none was provided, slashes ("/") are used to indicate missing data. For the purpose of anonymity, a Roman numeral from *I* to *IX* is associated with each DHS and used to reference each where they are sources of specific information.

To get an overall picture of the currently implemented approaches for DHS LD management, an interview was also conducted with a representative of LANCIER monitoring [23]. The company specialises in infrastructure monitoring worldwide and develops ILDS for DHS in particular for the German-speaking market [G. Roehl, personal communication, May 8, 2025].

E.3.1 General overview

DHSs in German-speaking countries vary greatly in size – from small local networks of just a few kilometres [G. Roehl, personal communication, May 8, 2025] to some of the largest in the European Union, exceeding 2.000 km of pipeline [11]. To balance the need

[3] While leaks in duct-based pipes will not cause the heated medium to spread to the surface, they are commonly caused by breaks in the ducts themselves which can be seen in TIRs. German Working Committee on District Heating (AGFW) [3] presumably highlight its pipe-agnostic applicability for this reason.

for generalisability with an accurate portrayal of network diversity, this study's DHSs are categorised into four groups according to their pipeline grid length. This has the additional benefit of preserving participant anonymity. The four-tier categorisation is defined as follows: 1. small (S) ($L_{\mathrm{DHS}} < 30km$), 2. medium (M) ($30km < L_{\mathrm{DHS}} < 100km$), 3. large (L) ($100km < L_{\mathrm{DHS}} < 300km$), 4. very large (XL) ($300km < L_{\mathrm{DHS}}$).

An overview of general network characteristics is provided in Table E.2. Each category is represented by at least two DHSs, for which averages or ranges are provided. Although some initially began as steam-based, all participating networks use hot water as a heat carrier, as is common to DHSs nowadays (see Section E.2.1). Various pipeline diameters are installed, ranging from $DN20$ to $DN500$ [*VIII*].

Table E.2: Characteristics of participating networks, generalised through categorisation into four groups according to their pipeline lengths.

Characteristic				Unit	S	M	L	XL
DHS grid length				km	< 30	$30 - 100$	$100 - 300$	> 300
Heat production				GWh / yr	$20 - 45$	$75 - 85$	$300 - 850$	> 900
Connected households				-	$50 - 400$	$1000 - 5000$	$20,000 - 50,000$	$> 50,000$
Share heat demand covered by DHS				%	$12 - 13$	$13 - 40$	$12 - 25$	$30 - 90$
Avg max flow temperature (as per Fig. E.1)				$°C$	$90 - 110$ (~ 100)	> 110 (~ 120)	> 110 (~ 125)	> 110 (~ 132)
Installation method	Above ground			%	0	0	<1	<1
	Underground duct-based			%	0	0	18	20
	Underground ductless	Pour-in-place		%	0	0	<1	<1
		PIP	Glide	%	0	0	<1	<1
			Rigid	%	95	94	70	51
			Rigid (dated)	%	0	0	4	<1
			Flexible	%	0	0	<1	5
			Twin	%	0	6	0	<1
	Cellar pipes			%	0	0	5	8
	Unknown			%	5	0	2	11

As expected, the heat produced for end use and the number of connected households are closely related to network size. Similarly, maximum flow temperature increases with pipeline length, although – with most temperatures above 110 °C – the participating networks are generally on the higher end of the spectrum of German DHSs (see Fig. E.1). Given these circumstances, the interviewed network operators may handle potentially more challenging conditions than on average. To better characterise each category, the percentages of pipeline types were averaged to match the installation methods introduced in Table E.1. When it comes to the types of installed pipelines, network size can be seen as an indicator of diversity. Larger networks show much greater variety than smaller ones due to their greater age, as demonstrated by the existence of duct-based systems and other outdated types. S to M DHSs, on the other hand, consist almost exclusively of the newer PIRPs in single or twin form (see Table E.1).

E.3.2 Leak occurrences

Table E.3 gives an overview of leaks and the impact of their occurrence in DHSs. Network losses vary greatly between the different DHSs, though they generally reflect the 10 % identified by German Working Committee on District Heating (AGFW) [2] as an average for German DHSs. Aside from heat losses [*IV*], leaks constitute a considerable part of these [*I, V, VI*]. As may be expected, the number of leaks per year increases with network size. Although observed leak rate ranges can be calculated using such values (as shown in Table E.3), these suffer from detection bias as not all leaks are found and repaired. This problem is particularly prevalent in larger networks, as highlighted by the standard losses some operators were able to provide. In these instances, operators either do not have the resources available to locate the remaining leaks, are currently not able to repair them (see Section E.3.3), or do not believe the efficiency increase to outweigh the localisation and repair effort. It may therefore be assumed that true leak rates for larger networks are higher than those in Table E.3. Beyond typical loss levels, DHS operators report leaks to cause more than a quadrupling in losses. Even a XL network mentions more than double their already heightened baseline losses to occur before LD and repairs are initiated.

Table E.3: Characteristics related to leak occurrences of participating networks.

Characteristic	Unit	S	M	L	XL
Network losses	%	8 – 16	12	12 – 25	10 – 19
Standard losses	m^3/day	/	6	96	480
High losses	m^3/day	/	25	/	960
Repaired leaks	$\#/yr$	≤ 1	2 – 4	9 – 24	60 – 140
Leak rate ranges	$\#/km$	0.02 – 0.1	0.05 – 0.1	0.04 – 0.1	0.07 – 0.1
Avg. leak rate	$\#/km$	0.06	0.09	0.09	0.09

Pipeline leaks generally start as hairline cracks before growing to greater sizes [50]. According to practical findings from Fuchs and Frommhold [10], these can range from 1.2 m^3/day to 85 m^3/day. Here, hairline cracks are estimated to cause about 2 m^3/day to 3 m^3/day [*V*], while larger leaks may incur 10 m^3/day [*VI*] or grow further to near 20 m^3/h [*I*]. A leak can culminate in a rupture of 1000 m^3/h [*IX*].

E.3.3 Leak detection

As hinted at in the previous section, DHS losses vary greatly depending on the implemented monitoring strategy, which tends to align with size-based category. Table E.4 summarises the different LD methods used by the interviewed operators, demonstrating commonalities and dissimilarities.

The first indicator of system leaks are changes in operational parameters, such as pressure drops and the requirement for make-up water. While these are monitored by all DHS operators, they cannot reveal the concrete location of a leak. For that purpose, other methods are listed as having been implemented, each with a varying degree of emphasis and priority. For example, several operators stated using calls from citizens as an alert and means of

Table E.4: Characteristics related to LD of participating networks.

Characteristic			Unit	S	M	L	XL
LD methods	Conventional	operational change	-	✓	✓	✓	✓
		public notification	-	✓		✓	✓
		visual checks	-		✓	✓	✓
		ILDS	-			✓	✓
	Novel	TLD	-	✓	✓	✓	✓
		acoustic	-		✓		✓
		tracer-based	-			✓	
ILDS	Installation		%	98	100	58	29
	Type	Nordic	-	✓	✓	✓	✓
		Brandes	-	✓		✓	✓
		other	-			✓	
	Usage	not used	-	✓	✓	✓	✓
		decentral / manual	-			✓	✓
		central	-			✓	
Leak duration	days		-	✓	✓	✓	✓
	weeks		-	✓	✓	✓	✓
	months		-		✓	✓	✓
	years		-		✓		

locating leaks. However, a XL network is the only one to indicate that this is the main localisation method they rely upon. Only in the rare cases where no notification is provided and losses are exceedingly high is an active search initiated. While another XL operator describes a much more stringent approach for their smaller secondary networks, they face similar problems in their primary which services half the city. While DHSs of other categories also mention having made use of public notifications, they are generally able to follow a more proactive maintenance strategy by trying to locate leaks of smaller sizes throughout their systems. To this end, visual checks[4] in areas of interest are implemented across the board. Such methods may be useful for sporadic leak localisation, but are not viable as holistic monitoring approaches due to the involved effort. The popularity of their usage highlights a surprising aspect of LD in practice.

Given the prevalence of PIRPs indicated in Table E.2, one would expect there to be almost no need for manual methods. In practice, however, the integrated monitoring systems are put to surprisingly little use – paradoxically in particular in S and M networks where their prevalence is highest. The reason provided for this is simple: the operators do not have the (full) circuit diagrams of the monitoring wires and thus lack the capability to interpret the measured results as intended [*I, III, V, VI, VIII*]. Even when plans exist, their usage can still be hampered by wiring errors that occurred during installation [*II, III*] or a lack of personnel [*VI, XI*]. These observations are not just limited to the interviewed DHSs – carelessness during pipeline installation has caused a disuse of integrated monitoring

[4] This includes inspecting the ground above pipelines under suspicion to see if they are dry and warm [*V*] or checking for steam rising from manhole covers, which can be an indicator of leaked DHS water heating the sewage system [*I, II*].

throughout German-speaking countries[5] [G. Roehl, personal communication, May 8, 2025]. However, this is not true for all networks. *IV*, for instance, reported restoring their ILDS recently so that it has now become their main LD and localisation method.

Such examples bring to light a new trend that is mirrored in Table E.4, namely a recent focus on monitoring and LD in general. As implied by the previous descriptions and listed repair time frames, several operators show a more laissez-faire approach to leak repair. This contradiction in terms of business economics may be attributed to the fact that many networks, especially S and M ones, originated as a means of using readily available excess heat (such as from incineration plants or production processes) and were not intended or managed as a regular business [*VIII*]. While many operators did not look beyond function fulfilment and network expansion, the energy crisis has placed a new spotlight on DHSs as a means to provide climate neutral heat, causing a change in mindset and interest in a more economical operation [G. Roehl, personal communication, May 8, 2025]. Given these circumstances, various network operators have reported trying alternative methods, first and foremost among these thermography. This was implemented in various ways, including via aeroplane [*I, IV, XI*], UAS [*II, III, VII, XI*], vehicle-mounted [*XI*], or hand-held sensor [*IV, VI, XI*] – all of which were successful in discerning leaks and their locations. *XI* in particular report having incorporated all forms of thermography into their network monitoring strategy. Information obtained through TIR acquisition flights is mentioned as particularly useful, having helped identify leaks that accumulated to 40 % of network losses [*XI*].

Other reported methods include acoustic techniques, such as the unsuccessful use of correlators by an M network. One XL DHS reported using an ultrasonic listening probe, though they stressed the sensor's ineffectiveness in urban environments given interfering ambient noise. Another XL network reported the the continuous use of Uranin-based tracer dye, which is generally helpful to differentiate leaks when they occur at construction sites or near water distribution networks.

E.3.4 Causes for leaks

As discussed in Section E.2, the reasons why leaks occur are varied, with corrosion mentioned as the main factor by many studies in literature. Thanks to a detailed breakdown provided by *II* for the years 2019 to 2024, Table E.5 can give an overview of the prevalent causes and most affected pipeline types in an exemplary German DHS consisting of various different pipeline types. Interestingly, corrosion is only the second most common cause, with installation and manufacturing errors being the source of more than half of all repaired leaks. Among these, welding seam defects are most prevalent.

[5] The reasons for this are two-fold according to G. Roehl [personal communication, May 8, 2025]. Firstly, if circuit diagrams are not explicitly listed in the installation Statement of Work, they simply are not provided. Reproducing them post-construction is possible, but comes with massive effort and cost. Secondly, very few companies document the circuits with enough precision that they can be used reliably. In Germany, experience has shown only a single company to provide plans of sufficient detail.

Table E.5: Causes for leaks with respect to pipeline types (based on data from *II*). All values are given in %.

