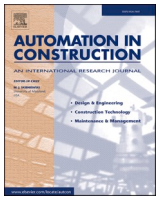
Automation in Construction 168 (2024) 105709

Contents lists available at ScienceDirect

Automation in Construction

journal homepage: www.elsevier.com/locate/autcon



Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods

Elena Vollmer *, Julian Ruck, Rebekka Volk, Frank Schultmann

Karlsruhe Institute of Technology (KIT), Institute for Industrial Production (IIP), Hertzstr. 16, Karlsruhe 76187, Baden-Wuerttemberg, Germany

ARTICLE INFO

Keywords:
Anomaly detection
Thermography
Image processing
Unmanned aircraft
Remote sensing
District heating networks

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.

1. Introduction

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 [1]. Accounting for approximately 30% of the world's energy demand, building operation – specifically heating – is primarily responsible for these sizeable requirements [1]. 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 [2]. 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 [3].

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 [4]. In Denmark, for instance, such networks currently already supply two-thirds of the population with 89% climate-neutral heat [3]. 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 [3]. 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 [5,6]. 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 [5]. 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 [5]. It is centred around Ljungberg and Rosengren [7]'s and Axelsson [8]'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

* Corresponding author.
E-mail address: elena.vollmer@kit.edu (E. Vollmer).

https://doi.org/10.1016/j.autcon.2024.105709
Received 2 May 2024; Received in revised form 12 August 2024; Accepted 12 August 2024
Available online 2 October 2024

Acronyms

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

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 [9]. 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 [5,6]. Moreover, the selection of a robust leakage detection method is essential for success [5]: 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 [10]. 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.

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. [6] 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. [11] instead use building data from OpenStreetMap and additionally integrate a feature-based machine learning (ML) classifier to improve false alarm reduction.

Sledz et al. [12] 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. [13] and Zhong et al. [14] implement Itti et al. [15]'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 [16] instead modify Itti et al. [15]'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 [17]'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. [19] and Hossain et al. [20] 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. [11]'s.

Most recently, Vollmer et al. [10] 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. [21] for image binarisation. False alarms are removed by size, shape, and temperature difference ΔT to surroundings and classified by ΔT severity.

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 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. [10] have made both their code and a dataset available online [22].

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. To do so, we select three approaches from Section 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. [10] for identical data pre- and post-processing. This, too, is

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

Publication	Geographical information	Aerial vehicle	Flight height [m]	Infrared camera / sensor	Image count	Image resolution [pixels]
Friman et al. [6] Berg et al. [11]	15 Swedish and Norwegian cities	airplane	800	FLIR SC7000	>50,000	640 × 512
Xu et al. [13]	Gävle, Sweden	airplane	–	FLIR X8000sc	–	1280 × 1024
Zhong et al. [14]	Gävle, Sweden	UAV DJI S1000	120	FLIR Tau 2640	–	640 × 512
	Datong, China	UAV DJI S1000	150	FLIR Tau 2640	–	640 × 512
Hossain et al. [19]	7 Danish cities	UAV	–	–	27,050	–
Hossain et al. [20]	12 Danish cities	UAV	–	FLIR Tau 2640	243,082	640 × 512
Sledz et al. [12] Sledz and Heipke [16]	Hannover, Germany	UAV DJI M200	40	DJI Zenmuse XT2	290	640 × 512
Vollmer et al. [10]	Munich, Germany	UAV DJI M600	60	DJI Zenmuse XT2	3365	640 × 512

enhanced with a novel, universally applicable vignetting correction (VC) which significantly improves performance of, for instance, Vollmer et al. [10]'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 [23].

The paper is divided into five parts. Section 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 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. [10], including new imagery from Munich as well as Karlsruhe to create a more substantial basis for analyses. Section 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 5. Section 6 draws conclusions from the study, details limitations, and presents an outlook for future work.

2. Implemented methodologies

The following anomaly detection methods are implemented in this study: 1. Vollmer et al. [10]'s triangle-histogram-thresholding (THT) approach, inspired by Friman et al. [6], 2. Hossain et al. [20]'s local thresholding (LT) method based on filter kernels, 3. Sledz and Heipke [16]'s adaptation of Itti et al. [15]'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. [10]. 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 1) in three steps:

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

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. [10] 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 2.1.1), preceding all other steps.

2.1.1. Vignetting correction (VC)

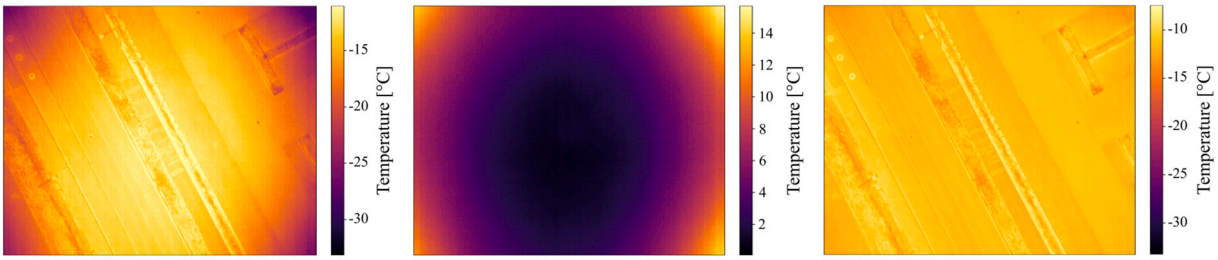
In thermography, the "vignetting" effect refers to a radial distortion where image corners and edges exhibit colder values than the centre. Despite thermal cameras including automatic non-uniformity correction, the case study's TIRs significantly suffer from vignetting – an observation that aligns with field tests conducted by Yuan and Hua [24]. 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 [24].

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.

$$CM_{VC} = PWM - min(PWM) \quad (1)$$

$$T_{u_{VC}} = T_u + CM_{VC} \quad (2)$$

[1] For this reason, Friman et al. [6]'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.



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

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

The simplified approach is based on Yuan and Hua [24]'s observation that a pixel-by-pixel temperature average over all corrected images of a 30 min 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. (1)). A raw thermal image T_u can be corrected to $T_{u_{VC}}$ via Eq. (2)[3], as visualised in Fig. 1. Vignetting masks such as Fig. 1b vary greatly depending on the acquisition conditions, so it is paramount to calculate one for each dataset.

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 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 4.3, where optimal method variants are found.

2.2.1. Triangle-histogram-thresholding (THT)

2.2.1.1. Original Methodology by Vollmer et al. [10]. First implemented by Vollmer et al. [10] in the context of leakage detection, the adapted THT approach functions as this study's baseline. As suggested by Friman et al. [6], 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. [6] 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. [10] instead calculate a histogram and image-wise threshold per masked image T_m using an approach based on Zack et al. [21].

Fig. 2 shows how a histogram is created based on the temperature distribution of a masked image, with each class covering an interval of $0.5^{\circ}C$. 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.

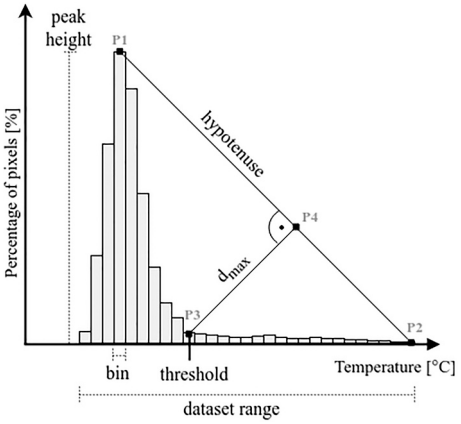


Fig. 2. Visualisation of the THT method [10].

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%.

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

2.2.2. Local thresholding (LT)

2.2.2.1. Original methodology by Hossain et al. [19,20]. While Hossain et al. [19,20]'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. (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. [20]. 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. [20].