Component				Share of pipe length [%]	Cause [%]													SUM [%]
					Wear and Ageing		Corrosion			Installation and Manufacturing Errors					External		Unknown	
					Defective component	Leaky component	External corrosion	Crevice corrosion	Steel pipe corrosion	Welding seam defect	Incorrect insulation	Installation error	ILDS defect	Faulty ILDS reading	Construction Damage	Other Damage		
Installation method	Above ground			1.7	0	0	0	0	0	0	0	0	0	0	0	1.1	0	1.1
	Underground in ducts			12.3	0	1.1	2.1	0	0	0	0	0	0	0	0	0	1.1	4.3
	Underground ductless	Pour-in-place		0.3	0	0	1.1	0	0	0	0	0	0	0	0	0	0	1.1
		PIP	Glide	0.4	1.1	0	5.3	1.1	0	1.1	0	0	0	0	0	0	0	8.6
			Rigid	63.9	1.1	3.2	3.2	0	0	38.3	2.1	3.2	6.4	2.1	4.2	1.1	1.1	66.0
			Rigid *(dated)*	11.2	0	2.1	4.2	0	0	0	0	0	0	0	0	0	0	6.3
			Flexible	0.4	0	0	0	0	0	0	0	0	0	0	1.1	0	0	1.1
			Twin	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	Cellar pipes			5.1	0	0	0	0	1.1	1.1	0	0	0	0	1.1	0	0	3.3
	Unknown			4.6	2.1	0	2.1	0	0	0	0	0	0	0	0	0	0	4.2
Other components (flange, compensator)					0	2.1	0	0	0	0	0	1.1	0	0	0	1.1	0	4.3
SUM					4.3	8.5	18.0	1.1	1.1	40.5	2.1	4.3	6.4	2.1	6.4	3.3	2.2	
SUM (by category)					12.8		20.2			55.4					9.7		2.2	

A closer look reveals the source of this to be the PIRPs, which seem disproportionately affected[6]. While 66 % of leaks throughout the provided 6 year time frame occurred in these most common of pipelines, 79 % of them were caused by installation and manufacturing errors. This observation coincides with reports made by other interviewed DHS operators that mention such errors as a dominant reason for repairs to become necessary long before the officially estimated end of lifetime[7] [*I, III, IV, V*]. In earlier years, this was attributed to underdeveloped sleeve technology and substandard steel quality from the postwar years, although these aspects have become less problematic since the 1990s [*I, IV*]. Current errors are assumed to be the result of construction defects and inadequate site supervision [*II, III, V*]. Incidentally, *II* report up to 40 % of annual leaks (to which such errors are counted) still occur within the warranty, customarily a period of 5 years [G. Roehl, personal communication, May 8, 2025]. Other operators such as *VI* also describe the occurrences of multiple leaks annually starting just over 10 years after initial installation. It may be for these reasons that operators such as *VII* go as far as to refute the correlation between pipeline age and occurrence entirely. The fact that these circumstances can be observed across all interviewed DHS operators further underlines the need for active LD.

[6] *II* report leaks in these pipes to make up between 58 % to 92 % of yearly occurrences.
[7] The technical service life of PIRPs in Germany is estimated at 36 yrs to 70 yrs, on average 51 yrs [1].

Table E.5 also shows other interesting aspects. External damage causes a comparatively small amount of leaks, about 10 % throughout the regarded time frame. Older systems, such as duct-based and dated PIRP installations are significantly less fault-prone than the previously discussed newer types. Leaks in these pipelines are caused only by corrosion or otherwise wear. This coincides with observations made by other operators, who attest to their general reliability as long as certain conditions are upheld – in particular ensuring the pipe surroundings remain dry [*III, IV*].

With regard to other installation methods not represented in *II*, twin pipes are reported by *V* to be particularly fault-prone. The high flow temperatures are found to induce mechanical stresses that lead to the crack formation and growth, a process often exacerbated by production errors [*V*]. This practical experience contradicts literature such as Mazhar et al. [31], which suggest "to use twin-pipes wherever applicable since their performance to price ratio is the best".

Aside from these pipeline type-related aspects, some general observations regarding DHSs and leak causes should be mentioned. While the main focus often lies on the heating period and operation at high temperatures, difficulties may arise owing to the longer periods of colder flow temperatures. It can cause damage to the installed press technology or sleeves to loosen given the shear forces that occur when the pipe contracts [*V*]. Maintenance is also a key aspect to prevent leak-enabling circumstances. Without it, draining and venting can be compromised, shut-off valves clogged, or salt and water enter the shafts [*V*]. This, again, emphasises the necessity for active LD.

E.4 Economic analysis

E.4.1 Estimating leak costs

The cost incurred by DHS leaks can fundamentally be divided into two parts: 1. ongoing, namely the cost of treated and heated make-up water required to compensate for the incurred loss, and 2. one-time, namely the costs of repair. As highlighted in Section E.2.1, the longer a pipeline leak remains undetected, the greater the risk of rupture and destruction of surrounding infrastructure [50], a result of which would be very high repair costs. Economically speaking, an earlier detection has the benefit of minimising the considerable ongoing costs DHS operators have to bear while a leak persists [50]. However, both types of cost are difficult to quantify given their dependence on the unique circumstances of each DHS. Nevertheless, this study makes an effort to estimate them given a combination of theoretical sources and practical knowledge from the conducted surveys described in Section E.3.

E.4.1.1 Ongoing costs

Leaks incur ongoing costs chiefly due to the make-up water that must be fed into the system to counteract the ensuing loss. This mainly stems from the required heating and treatment of the water before it can enter the pipelines [50]. The latter is essential to prevent internal corrosion and includes the removal of particulate matter, hardness minerals, salts, and oxygen, as well as addition of chemicals to raise the pH level, stabilise hardness, and prevent particle coalescence [14, 51]. The addition of make-up water is outsourced to the DHS water carrier providers, meaning that operators pay according to their individual contracts and circumstances [V]. Most interviewed operators were unable to provide costs per m^3, indicating that these expenses are not taken note of as carefully as one might expect. *IX* mention 7 €/m^3 as a value from another German DHS, while *V* provide an estimate of 5 €/m^3 for their network with 125 °C flow temperature. *VIII* report a precise value 5.20 CHF/m^3, equivalent to approximately 5.60 €/m^3, for their 20 °C colder network. Following a conservative approach, 5 € is assumed as the cost $C_{make-up}$ incurred per m^3 of make-up water.

The amount of required water depends on the size of the leak, which in turn depends on the duration of its existence (see Section E.2.2). As discussed in Section E.3.3, size varies greatly depending on all manner of influencing factors. However, for the purpose of this study, a simplification must be made to quantify ongoing monetary losses. To this end, an exemplary leak and its growth are modelled. While there is a significant lack of literature in this field for DHS pipelines, leak growth models have been developed for those of water distribution networks, which share the same medium and similar characteristics. Specifically, Guo et al. [12] compared different growth functions, including Logistic, Gompertz, and Richards, to identify which is most adept at simulating a leak's growth based on a real leak dataset. They find that the Richards function performs best, which is generally defined as

$$f(t) = \frac{a}{\left(1 + e^{(b-c\cdot t)}\right)^{\frac{1}{N}}} \tag{E.1}$$

where $f(t)$ is the leak flow rate (m^3/h), t defines the leak duration, and a is equal to the maximum leak flow rate. The other parameters – b, c, and N – are tuned to best simulate the real-world dataset. Guo et al. [12] performs these optimisations for diameters $DN100$ to $> DN400$ and different pipe materials such as cast iron, cement, and steel. The growth coefficient, c, is found to be best defined as 0.20 throughout. The position of the inflection point, b, and the curve steepness, N, are defined by the diameter: $b = 3$ and $N \approx 0.8$ for $\leq DN300$ and $b = 1$ and $N \approx 0.2$ for $> DN300$. a generally grows with diameter and is smaller for steel pipes than cast iron. [12]

An exemplary leak is modelled on the basis of the theoretical and practical information described in Section E.3. To this end, Guo et al. [12]'s function definition is adapted to the circumstances of this study and to account for the differences between water distribution

networks and the systems at hand[8]. To better suit the time frames for DHS leaks (s. Section E.3.3), growth is measured in m^3/day instead of m^3/h. The inflection point position b is based on larger pipe diameters with a value of 1. A low growth coefficient c of 0.01 is selected to reflect much longer leak progressions, while the curve steepness of $N = 0.4$ similarly allows for a gradual S-curve. The maximum leak flow rate a is selected as 30 m^3/day to simulate the following behaviour: The leak starts as a small hairline crack of 1 m^3/day [10], a more conservative choice than the reported 2 m^3/day to 3 m^3/day [V]. It reaches 10 m^3/day after 6 months, equivalent to a leak of a somewhat larger size [VI], and arrives at 20 m^3/day after a year, thereby equating the additional network losses reported by M networks. Although the modelled leak does increase progressively over time, in an effort to balance out the possible extremes, it stays well below the more major occurrences reported in practice and literature. It therefore simulates a leak that will not necessarily be called in by the public.

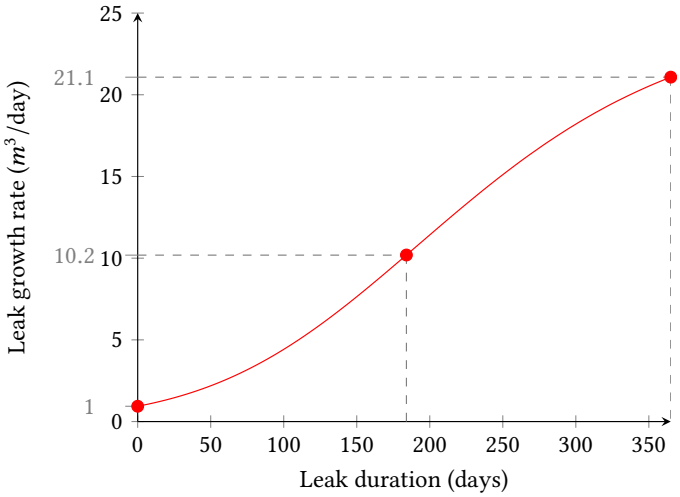


Figure E.2: Leak growth of an exemplary leak with $a = 30, b = 1, c = 0.01, N = 0.4$.

The leak growth curve can be translated into cost via Equation E.2. The cumulative summation helps obtain an amount in m^3 for a given number of days that the leak exists, while multiplication with the cost of make-up water $C_{make-up}$ translates to a quantification of economic value attributable to the ongoing loss. For the exemplary leak growth shown in Figure E.2, the cost can be defined as a function of time – specifically leak duration (s. Equation E.3). The cumulative cost is visualised in Figure E.3.

$$C_{\text{ongoing}}(t) = C_{\text{make-up}} \cdot \sum_{t=0}^{T} f(t) \tag{E.2}$$

[8] Specifically, DHSs have a significantly higher medium temperature to fulfil their purpose, which necessitate other pipe materials.

$$C_{\text{ongoing}}(t) = 5\ \frac{€}{\text{m}^3} \cdot \sum_{t=0}^{T} \left(\frac{30\ \frac{\text{m}^3}{day}}{\left(1 + e^{(1-0.01 \cdot t)}\right)^{\frac{1}{0.4}}} \right) \tag{E.3}$$

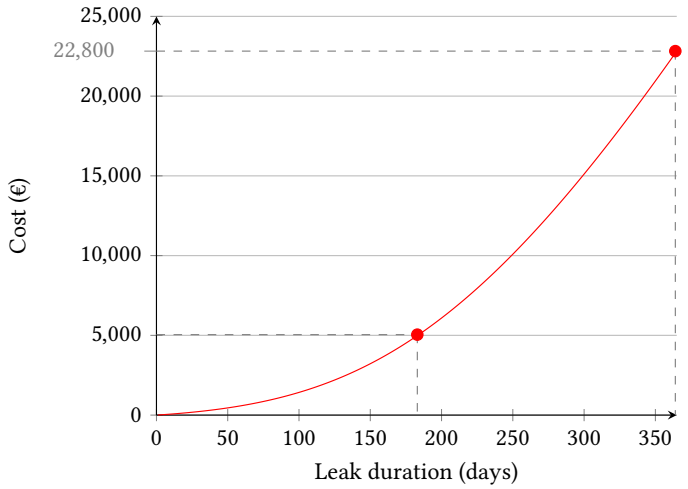


Figure E.3: Cumulative ongoing cost of the exemplary leak defined by Equation E.3.

E.4.1.2 Repair costs

Beyond the costs incurred while leaks persist, additional expenses arise from repairs. On this aspect, the interviewed DHS operators were able to provide significantly more detailed information than they could about ongoing losses. In general, repairs may consist of several cost categories, such as those for materials, civil engineering work, pipe construction, welders, construction cite coordination, and traffic re-routing [*II, V*]. The required services are commonly outsourced to specialised companies [*V*].

The precise requirements are unique to each leak and costs can vary greatly depending on a multitude of factors, such as the surrounding infrastructure, pipeline installation method, cause of failure, time of occurrence relative to the heating period, and many more. Repair costs registered by the participating DHSs described in Section E.3 spanned an incredibly broad spectrum from 2000 € to over 1,000,000 €. In general, the more complex the circumstances, the greater the expense [G. Roehl, personal communication, May 8, 2025]. As Table E.6 shows, such complicating conditions have occurred in DHSs of all sizes, given their placement in urban, and thus structurally speaking complex, areas.

Table E.6: Estimated cost ranges for a single leak repair.

Characteristic	Unit	S	M	L	XL
Repair cost		€ 5,000 – 250,000	3,500 – 100,000	2,000– > 1,000,000	avg. 25,000; mostly < 100,000

The relationship between leak repair costs, their cause, and affected pipeline type is displayed in Figure E.4, largely thanks to a detailed breakdown of occurrences throughout 2024 provided by *II*. Despite the comparatively small number of data points and thus arguable statistical relevance, some important observations can be gleaned from it. Similarly to Section E.3.4, the figure highlights the prevalence of leaks in the most common form of pipelines (PIRP), mainly on account of installation and manufacturing errors. The repair costs for this pipeline type also spans the greatest breath, from 2000 € to 1,000,000 €. Even PIRP leaks attributed solely to installation and manufacturing errors can range from 2500 € to 35,000 €, highlighting again the relevance of situational circumstances. Unsurprisingly, leaks in underground pipelines are more expensive than above-ground or cellar variants. Duct-bound pipes, in particular, require high repair cost due to the nature of the installation method [*II*]. Aside from this, leaks caused by corrosion seem to generally cause greater expenses than wear and ageing or external influences.

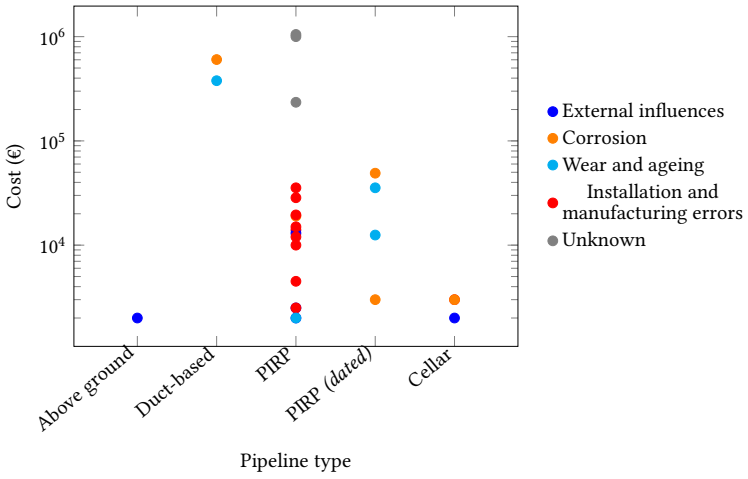


Figure E.4: Repair costs for leaks reported by *II, III, VIII*.

Several DHS operators highlight a correlation between the duration for which a leak exists and its repair cost. For PIRPs, *V* list smaller leaks to require 3000 € to 8000 € for civil engineering and 500 € for materials (mainly sleeves), but these costs increase for leaks that exist for longer time frames to a total of around 20,000 €. This is because material costs escalate to 3000 € to 4000 € when re-insulation becomes necessary due to a larger leak size. These numbers showcase how fixing leaks earlier can save between 1.7 to 5.7 times in repair costs. Similarly, costs of extreme magnitudes of >1,000,000 € often stem from infrastructure damage caused by the prolonged existence of leaks [*III*]. *III* list two such events in which entire crossings or network sections needed replacing due to 6 months of persistent moisture leakage.

To quantify exemplary repair costs for this economic analysis, several assumptions and generalisations must be made. PIRPs are considered for the example leak due to their current and future prevalence and high occurrence both in the given DHSs (Section E.3.1) and leak statistics (Section E.3.4). The key information provided by *V* on the effect of leak duration on repair costs is used as reference to model the time-based dependence. As per *I*, repair costs commonly do not exceed 100,000 €, meaning instances above these are viewed

as outliers and not considered for our conservative example leak. Taking into account the remaining values shown in Figure E.4, the repair cost for a leak of this pipeline type averages at around 12,900 €. This is close to the average of the reported range by *V* for longer existing leak repairs, namely 13,000 €. We therefore assume this to be the cost for repairs after 6 months. The starting point is selected using the average of lower repair costs reported by different operators for PIRPs. These encompass the cluster of instances of lower repair costs from Figure E.4 – that average to 2700 € – as well as lower bounds of various provided value ranges, including 3500 € [*V*] and 5000 € [*VI, VII*]. Together, these combine to an average of around 3400 €, close to equal to the lower bound of the fast repair range defined by *V*. To approximate the extent of the duration-induced cost increase, the repair expenses after a year's existence are also estimated. The cluster of higher leak repair costs from Figure E.4 amount to an average of 18,500 €, just short of the 20,000 € value defining the upper bound of the slower repair range [*V*].

Given these data points, a function is defined to model the time-based cost progression of leak repairs. As simpler functions (such as linear, exponential, or logarithmic) are not able to properly capture the described circumstances, a growth function similar to the one from Section E.4.1.1 is chosen. For the given purpose, we select the simplest and most well-known sigmoid function, the Logistic function [12, 41], generally defined as in Equation E.4.

$$C_{\text{repair}}(t) = \frac{a}{\left(1 + e^{(b - c \cdot t)}\right)} \tag{E.4}$$

Here, the maximum cost a is defined as the upper bound 20,000 €. The remaining parameters, b and c, are chosen as 1.6 and 0.012 respectively to fit the function to the circumstances. This results in a time-dependent repair cost progression as defined by Equation E.5 and shown in Figure E.5.

$$C_{\text{repair}}(t) = \frac{20{,}000\,€}{\left(1 + e^{(1.6 - 0.012 \cdot t)}\right)} \tag{E.5}$$

E.4.1.3 Total leak costs

The ongoing and repair costs from Sections E.4.1.1 and E.4.1.2 can be combined into total leak-induced costs by summation. Figure E.6 displays the resulting cost for the derived exemplary leak, for which a conservative growth and repair were estimated. For any point in time t, the plot shows the cost that the leak accumulated throughout its lifetime, were it be repaired that day. Although repair costs make up the lion's share of expenses in the beginning, the figure clearly shows how ongoing costs replace these in severity if the leak persists for long enough (e.g. contributing more than 50 % after 1 year). It is important, therefore, that these costs are not underestimated.

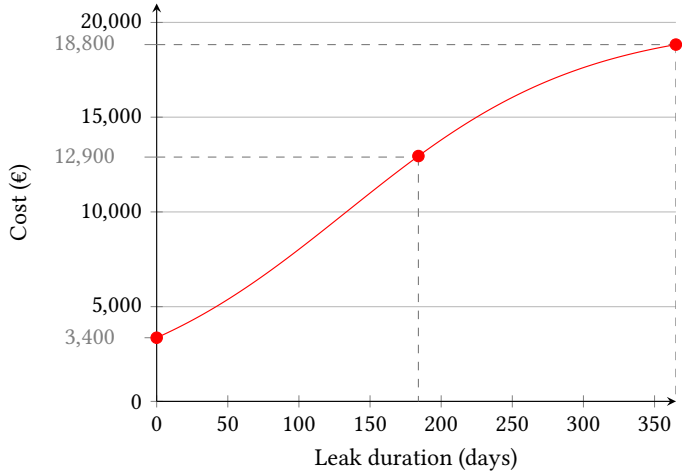


Figure E.5: Repair cost growth of an exemplary leak defined by Equation E.5.

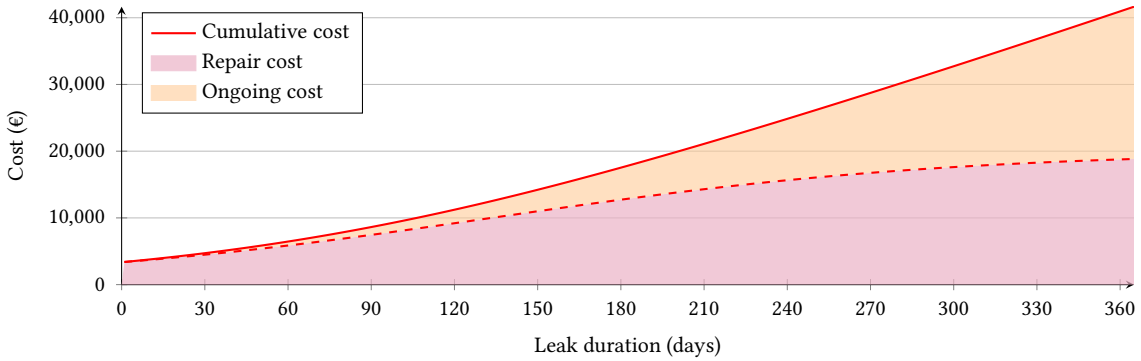


Figure E.6: Combined cumulative ongoing and repair costs for an exemplary DHS leak (own figure from Study E).

E.4.2 The economic necessity for alternative leak detection

In terms of economic viability and cost efficiency, the interviews with DHSs operators also highlight a current need for alternative LD techniques despite commonly installed ILDS or use of other traditional methods. This can be attributed to the following:

1. Lack of usability: ILDS either do not exist or have not been installed in a way that allows them to be used (incorrect wiring, missing circuit diagrams) [*V, VI, VIII*].
2. Lack of usage: Systems are not used as required, i.e. manual readings are not performed regularly. For *III*, this caused a leak to go undetected for 6 months, culminating in 1,000,000 € repair cost.
3. System malfunction: The ILDS either does not give a warning despite the presence of a leak, returns a false positive or provides an incorrect location. The first caused *III* to fail to localise a leak for 6 months, again causing costs of 1,000,000 €. The second can lead to unnecessary excavations and cost, such as for 2500 € as described by *II*.

Incorrect localisations are attributed with 1000 €/m by *I*, though examples show they can grow to over 2000 €/m[9] [*VIII*].