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

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

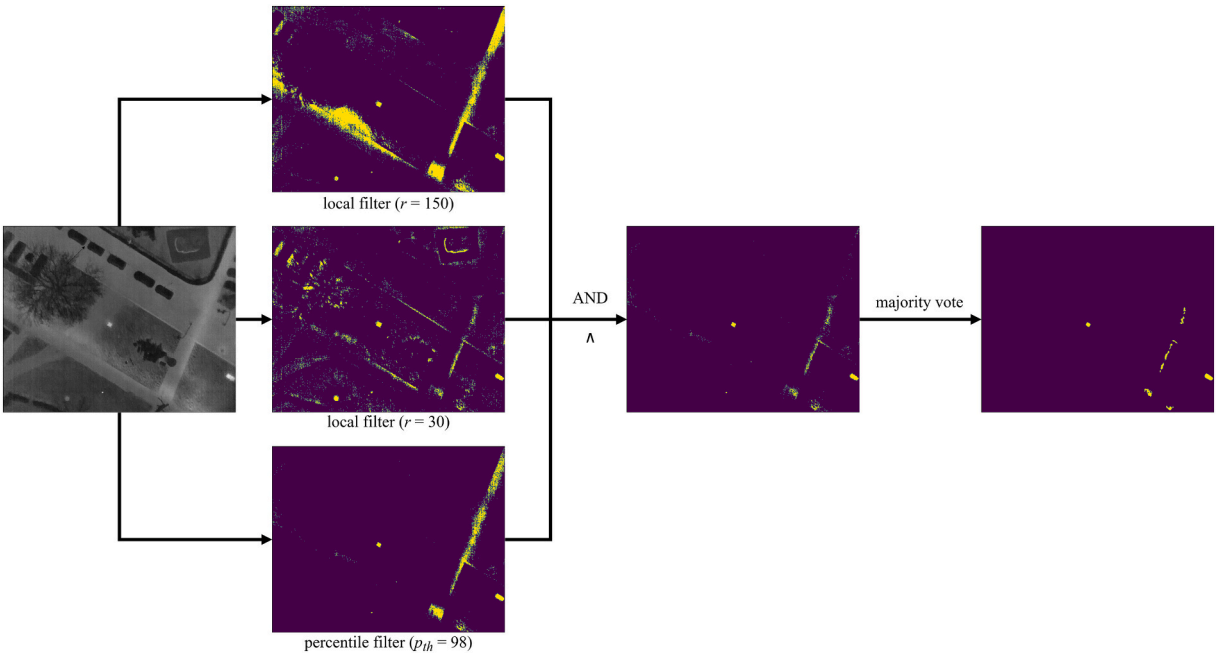


Fig. 3. Visualisation of the adapted LT process.

A second filter combining vertical and horizontal Sobel edge detectors calculates a gradient image G. Hossain et al. [20] 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}.

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

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

2.2.2.2. Adaptations for this study. An analysis in the context of this case study reveals several method problems. Hossain et al. [20]'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. 3).

Instead of a single temperature-based mask generated with $r = 100$, masks are created for every radius $r \in R, R = \{r_1, \ldots, r_n\}$. 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.

2.2.3. Saliency mapping (SM)

2.2.3.1. Original methodology by Itti et al. [15], Xu et al. [13], and Zhong et al. [14]. Saliency analyses model the human brains' attention directing mechanisms for visual stimuli to find conspicuous image regions [25]. As thermal anomalies stand out in TIRs, Xu et al. [13] are the first to propose this method for leakage detection in DHSs. Based on Itti et al. [15], 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 [15]: 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 $\overline{C}$, and orientation $\overline{O}$ 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 $\sigma \in \{0, \ldots, 8\}$ 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. [14] 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 $\theta \in \{0^\circ, 45^\circ, 90^\circ, 135^\circ\}$.

Every feature map F is normalised via Eq. (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 $\overline{m}$ 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.

$$N(F) = \langle F \rangle_0^M \cdot (M - \overline{m})^2 \quad (5)$$

Xu et al. [13] and Zhong et al. [14] 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. [19] report similarly unsatisfactory results and attribute the deviations to the simplistic nature of Zhong et al. [14]'s imagery.

To binarise the saliency map, Zhong et al. [14] implement an adaptive thresholding technique called maximum entropy segmentation

[26]. 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.