These aspects call into question the advisability of relying solely on ILDS. As described in Section E.3.3, the monitoring strategies shown by DHS operators in practice seem to reflect similar conclusions, with many of them turning to a combination of methods and alternative options, such as thermography.

E.4.3 Estimating TLD costs

The usage of TLD for DHS incurs its own cost, which is quantified in this subsection. However, the method as described in Section E.2.3.4 is characterised by some factors that require consideration for cost estimation. Mainly, these consist of the time frame limitation and the form of acquisition itself.

Given the stringent requirements for TIR acquisition discussed in Section E.2.3.4, the time frame for data collection is limited. This naturally sets boundaries for the applicability of the method and must therefore be quantified as part of the economic viability assessment. By means of a comprehensive analysis of historical weather data in Germany, included in Appendix E.III, the viable number of days per year in which TIRs can be acquired is estimated as 40. Clusters of days at the beginning (November), middle (January), and end (March) of the heating period are found to commonly exhibit suitable characteristics that allow the capture of high-quality images, meaning multiple acquisitions per year are theoretically possible.

While the technology can be implemented via ground-based (hand-held, vehicle-mounted) or aerial (UAS, aeroplane) sensor, practical experience has shown the latter to be most viable. Ground-based thermography allows for the highest GSD but at the cost of an extremely time-consuming acquisition[10]. Additionally, these methods are subject to access restrictions in many urban areas, preventing full coverage [*IX*].

The most suitable type of aerial acquisition is a matter of contention among the interviewed operators. While *IX* have incorporated aeroplane-based TLD into their monitoring strategy, *IV* report images from this form of acquisition to provide only superficial overviews and not enough detail for comprehensive LD. This is due to the lower GSD, a side effect of the great flight heights. In comparison, UASs can fly much closer to the ground[11], thereby combining the benefits of high resolution and full coverage in one. At the same time, Heipke and Tödter [15] state around 20 km of pipeline can be covered via UAS per night, making it 10

[9] An instance of incorrect localisation reported by *VIII* lead to a tripling of repair costs when 30 m of pipeline required excavation until the true leak position was pinpointed.

[10] According to *VI*, the hand-held approach takes around 4 h/km, meaning that only around 2 km can be covered in one night.

[11] Sledz and Heipke [40] achieve a GSD of 5.2 cm for their UAS flight at 40 m, representing nearly a 5-fold improvement in resolution compared to the 24 cm GSD reported by Friman et al. [9] for their 800 m aeroplane flight.

times faster than the hand-held approach. While this speed may not be sufficient to cover of the largest XL networks in their entirety, its flexibility and applicability to most forms of network make UAS-based acquisition the main focus of this study.

The cost of UAS-based thermography comprises several key elements. These include the hardware (thermal camera and UAS), the acquisition flights (trained pilot), and the evaluation of the data (expert image analysis). While it is theoretically possible for operators to invest in the required technology and train or hire expert staff, this would mean a substantial initial investment. In terms of hardware, the costs are driven by thermal sensors, which are considerably more expensive than standard colour cameras [7].

As an alternative, several companies in Germany have begun offering the UAS-based thermography as a package service. In an effort to find specific leaks in their networks, some interviewed operators report trying such services. They list expenses of between 300 and 600 /€km for network coverage [*VI, VII*]. To further disaggregate these costs, offers were requested from different companies for an exemplary region. Three areas of expense are thereby identified: 1. flight costs C_{flight} (20 % to 30 %), 2. operational expenses $C_{\text{operation}}$, ranging from travel and accommodation to generators for loading the UAS batteries (30 % to 60 %), and 3. image analysis effort C_{analysis} to identify leak candidates for report creation (25 % to 40 %). For comparison, in the case of aeroplane-based acquisition – which is listed by *IX* as costing 143 $\frac{€}{km}$ – roughly 50 % are attributed to image analysis and report generation.

Recent advances in the field of TIR analysis for DHS LD have enabled the automation of said analysis step [6, 9, 15, 18, 40, 45, 46]. This not only allows for the reduction of a considerable portion of service cost for both kinds of aerial acquisition, but also removes the time-related bottleneck meaning results can be provided faster to the operators. Additionally, it may supply DHS operators with the necessary tool to implement the approach themselves by eliminating the requirement for extensive thermographic knowledge.

Taking all afore-mentioned aspects into account, the cost of UAS-based thermography as a service can be quantified in general according to Equation E.6, with n_{service} the number of DHS flyovers planned per year, C_{service} the service cost per km, and L_{DHS} the network pipeline length:

$$C_{\text{TLD}} = n_{\text{service}} \cdot C_{\text{service}} \cdot L_{\text{DHS}} = n_{\text{service}} \cdot (C_{\text{flight}} + C_{\text{operation}} + C_{\text{analysis}}) \cdot L_{\text{DHS}} \tag{E.6}$$

Assuming the average of expenses reported by DHS operators, the cost to requisition UAS-based TLD C_{service} is defined as 450 €/km. However, since the analysis step can be automated, the associated cost C_{analysis} can be eliminated, thus reducing the total TLD cost. Given the previously described range for C_{analysis}, C_{service} can drop to 270 €/km to 337.5 €/km. For the purpose of this study, the most conservative value is chosen – meaning 25 % savings. The TLD costs can therefore be summarised as shown in Table E.7 for this study's four-tier categorisation (see Section E.3.1). Given the average number of viable days for TIR acquisition per year and kilometres of pipelines that can be covered by

night, all sizes of DHS (up to 800 km) can be checked via UAS-based thermography at least once per heating period ($n_{\text{service}} = 1$). The TLD cost is therefore a function of DHS length and defined as Equation E.7.

$$C_{\text{TLD}}(L_{\text{DHS}}) = 1 \cdot 337.5 \, \frac{€}{km} \cdot L_{\text{DHS}} \tag{E.7}$$

Table E.7: Estimated cost ranges for the TLD approach.

Characteristic	Unit	S	M	L	XL
L_{DHS}	km	< 30	$30 - 100$	$100 - 300$	> 300
C_{TLD}	€	$< 10,110$	$10,110 - 33,750$	$33,750 - 101,250$	$> 101,250$
Required time frame	nights	≤ 2	$2 - 5$	$5 - 15$	> 15

E.4.4 Break-even analysis

The previously estimated costs of exemplary leak and TLD are compared to assess the economic viability of the latter method. In particular, the objective of the analysis lies in identifying the BEP for TIR-based DHS pipeline LD. To this end, the following scenarios are contrasted in a BEA:

$$TC_{\text{inf}} = n_{\text{leaks}} \cdot \lim_{t \to \infty} C_{\text{ongoing}}(t) \tag{E.8}$$

$$TC_{\text{none}} = n_{\text{leaks}} \cdot \left(C_{\text{ongoing}}(T) + C_{\text{repair}}(T) \right) \tag{E.9}$$

$$TC_{\text{TLD}} = n_{\text{leaks}} \cdot \left(C_{\text{ongoing}}(t_d) + C_{\text{repair}}(t_d) \right) + C_{\text{TLD}}(L_{\text{DHS}}) \tag{E.10}$$

In the reference scenarios (Equations E.8 and E.9), leaks are not actively sought, meaning no expenses are incurred by detection methods. Instead, leaks are assumed to grow – either infinitely (Eq. E.8) or until day T when, for instance, the DHS operators are informed about their location by the public (Eq. E.9). This means they continuously incur time-dependent ongoing C_{ongoing} and – in the case of TC_{none} – repair C_{repair} expenses once found. The costs are combined as discussed in Section E.4.1.3. In the TLD scenario (Equation E.10), leaks are localised via UAS-based thermography. Here, ongoing and repair costs only increase until the leaks have been removed on day t_d, though at the additional price of the thermography-based survey C_{TLD}. At minimum, t_d consists of the number of nights required to survey the DHS – which depends on network size – and the time needed for image analysis and leak repair. The latter duration is estimated at 7 days, assuming the analysis is automated.

The number of leaks n_{leaks} that we assume occur for the purpose of this analysis depend on the size of the network. To quantify this number, the four categories of DHS length are viewed individually. Table E.8 presents the defining characteristics for the BEA based on findings from Section E.3.2 as well as the following additional assumptions. XL networks

are capped at 800 km, as this is the upper limit for a TLD coverage within the viable amount of days (see Section E.4.3). Given the detection bias discussed in Section E.3.2, we assume a leak rate that increases slightly with DHS size from 0.06 leaks/km to 0.1 leaks/km. While this amplifies the empirically found value for XL networks, it still works conservatively, as we may assume a much higher unrecorded leak number given standard losses in such systems. Combining network size with leak rate provides estimates for a range of n_{leaks} in each system.

Table E.8: Network characterisations for BEAs.

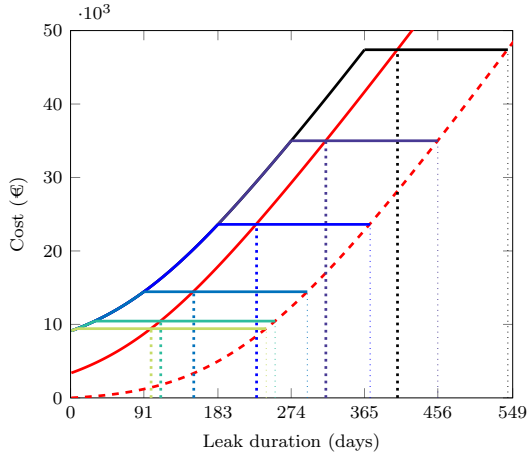
Characteristic		Unit	S	M	L	XL
General	L_{DHS}	km	0 − 30	30 − 100	100 − 300	300 − 800
	Leak rate	$\#/km$	0.06	0.09	0.09	0.1
	n_{leaks}	$\#/yr$	0 - 2	2 - 9	9 - 27	30 - 80
Example	L_{DHS}	km	17	56	200	550
	n_{leaks}	$\#/yr$	1	5	18	55
	t_{TLD}	nights	1	3	10	28
	$t_{d_{\text{min}}}$	days	8	10	17	35

Using the thus determined general value ranges, a representative DHS is defined for each of the four network categories (see Table E.8, "Example"). In each case, average values for L_{DHS} and n_{leaks} are selected. The former helps define the number of nights required for TLD, t_{TLD}, which in turn helps to estimate a minimum $t_{d_{\text{min}}}$.

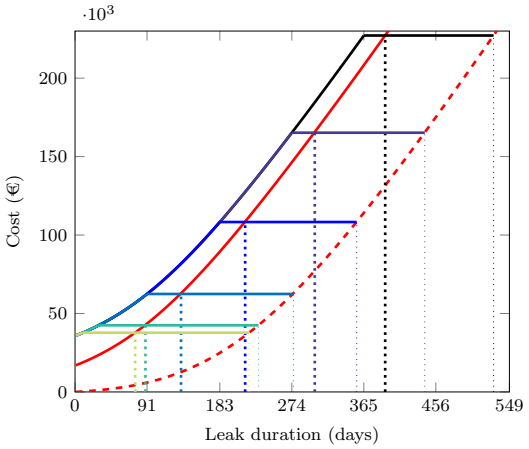
Comparative cost plots for each of the four network categories are provided in Figure E.7. To improve interpretability, it is assumed that all leaks occur simultaneously and that in thermography scenario the C_{TLD} expense is incurred at the beginning – regardless of when exactly the flights take place. The figure visualises the BEA for leak localisation and repair at $t_d \in \left\{t_{d_{\text{min}}},\ 30,\ 91,\ 183,\ 274,\ 365\right\}$ days, thus spanning from weeks over quarter-year increments to a year. These time frames take into account that delays can occur during localisation (as TLD is only implemented once a year) or repair (see Section E.3.3). For reference, the precise values visualised in Figure E.7 are provided in Appendix E.IV.