2.2.3.2. Adapted Methodology by Sledz and Heipke [16]. Building on Itti et al. [15]'s approach, Sledz and Heipke [16] 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:

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

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 [16] binarise the saliency maps by implementing Dempster [17]'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.

2.2.3.3. Adaptations for this Study. An in-depth analysis highlights a shortcoming of Sledz and Heipke [16]'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 $\left[percentile\left(I, p_{clip}\right), \infty\right]$, limiting the lower bounds. Fig. 4 shows the impact this has.
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. 5 illustrates this.

A qualitative assessment of Zhong et al. [14]'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):

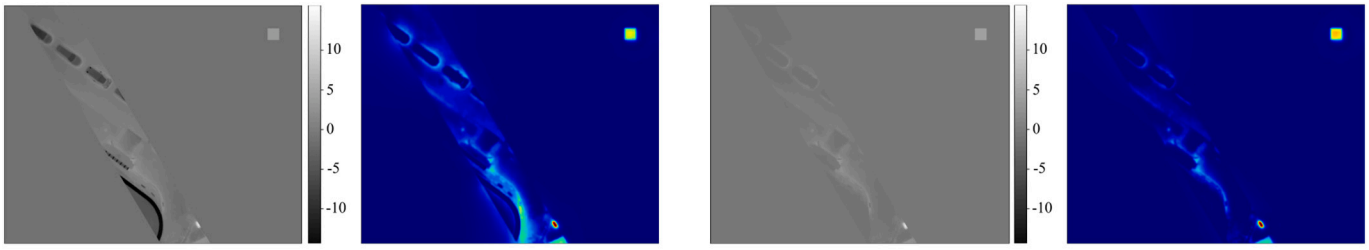
$$s_{\text{th}} = min\left(max\left(s_{ME}, s'_{min}, s_{ref} - \Delta s_{neg}\right), s_{ref} + \Delta s_{pos}\right) \tag{7}$$

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

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}.

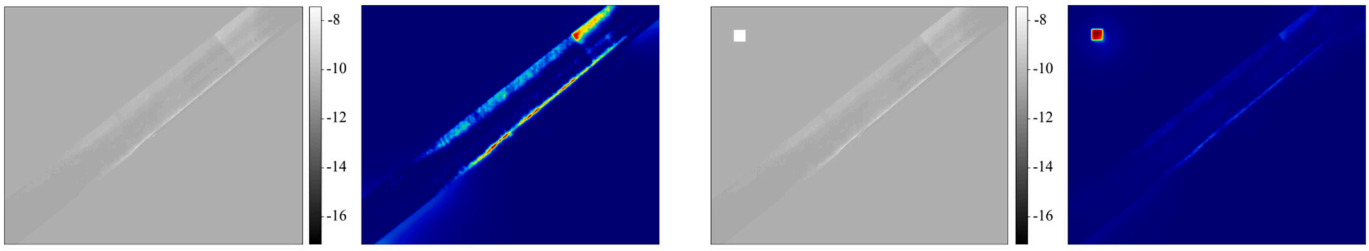
2.3. Leakage identification

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



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

Fig. 4. Comparison of generated saliency maps with and without active clipping.



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

Fig. 5. Influence of a reference point on the generted saliency maps.

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. [10], all labelled regions are therefore classified based on their size as small (≤ 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.

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. [10], 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. [12] assume a required delta of at least 5 °C based on literature research, while Berg et al. [11] report a confirmed leak with only 3 °C under specific ambient conditions. Vollmer et al. [10] 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 ($\Delta T < 5^{\circ}C$), moderate ($5^{\circ}C \leq \Delta T < 10^{\circ}C$), pronounced ($10^{\circ}C \leq \Delta T < 15^{\circ}C$), or critical ($15^{\circ}C \leq \Delta T$). We refrain from further categorisation (by classifier or geographical positions) as the purpose of study lies in comparing the anomaly detection methods themselves.

3. Case study

3.1. Data

The case study comprises 3750 UAS images of two DHSs from the German cities of Munich and Karlsruhe, as illustrated in Table 2. 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.