A comparison with the reference scenarios shows that, regardless of network size or leak removal date, the BEP t_{BEP} is reached in a surprisingly short period of time after t_d – for TC_{none} no longer than three months. Even if one assumes that leaks are never found (TC_{inf}), the extra expense of TLD and repair is shown to always pay off in less than three quarters of a year after t_d.

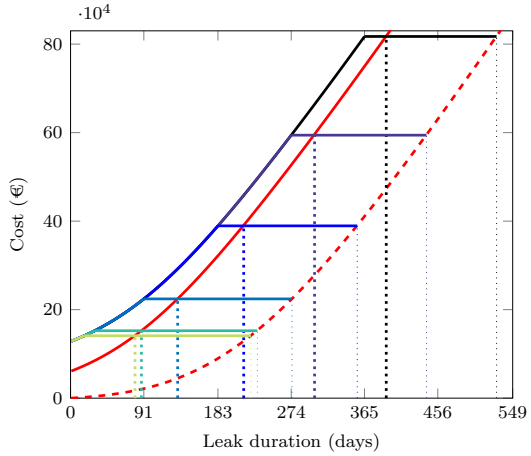
When comparing the different t_d values, an interesting trend can be observed mirrored through all network sizes. Faster leak detections and removals take slightly longer to amortise than slower ones, specifically the delta $\Delta t_{\text{BEP}} = t_{\text{BEP}} - t_d$ diminishes with larger t_d values. This is caused by the non-linearity of both C_{ongoing} and C_{repair}, which means that the relative effect of the added constant C_{TLD} decreases over time. For $t_{\text{BEP}_{\text{none}}}$, for instance, the delta in S networks decreases by 55.4 % from the initial 92 days for $t_{d_{\text{min}}}$ to 41 days for $t_d = 365$. This aspect can be seen as favourable for TLD, as one may assume that the probability of a leak being localised by other means (such as public notification) grows over time. Later localisation via TLD should therefore amortise faster for the method to remain viable.



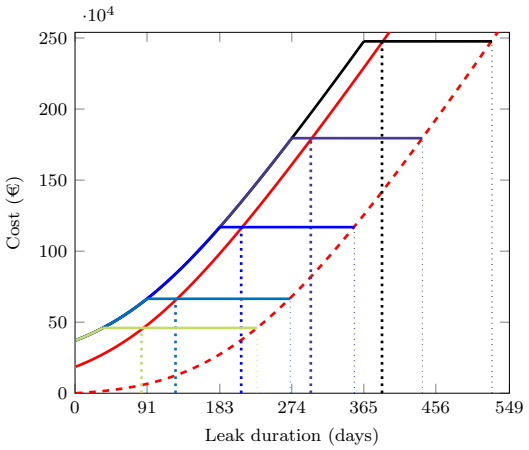
(a) BEA for S example network ($L_{\text{DHS}} = 17$ km, $n_{\text{leaks}} = 1$, $t_{d_{\text{min}}} = 8$).



(b) BEA for M example network ($L_{\text{DHS}} = 56$ km, $n_{\text{leaks}} = 5$, $t_{d_{\text{min}}} = 10$).



(c) BEA for L example network ($L_{\text{DHS}} = 200$ km, $n_{\text{leaks}} = 18$, $t_{d_{\text{min}}} = 17$).



(d) BEA for XL example network ($L_{\text{DHS}} = 550$ km, $n_{\text{leaks}} = 55$, $t_{d_{\text{min}}} = 35$).

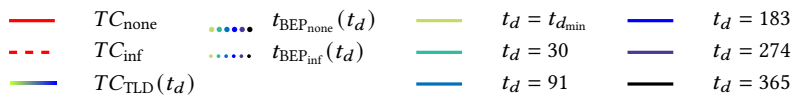


Figure E.7: BEAs across different network sizes. Red lines depict the reference scenario costs, aqua-coloured lines the TLD costs for $t_d \in \{t_{d_{\text{min}}}, 30, 91, 183, 274, 365\}$ days. To show when TLD costs amortise, the BEPs are marked by dotted vertical lines – thick ones for $t_{\text{BEP}_{\text{none}}}$, thin ones for $t_{\text{BEP}_{\text{inf}}}$.

Another interesting observation from Figure E.7 is that TLD pays off faster with increasing network size. This may, in particular, be attributed to the associated leak rates. For instance, XL DHSs have $\Delta t_{\text{BEP}_{\text{none}}}$ values that range from 52 days ($t_{d_{\text{min}}} = 35$) to 25 days ($t_d = 365$), which are 43 % and 39 % smaller, respectively, than those previously listed for S networks. This indicates that the method becomes more viable with higher leak rates.

These findings have an impact on the economic viability of TLD in general. In the best-case scenario, where leaks can be removed in the shortest time frame ($t_d = t_{d_{\text{min}}}$), amortisation of the UAS-based TLD service cost takes no more than 3 months after leak removal and the additional repair costs are paid off in less than three quarters of a year. In the worst-case

scenario, where leaks are located and repaired only after a year ($t_d = 365$), the cost for TLD is amortised almost within a month after t_d and the entire localisation and repair via TLD pays off in less than half a year after t_d. In summary, this means that the true economic benefit of TLD, namely faster amortisation over time, complements its methodological strength as a LD method, where its advantage lies in the ability to help locate leaks that are otherwise not pinpointed quickly.

In the context of this BEA, an additional sensitivity analysis (SA) is performed to show the impact of automation on TC_{TLD}. As detailed in Appendix E.V, the analysis shows that automating C_{analysis} — and thereby reducing TC_{TLD} by a given percentage — leads to a reduction in the amortisation period Δt_{BEP} by at least the same percentage. This means that even a conservatively estimated automation significantly shortens the amortisation time beyond the savings percentage. Given future technological advancements and growing UAS market [17], this aspect can be expected to provide an even greater potential for cost savings and increased amortisation, as the share of C_{flight} can be expected to drop.

An aspect that is disregarded in this direct cost comparison (as it is not easily quantified) is the additional benefit afforded by the investment in TLD. Acquired TIR images can be used beyond mere LD to assess the condition of the DHS pipelines [*IV, IX*]. If flights are performed regularly, a temporal analysis can provide valuable information on network degradation [6]. *IX* highlight these aspects as other main motivators for using aerial TLD.

E.5 Recommendations

In light of the findings from Sections E.3 and E.4, several important statements can be made regarding DHS LD and specifically TLD from an economic standpoint. These elicit general as well as DHS-unique recommendations.

The considerable expenses caused by leaks in practice (Section E.4.1) highlight the critical importance of implementing LD in general. At the same time, the empirical study shows that this aspect is attributed with a varying degree of urgency and importance, especially where less critical leaks are concerned (Section E.3.3). Similarly to other publications [15, 50], this study's findings urge for the general adoption of regular LD in DHS monitoring. Beyond that, leaks should be viewed as part of the system's lifecycle, meaning that LD should already be considered during network design.

Modelling leak growth and ongoing cost as part of this study has exposed the critical impact of leaks on economic efficiency. This aspect is key in the fast amortisation of TLD in the BEA, yet – as discussed in Section E.4.1.1 – a considerable lack of provided data in this regard makes apparent the current focus on repair. Operators should incorporate ongoing cost calculations into their system assessment when selecting LD methods.

The BEA clearly highlights the economic advantage that aerial TLD can offer for DHS efficiency, but the decision of its usage depends on each network's unique characteristics. The following factors require consideration:

- Defining L_{DHS}: While TLD is pipeline-agnostic (E.2.3.4), the applicable area is limited by some factors. This includes installation depth (s. Appendix E.I) and pipe surroundings, such as barriers that obstruct the line of sight between sensor and ground (train tracks, excessive foliage, etc.). Such areas fall outside the scope of remote sensing-based TIR acquisition and cannot be inspected.
- Defining n_{leaks}: Although the number of leaks has been found to scale with L_{DHS} (Section E.3.2), this may not apply to all DHSs. The individual leak rate, assignment to one of the four network categories, and – by extension – applicability of the derived BEA must therefore be checked.
- Defining C_{service}: As shown in Section E.4.3, the derived cost for TLD as a UAS service can vary f.e. depending on DHS location. Requesting specific estimates from companies for the given network will provide precise values for individual cost estimations.
- Alternate LD: If other techniques are being implemented, comparisons should be drawn between their costs and application areas (Section E.2.3). Given the unique circumstances of each DHS, another method might outweigh TLD in terms of economics or applicability.

Taking these as well as previous chapters into account, DHS operators can assess the economic viability of TLD given their unique circumstances. Marginal cases may also occur, where other forms of implementations are more cost-effective. For S networks with few leaks, hand-held thermographic sensors might be an alternative. Larger networks may consider acquiring the ownership and know-how to carry out UAS flights independently. For XL DHSs, the use of aeroplanes might be more economically viable than UAS [*IX*].

While Section E.4.4 highlighted the economic viability of TLD for leak detection, especially given a degree of automation (s. Appendix E.V, the method also provides other benefits. In the long run, TLD has the potential to be used for large-scale leak monitoring through temporal analyses [6]. While a single TIR acquisition provides only a snapshot of the DHS, data collected across regular UAS flights can be compared to help operators assess their DHS's general condition. However, as with all methods discussed in Section E.2.3, TLD is subject to limitations and may not be capable of assessing a DHS in its entirety. A truly comprehensive LD monitoring strategy should rely on the combination of different techniques to ensure full coverage [42, 52].

E.6 Conclusion and outlook

This study addressed a critical research gap by evaluating the economic viability of TLD for leak detection in DHS pipelines, with a specific focus on networks in German-speaking

regions. In contrast to other established methods, TLD offers a flexible, scalable, and non-invasive leak localisation approach for all manner of pipe types. Through the integration of new empirical data, a novel form of leak-associated cost was modelled and first-of-its-kind BEA performed. The findings presented in Section E.4.4 highlight TLD's potential as a cost-effective tool for leak localisation in networks of all sizes given their short amortisation periods. This study was thereby able to justify the use of TLD for increased system efficiency – not only in terms of cost.

Naturally, it is subject to several limitations. As a case study of German-speaking countries, the observations and findings may not be applicable to other nations, in particular outside of Europe. Much of the content was developed based on the empirical study, which consisted of a relatively small number of DHS operators. The occasionally sparse information required generalisations that may not be accurate for the majority of DHSs. In order to perform the economic analysis, assumptions needed to be made to model the exemplary leak. In practice, leak growth and repair is defined by a myriad of influencing factors, meaning the cost of a specific leak may differ considerably from the simulated example. The same can be said for DHSs, where each has its own defining characteristics that may differ from the defined four-tier categorisation.

While this study has provided an important initial economic assessment of TLD for DHS leak localisation, there is much potential for future research. To better replicate real-world conditions, exemplary leak modelling can be expanded to simulate different kinds of leak growth in various pipe types. The BEA can be further developed by adjusting for currently disregarded time-dependant changes in cost, such as season-based fluctuations and interim leak repairs. Drawing economic comparisons with other LD methods, such as ILDS, could provide additional weight to the results. In this regard, expanding the questionnaire focus and empirical study – for instance to a European level – would help provide a more solid and more holistic data foundation. On a broader scale, the economic assessment of approaches such as TLD not just for leak detection but also condition monitoring could help shape the future of DHS monitoring.

Acknowledgments

The authors gratefully acknowledge all study participants for their valuable contributions. This work is supported by funding from the European Union through the AI4EOSC project (Horizon Europe) under Grant number 101058593.

Declarations

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

During the preparation of this work, the authors used ChatGPT (OpenAI) to support technical clarification and expression, as well as explore options for data visualisation and interpretation. After using this tool, the authors reviewed and edited the content as needed and assume full responsibility for the content of the published article.

CRediT author statement

Elena Vollmer: Conceptualisation, Data Curation, Formal Analysis, Investigation, Methodology, Validation, Visualisation, Writing - Original Draft, Writing - Review & Editing. **Rebekka Volk**: Funding Acquisition, Supervision. **Frank Schultmann**: Writing - Review & Editing, Supervision.

Appendices

Appendix I. Conditions for TLD

The acquisition-dependent conditions required for collecting high quality TIR images can be summarised as follows:

- Weather-related disturbances can impact the temperature signature on the earth's surface, causing low quality or erroneous TIRs. These disturbances include solar radiation, cloud cover, wind, rain, snowfall, and fog [9, 15, 46]. Direct sunlight can already cause significant errors in TIR sensor measurements in the far infrared spectrum [29], persistent rain and wind cool down the surface [15], and water droplets in various forms of precipitation greatly reduce resulting image quality [46]. Acquisition should therefore be carried out in dry weather and at low wind speeds [9].
- The camera should have an unobstructed view of the ground to ensure no leaks are missed. This includes no insulating layers of foliage or snow [9, 46].
- Acquisition should only take place at nighttime or in the early morning when thermal reflectance is lowest, the delta of surrounding temperatures to the network highest, and the TIR sensor most accurate[12] [15, 46].
- The carrier temperature in the flow pipe should correspond to the maximum permissible continuous operating temperature and should be constant over a longer period of time so that the heat loss is as constant as possible [15].
- The flow pipe installation depth should be near constant and not exceed 1 m [15].

[12] Pech et al. [34] compare different flight times – 7 AM, 2 PM and 9 PM – and find that midday flights cause considerable variation to ground measurements

Appendix II. Questionnaire

This appendix lists the set of questions posed to the DHS operators who participated in the study. The first group consists of general questions:

- What is the size of your network (pipe length in km)?
- What types of pipes are installed (percentage of the total network or length in km)?
- How many kilometers of your district heating network are equipped with sensors for leak detection? What specific sensor technology is installed? How reliable is it?
- Do you know of any correlations between leaks and the pipeline age, pipeline type, etc.?
- Have you already had a particularly damaging leak in the network? If so: When was this? What pipeline type did it affect? How high were the losses and repair costs?

A second group of questions revolved around information for a time period (such as per year or higher granularity if available):

- What is the heat production for your network (GWh)?
- What losses are registered (MWh)? What costs are incurred (/ € MWh or / em^3)?
- If there is a general energy loss without leaks, what proportion of the loss is due to leaks?
- How many leaks does your network have per year? How large are they?
- How long does it take to find and repair leaks? What costs are incurred with the methods currently used to locate leaks in pipes?
- What costs are incurred in the event of a falsely located leak, i.e. by excavating at the incorrect place?

Appendix III. Estimating the viable day count for TLD

Historical weather data from Meteostat [32] are analysed to estimate the number of days that high-quality TIR images can be acquired. To this end, weather stations located within the largest city in each German state are selected to represent the different regions. For a representative assessment, an investigation period of three consecutive years is selected, specifically 2019, 2020, and 2021[13]. Daily measurements are used to identify when the stringent conditions as per Appendix E.I were met. Specifically, these are defined as:

1. a maximum temperature below 10 °C ($T_{\max} < 10\ ^{\circ}C$),
2. a minimum temperature below 5 °C ($T_{\min} < 5\ ^{\circ}C$),

[13] These specific years were chosen as they were the only ones in the past decade for which data was consistently available for all days [32].

3. no precipitation ($prcp = 0\ mm$),
4. no snowfall ($snow = 0\ mm$),
5. an average wind speed below 5 m/s ($wspd < 5\ m/s$).

Figure E.8 displays the results of the analysis averaged for each of the different states as well as across Germany. Acceptable time frames generally range from fall to spring, whereby periods of consistently favourable conditions are found in November, mid-January, and mid-March. This indicates that multiple acquisitions could be organised for DHS assessment, such as at the beginning and end of the main operating season.

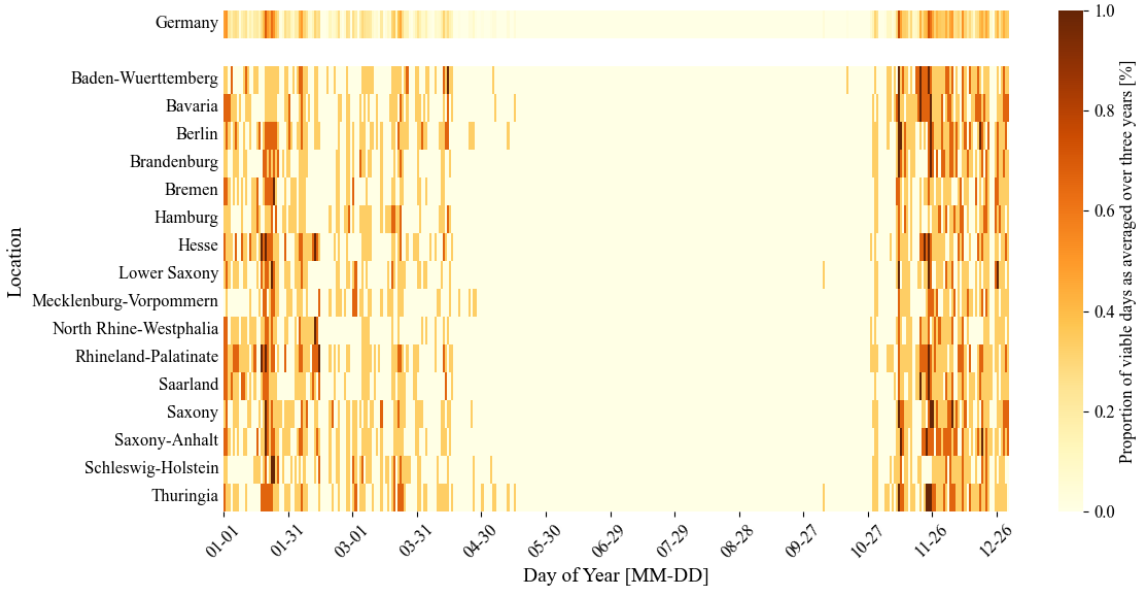


Figure E.8: Viable days for TIR image acquisition across Germany and its various states (based on weather data from Meteostat [32]).

In terms of quantification, Figure E.9 visualises the number of viable days per year for data acquisition throughout Germany. While the box plots highlight to what extent this value may differ depending on the year in question, the median is found to lie within a comparatively similar range of between 35 and 43 days. The combination of these values, shown in the overall box plot, depicts a median of around 39, which is almost identical to the overall average of 40 days. This latter value will therefore be used as an underlying assumption for the economic viability analysis.

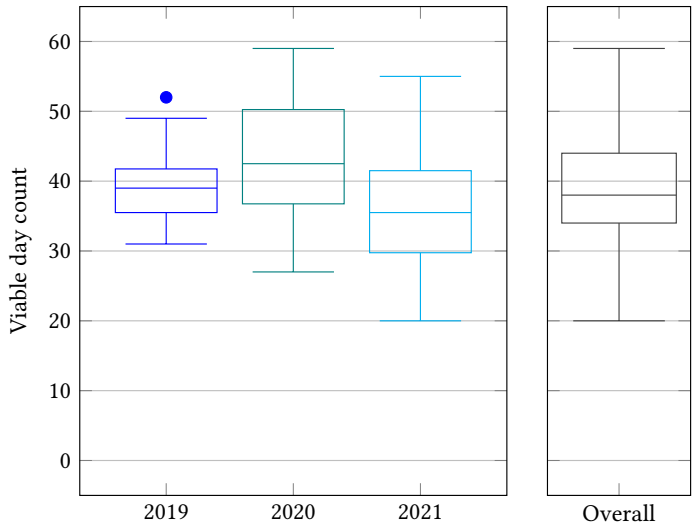


Figure E.9: Box plots of viable day counts for TIR image acquisition per year across all German states (based on weather data from Meteostat [32]).

Appendix IV. Breakdown of the break-even analysis

Table E.9 gives a quantitative overview of the BEA results, detailing the values that are visualised in Figure E.7. Similarly to there, TC_{none} (Eq. E.9) and TC_{inf} (Eq. E.8) are compared to TC_{TLD} (Eq. E.10). t_{BEP} is the number of days at which the BEP is reached for both of those comparisons, while Δt_{BEP} is the number of days after which the BEP is reached after localisation and repair ($\Delta t_{\text{BEP}} = t_{\text{BEP}} - t_d$). The total cost at BEP is the same for both comparisons, as it is equal to $TC_{\text{TLD}}(t_d)$.

Table E.9: Breakdown of BEA results for all networks sizes.

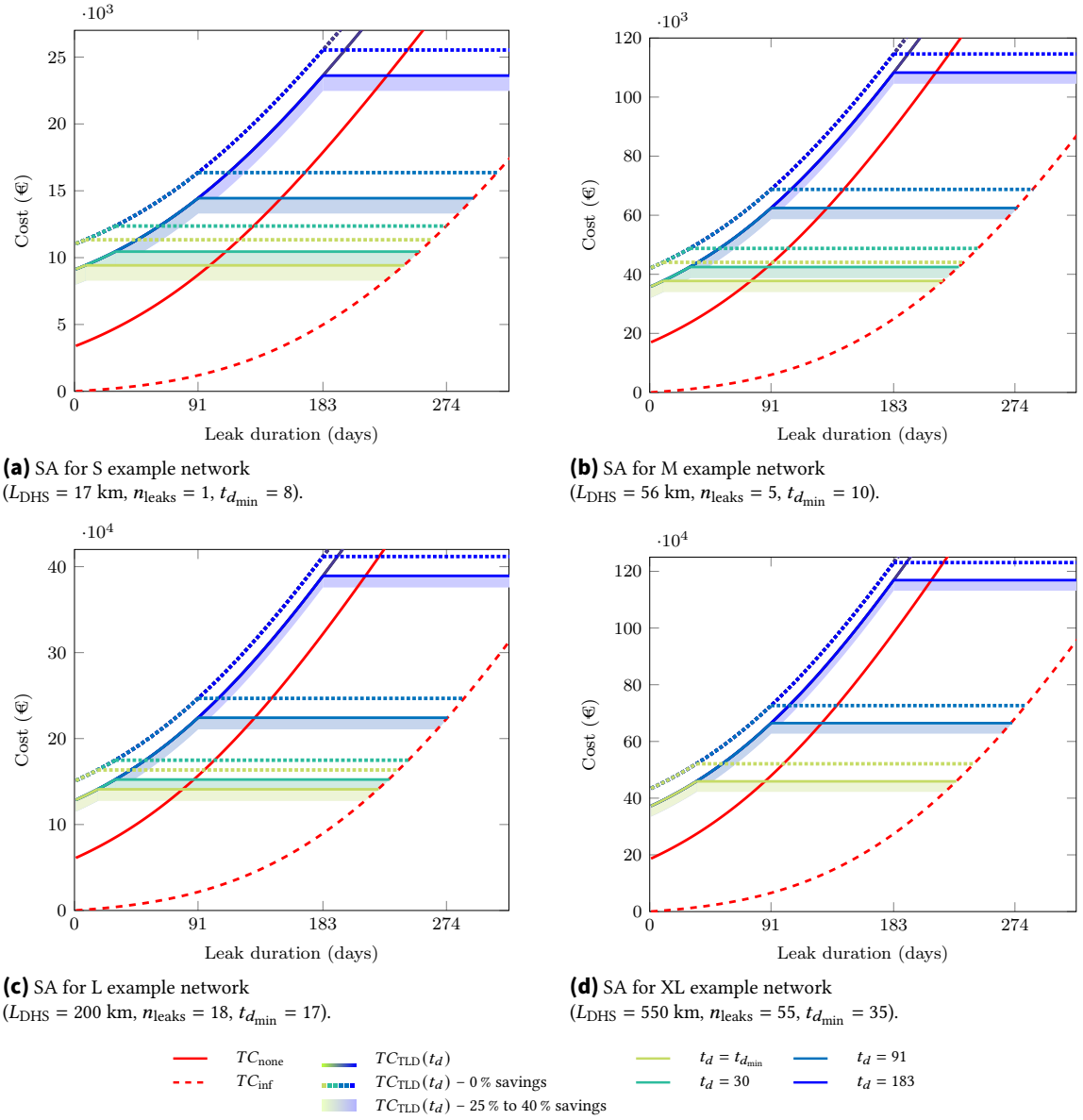
t_d scenario	Parameter	Unit	small		medium		large		very large	
			TC_{none}	TC_{inf}	TC_{none}	TC_{inf}	TC_{none}	TC_{inf}	TC_{none}	TC_{inf}
$t_{d_{min}}$	t_{BEP}	days	100	243	76	220	80	224	87	230
	Δt_{BEP}	days	92	235	66	210	63	207	52	195
	TC_{TLD}	€	9422.97		37753.95		140974.56		459298.95	
30	t_{BEP}	days	112	254	89	232	88	232	-	-
	Δt_{BEP}	days	82	224	59	202	58	202	-	-
	TC_{TLD}	€	10451.05		42467.75		152343.9		-	
91	t_{BEP}	days	153	294	134	276	133	275	129	272
	Δt_{BEP}	days	62	203	43	185	42	184	38	181
	TC_{TLD}	€	14446.79		62446.45		224267.22		664635.95	
183	t_{BEP}	days	231	372	215	356	215	356	212	353
	Δt_{BEP}	days	48	189	32	173	32	173	29	170
	TC_{TLD}	€	23615.91		108292.05		389311.38		1168937.55	
274	t_{BEP}	days	317	456	303	442	303	442	300	439
	Δt_{BEP}	days	43	182	29	168	29	168	26	165
	TC_{TLD}	€	34999.9		165212		594223.2		1795057	
365	t_{BEP}	days	406	543	392	529	392	529	390	527
	Δt_{BEP}	days	41	178	27	164	27	164	25	162
	TC_{TLD}	€	47390.56		227165.3		817255.08		2476543.3	

Appendix V. Impact of automation on TLD cost

This appendix provides details on the impact that automating the analysis step has on TLD cost given the associated reduction of C_{analysis}. Figure E.10 shows how this choice influences each BEA for the various network sizes as visual SAs. As in Figure E.7, the aqua-toned lines show the TC_{TLD} costs depending on t_d (Eq. E.10, with $C_{\text{service}} = 337.5\ \frac{€}{km}$ given a 25 % savings through C_{analysis} automation (Eq. E.7). As described in Section E.4.3, this was the most conservative choice given the possible range of 25 % to 40 % that manual C_{analysis} currently contributes to the total cost. To assess the impact of said choice, this value range is depicted visually as the coloured, filled in areas below each aqua-toned line, where the 40 % maximum savings scenario are at the lower edge of each area. The dotted lines above the 25 %-savings scenario represent the current status quo – no automation of C_{analysis} – as a baseline.

The comparison is based solely on monetary savings, without taking into account the time delay to which the baseline scenario is inherently subject, given that manual image

viewing may require up multiple weeks depending on the network size[14]. Despite this, Figure E.10 shows a significant difference between baseline and the savings scenarios. Even the most conservatively assumed automation enables TLD costs to amortise faster for $t_d = 1$ month than they do for the $t_{d_{\min}}$ baseline.



(a) SA for S example network ($L_{\text{DHS}} = 17$ km, $n_{\text{leaks}} = 1$, $t_{d_{\min}} = 8$).

(b) SA for M example network ($L_{\text{DHS}} = 56$ km, $n_{\text{leaks}} = 5$, $t_{d_{\min}} = 10$).

(c) SA for L example network ($L_{\text{DHS}} = 200$ km, $n_{\text{leaks}} = 18$, $t_{d_{\min}} = 17$).

(d) SA for XL example network ($L_{\text{DHS}} = 550$ km, $n_{\text{leaks}} = 55$, $t_{d_{\min}} = 35$).

TC_{none} TC_{inf} $TC_{\text{TLD}}(t_d)$ $TC_{\text{TLD}}(t_d)$ – 0 % savings $TC_{\text{TLD}}(t_d)$ – 25 % to 40 % savings $t_d = t_{d_{\min}}$ $t_d = 30$ $t_d = 91$ $t_d = 183$

Figure E.10: SAs for the impact of automation on TLD cost across different network sizes.

Given the linearity of C_{TLD} and its sole dependency on L_{DHS}, the absolute cost difference to the baseline is constant for a network, regardless of t_d. This means that in terms of percentage, savings are more impactful earlier on. In the S network, for instance, cost savings for $t_{d_{\min}}$ amount to 17 % in the minimum savings scenario (25 %) and can reach up

[14] According to [47], around 1300 images are acquired per km pipeline, which assuming a fast viewing speed of 1 s/imagealready results in 25 days for XL networks

to 27 % for the maximum scenario (40 %). Given the much higher cost at $t_d = 365$ days, the savings here comprise only 4 % and 6 % of the total, respectively. However, amortisation is a different matter. As indicated above, the relevance of automation and associated savings plays a key role. Compared to TC_{none}, the most conservative automation reduces the time Δt_{BEP} by at least 19.3 % ($t_{d_{\min}}$) and up to 25.45 % ($t_d = 365$ days) for S networks. These numbers increase with network size, amounting to 21.2 % to 26.5 % for XL networks.

A comparison between the different savings scenarios shows that increased automation is greatly beneficial and increases in relevance as time passes. For S networks, for instance, an additional 15 % automation achieves 12.3 % faster amortisation for Δt_{BEP} compared, which increases to 14.6 % at $t_d = 365$ days. These values also grow with network size, reaching 13.6 % to 14.7 % for XL networks – and therefore a 34.9 % to 41.2 % reduction through maximum automation. This means that, given time, automating C_{analysis} by a given percentage results in an even greater percentage reduction in the amortisation time Δt_{BEP}.

Bibliography

[1] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2013). *Instandhaltungsstrategien und Rehabilitationsplanung - Mindestanforderungen [Maintenance strategies and rehabilitation planning - minimum requirements]*. Merkblatt [Fact sheet] FW 114. AGFW.

[2] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Hauptbericht 2022 [Main report 2022]*. Report. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht (visited on 1 May 2025).

[3] Arbeitsgemeinschaft Fernwärme [German Working Committee on District Heating] (AGFW) (2023). *Verfahren zur Zustandsermittlung von Fernwärmeleitungen und zur Feststellung / Einmessung von Abweichungen (Leckortung) [Methods for Assessing the Condition of District Heating Pipelines and for Locating Deviations (Leak Detection)]*. Technical report FW 435. Frankfurt am Main, Germany: AGFW. URL: https://www.agfw-shop.de/regelwerk/fw-435c-verfahren-zur-zustandsermittlung-von-fernwarmeleitungen-und-zur-feststellung-einmessung-von-abweichungen-leckortung-druckfassung.html.

[4] Arbeitsgemeinschaft QM Fernwärme [Swiss Working Committee on District Heating] (2021). *Planungshandbuch Fernwärme [Handbook on Planning of District Heating Networks]*. Handbook. Version 1.3. Bern, Switzerland: Bundesamt für Energie [Swiss Federal Office of Energy]. URL: https://www.verenum.ch/Dokumente/PHB-FW_V1.3.pdf.

[5] Axelsson, S. (1988). "Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 26(5), pp. 686–692. DOI: 10.1109/36.7695.

[6] Berg, A., Ahlberg, J., and Felsberg, M. (2016). "Enhanced analysis of thermographic images for monitoring of district heat pipe networks". In: *Pattern Recognition Letters* 83, pp. 215–223. DOI: 10.1016/j.patrec.2016.07.002.

[7] Clouet, A. (2023). "The thermal imaging and sensing market since 2019". In: *Thermosense: Thermal Infrared Applications XLV*. Ed. by Avdelidis, N. P. Orlando, United States: SPIE, p. 34. DOI: 10.1117/12.2664172.

[8] Euroheat & Power (2024). *DHC Market Outlook*. Report. Brussels, Belgium: Euroheat & Power. URL: https://api.euroheat.org/uploads/DHC_Market_Outlook_Insights_Trends_2023_81498577a7.pdf.

[9] Friman, O., Follo, P., Ahlberg, J., and Sjokvist, S. (2014). "Methods for Large-Scale Monitoring of District Heating Systems Using Airborne Thermography". In: *IEEE Transactions on Geoscience and Remote Sensing* 52(8), pp. 5175–5182. DOI: 10.1109/TGRS.2013.2287238.

[10] Fuchs, H. and Frommhold, W. (1991). "Leckortung auf Fernwärmeleitungen [Leak localisation in district heating pipes]". In: *Fraunhofer IBP Mitteilungen [Fraunhofer IBP Communications]*. Neue Forschungsergebnisse, kurz gefasst [New research findings, in brief] 18(201). URL: https://www.ibp.fraunhofer.de/content/dam/ibp/ibp-neu/de/dokumente/ibpmitteilungen/1-400/201-300/201_IBPmitteilung.pdf.

[11] Gonzalez-Salazar, M., Langrock, T., Koch, C., Spieß, J., Noack, A., Witt, M., Ritzau, M., and Michels, A. (2020). "Evaluation of Energy Transition Pathways to Phase out Coal for District Heating in Berlin". In: *Energies* 13(23), p. 6394. DOI: 10.3390/en13236394.

[12] Guo, G., Liu, S., Jia, D., Wang, S., and Wu, X. (2021). "Simulation of a leak's growth process in water distribution systems based on growth functions". In: *AQUA - Water Infrastructure, Ecosystems and Society* 70(4), pp. 521–536. DOI: 10.2166/aqua.2021.021.

[13] Gurklienė, R., Hogland, W., Knutsson, H., Lukosevičius, V., Lundström, J., Ohlsson, M., Rogala, A., Rybarczyk, P., and Zajaczkowski, K. (2023). *BSAM data driven proactive maintenance handbook: Smart maintenance of district heating networks*. Kalmar, Sweden: Linnaeus University. ISBN: 978-91-8082-036-3.

[14] Heating, W. C. Q. D. (2020). *Handbook on Planning of District Heating Networks*. Handbook. Version 1.0. Translation of version 1.2 (German, 2018). Bern, Switzerland: Swiss Federal Office of Energy. URL: https://www.verenum.ch/Dokumente/Handbook-DH_V1.0a.pdf.

[15] Heipke, C. and Tödter, J. (2020). *Drohnengestützte Thermografie als Basis der Asset- und Instandhaltungsstrategie von Fern- und Nahwärmenetzen [Drohne-based Thermography as an Asset- and Maintenancestrategy for District Heating Systems]*. Schlussbericht [Final report] IGF-Vorhaben Nr. 19768 N. Fernwärme-Forschungsinstitut in Hannover e.V. [District Heating Research Institute in Hannover E.V.] and Leibniz Universität Hannover [Leibniz University Hannover]. URL: https://www.fernwaerme.de/pdfdata/Schlussbericht_IGF_19768N.pdf.

[16] Hlebnikov, A., Volkova, A., Dzuba, O., Poobus, A., and Kask, Ü. (2010). "Damages of the Tallinn District Heating Networks and Indicative Parameters for an Estimation of the Networks General Condition". In: *Environmental and Climate Technologies* 5(-1), pp. 49–55. DOI: 10.2478/v10145-010-0034-3.

[17] Höhrová, P., Soviar, J., and Sroka, W. (2023). "Market Analysis of Drones for Civil Use". In: *LOGI – Scientific Journal on Transport and Logistics* 14(1), pp. 55–65. DOI: 10.2478/logi-2023-0006.

[18] Hossain, K., Villebro, F., and Forchhammer, S. (2020). "UAV Image Analysis for Leakage Detection in District Heating Systems using Machine Learning". In: *Pattern Recognition Letters* 140, pp. 158–164. DOI: 10.1016/j.patrec.2020.05.024.

[19] International Energy Agency (IEA) (2020). *World Energy Investment 2020*. Report. Paris, France: IEA. URL: https://www.iea.org/reports/world-energy-investment-2020.

[20] IEA (2022). *Technology and Innovation Pathways for Zero-carbon-ready Buildings by 2030*. Tech. rep. IEA. URL: https://www.iea.org/reports/technology-and-innovation-pathways-for-zero-carbon-ready-buildings-by-2030.

[21] KMR Service GmbH (2021). *Zubehoer [Component parts]*. Technical report 4. Trollhagen, Germany: KMR Service GmbH. URL: https://kmr-fernwaerme.de/wp-content/uploads/4-Zubehoer-_hell-80_.pdf (visited on 1 May 2025).

[22] Konstantin, P. and Konstantin, M. (2024). *Praxisbuch der Fernwärme und Fernkälteversorgung: Systeme, Netzaufbauvarianten, Kraft-Wärme und Kraft-Wärme-Kälte-Kopplung, Kostenstrukturen und Preisbildung [Practical guide to district heating and cooling: systems, network configurations, combined heat and power and combined heat, power and cooling, cost structures and pricing]*. 3rd ed. Berlin/Heidelberg, Germany: Springer. ISBN: 978-3-662-69526-5. DOI: 10.1007/978-3-662-69526-5.

[23] LANCIER Monitoring GmbH (2025). *LANCIER Monitoring*. URL: https://www.lancier-monitoring.de/en/ (visited on 23 May 2025).

[24] Latif, J., Shakir, M. Z., Edwards, N., Jaszczykowski, M., Ramzan, N., and Edwards, V. (2022). "Review on condition monitoring techniques for water pipelines". In: *Measurement* 193, p. 110895. DOI: 10.1016/j.measurement.2022.110895.

[25] Ljungberg, S.-A. and Rosengren, M. (1987). "Aerial Thermography - A Tool For Detecting Heat Losses And Defective Insulation In Building Attics And District Heating Networks". In: *Thermosense IX: Thermal Infrared Sensing for Diagnostics and Control*. Vol. 780. Orlando, United States: SPIE, pp. 257–343. DOI: 10.1117/12.940525.

[26] Losi, E., Manservigi, L., Spina, P. R., and Venturini, M. (2024). "Data-driven approach for the detection of faults in district heating networks". In: *Sustainable Energy, Grids and Networks* 38, p. 101355. DOI: 10.1016/j.segan.2024.101355.

[27] Losi, E., Manservigi, L., Spina, P. R., and Venturini, M. (2024). "Data-driven approach for the detection of faults in district heating networks". In: *Sustainable Energy, Grids and Networks* 38, p. 101355. DOI: 10.1016/j.segan.2024.101355.

[28] Lund, H., Werner, S., Wiltshire, R., Svendsen, S., Thorsen, J. E., Hvelplund, F., and Mathiesen, B. V. (2014). "4th Generation District Heating (4GDH): Integrating smart thermal grids into future sustainable energy systems". In: *Energy* 68, pp. 1–11. DOI: 10.1016/j.energy.2014.02.089.

[29] Madura, H. and Kołodziejczyk, M. (2005). "Influence of sun radiation on results of non-contact temperature measurements in far infrared range". In: *Opto-electronics Review* 13, pp. 253–257.

[30] Mao, D., Wang, P., Fang, Y.-P., and Ni, L. (2024). "Understanding District Heating Networks Vulnerability: A Comprehensive Analytical Approach with Controllability Consideration". In: *Sustainable Cities and Society* 101, p. 105068. DOI: 10.1016/j.scs.2023.105068.

[31] Mazhar, A. R., Liu, S., and Shukla, A. (2018). "A state of art review on the district heating systems". In: *Renewable and Sustainable Energy Reviews* 96, pp. 420–439. DOI: 10.1016/j.rser.2018.08.005.

[32] Meteostat (2024). *Meteostat: Free Historical Weather and Climate Data*. Weather data for Germany sourced via the German Weatherservice (Deutscher Wetterdienst). URL: https://meteostat.net (visited on 21 May 2025).

[33] Murtazin, I., Kozhevnikov, M., and Starikov, E. (2021). "Development and application of methods of internal inspection of district heating networks". In: *International Journal of Energy Production and Management* 6(1), pp. 56–70. DOI: 10.2495/EQ-V6-N1-56-70.

[34] Pech, K., Stelling, N., Karrasch, P., and Maas, H.-G. (2013). "GENERATION OF MULTITEMPORAL THERMAL ORTHOPHOTOS FROM UAV DATA". In: *The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences* XL-1/W2, pp. 305–310. DOI: 10.5194/isprsarchives-XL-1-W2-305-2013.

[35] Persson, U. and Werner, S. (2011). "Heat distribution and the future competitiveness of district heating". In: *Applied Energy* 88(3), pp. 568–576. DOI: 10.1016/j.apenergy.2010.09.020.

[36] Rafati, A. and Shaker, H. R. (2024). "Predictive maintenance of district heating networks: A comprehensive review of methods and challenges". In: *Thermal Science and Engineering Progress* 53, p. 102722. DOI: 10.1016/j.tsep.2024.102722.

[37] Rafati, A. and Shaker, H. R. (2024). "Predictive maintenance of district heating networks: A comprehensive review of methods and challenges". In: *Thermal Science and Engineering Progress* 53, p. 102722. DOI: 10.1016/j.tsep.2024.102722.

[38] Roscher, H. (2023). "Rehabilitation von Fernwärmekanälen und Fernwärmeleitungen [Rehabilitation of district heating ducts and district heating pipes]". In: *Rohrleitungen 2 [Pipes 2]*. Ed. by Horlacher, H.-B. and Helbig, U. 3rd ed. Berlin, Germany: Springer, pp. 1199–1231. DOI: 10.1007/978-3-662-60804-3_69.

[39] Shan, X., Wang, P., and Lu, W. (2017). "The reliability and availability evaluation of repairable district heating networks under changeable external conditions". In: *Applied Energy* 203, pp. 686–695. DOI: 10.1016/j.apenergy.2017.06.081.

[40] Sledz, A. and Heipke, C. (2021). "Thermal Anomaly Detection Based on Saliency Analysis from Multimodal Imaging Sources". In: *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences* V-1-2021, pp. 55–64. DOI: 10.5194/isprs-annals-V-1-2021-55-2021.

[41] Thornley, J. H. and France, J. (2005). "An open-ended logistic-based growth function". In: *Ecological Modelling* 184(2-4), pp. 257–261.

[42] Tuikka, L. (2024). "Leak Detection Systems in Finnish District Heating Network". Bachelor thesis. Tampere, Finland: Tampere University of Applied Sciences. URL: https://www.theseus.fi/bitstream/handle/10024/872261/Tuikka_Lijun.pdf.

[43] Verband Fernwärme Schweiz [Association for District Heating in Switzerland] (2022). *Leitfaden Fernwärme / Fernkälte [Guide to district heating / cooling]*. Schlussbericht [Final report]. Version 1.3. Bern, Switzerland: Bundesamt für Energie [Swiss Federal Office of Energy]. URL: https://www.thermische-netze.ch/fileadmin/user_upload/Dokumente/Publikationen/Downloads/Leitfaden_Fernwaerme_Fernkaelte_03-2022.pdf.

[44] Vollmer, E., Benz, M., Kahn, J., Klug, L., Volk, R., Schultmann, F., and Götz, M. (2025). "Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering". In: *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 18, pp. 6206–6216. DOI: 10.1109/JSTARS.2025.3537330.

[45] Vollmer, E., Ruck, J., Volk, R., and Schultmann, F. (2024). "Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods". In: *Automation in Construction* 168, p. 105709. DOI: 10.1016/j.autcon.2024.105709.

[46] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[47] Vollmer, E., Volk, R., and Schultmann, F. (2023). "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems". In: *International Journal of Remote Sensing* 44(23), pp. 7263–7293. DOI: 10.1080/01431161.2023.2242586.

[48] Werner, S. (2017). "International review of district heating and cooling". In: *Energy* 137, pp. 617–631. DOI: 10.1016/j.energy.2017.04.045.

[49] Winkens, H.-P. (1993). *Fernwärmespeicherung, -transport und verteilung [District heating storage, transportation and distribution]*. Project report 4. Mannheim: Institut für Energiewirtschaft und Rationelle Energieanwendung, Universität Stuttgart [Institute of Energy Economics and Rational Energy Use, University of Stuttgart].

[50] Wojdyga, K. and Chorzelski, M. (2017). "Chances for Polish district heating systems". In: *Energy Procedia* 116, pp. 106–118. DOI: 10.1016/j.egypro.2017.05.059.

[51] Woods, P. (2023). *An Introduction to District Heating and Cooling: Low carbon energy for buildings*. IOP Publishing. ISBN: 978-0-7503-5286-4. DOI: 10.1088/978-0-7503-5286-4.

[52] El-Zahab, S. and Zayed, T. (2019). "Leak detection in water distribution networks: an introductory overview". In: *Smart Water* 4(1), p. 5. DOI: 10.1186/s40713-019-0017-x.

[53] Zhou, S., O'Neill, Z., and O'Neill, C. (2018). "A review of leakage detection methods for district heating networks". In: *Applied Thermal Engineering* 137, pp. 567–574. DOI: 10.1016/j.applthermaleng.2018.04.010.

[54] Zhou, S., O'Neill, Z., and O'Neill, C. (2018). "A review of leakage detection methods for district heating networks". In: *Applied Thermal Engineering* 137, pp. 567–574. DOI: 10.1016/j.applthermaleng.2018.04.010.

[55] Žutautaite, I., Augutis, J., Krikštolaitis, R., Dundulis, G., Valinčius, M., and Rimkevičius, S. (2016). "Risk and reliability assessment of the district heating network methodology with case study". In: *Risk, Reliability and Safety: Innovating Theory and Practice*. Ed. by Walls, L., Revie, M., and Bedford, T. Boca Raton, United States: CRC Press, pp. 2578–2585. DOI: 10.1201/9781315374987-391.