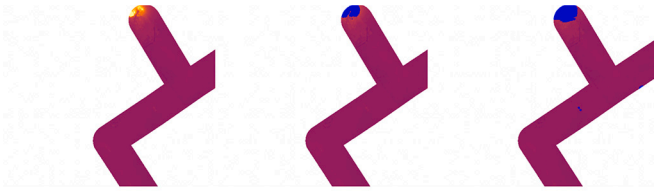
All flight routes were based on known DHS pipelines' positions. Both utilised Matrice unmanned aircrafts (UAs) supports automated flight along previously defined routes, with only take-off requiring manual handling. While nimble, the Matrice 300 RTK UA [27] is more susceptible to wind than the 600 Pro [28], 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 [29] – combines an optical 4 K 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 [29]. Images are stabilised via DJI's integrated gimbal [29]. 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.

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.

4. Preparing the evaluation

Comparing the anomaly detection methods from Section 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. 6 demonstrates, one could assign only the warmest pixels to an anomaly (as in 6b) or include less hot, neighbouring areas (as in 6c) – both fundamentally valid options. Even with strict annotation guidelines, manual labels will remain subject to some uncertainty. While the exact border between hot-



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

Fig. 6. Examples of possible segmentation masks for the same image.

Table 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

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.

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. 6. Both the created masks as well as the developed GUI labelling tool are published to Zenodo [23].

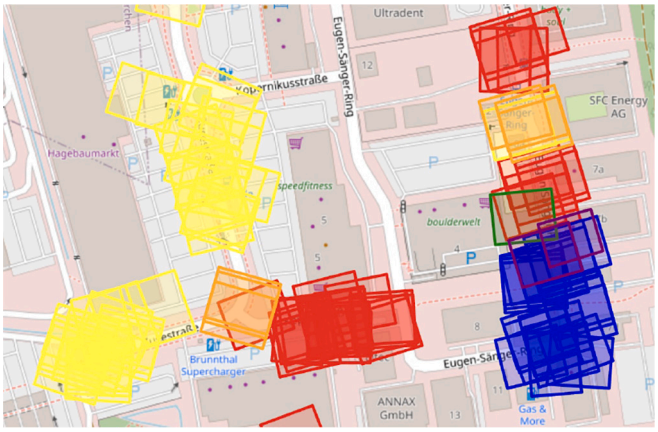


Fig. 7. Visualisation of the splitting procedure. Images in train are blue, validation red, and test yellow. Others are removed due to overlap. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 3
Overview of the evaluation dataset.

	Train	Validation	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		

4.1. Evaluation dataset

A total of 290 images are selected from the datasets presented in Section 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 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. 7 visualises the procedure for an exemplary area, while Table 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.

4.2. Metrics

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

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 [31]. 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

Table 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,[a] so F_2 is more suitable. [30]	$(1+\beta^2)\cdot\frac{P\cdot R}{(\beta^2\cdot P)+R}$
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**.[b]	
Detection rate 30 (DR_{30})	Corresponds to the DR of anomalies larger than 30 pixels.	

[a] Because anomaly existence takes precedence and FPs are removed in subsequent steps.
[b] 30% coverage ensures enough of the anomaly is included, also for candidate inspection.

Table 5
Overview of the parameters used in the grid search.

Method	Parameter	Description	Values
SM	ΔT	temperature delta between back- ground 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

Table 6
Grid search results for selected parameters and four metric optimisations.

Method	Parameter	MaxIoU	MaxIoU@85	MaxIoU@90	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

algorithms using 50 randomly chosen images from the training split, as this permits a feasible runtime. Table 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. [10]. Appendix A highlights the range the mentioned performance metrics assume on account of the parameter grid search via statistical analysis.

Four suitable parameter combinations are found for both methods by optimising Section 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 6 lists the identified, optimal parameter values.

5. Method evaluation and comparisons

5.1. Quantitative evaluation

The optimal method variants from Section 4.3 are quantitatively evaluated using the metrics described in 4.2. THT is included both in its original form [10] and this study's version, which incorporates VC (Section 2.1.1). The evaluation results (Table 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

Table 7
Results of the leakage detection method variants evaluated on validation and test sets. Best results are in bold.

Method	Configuration	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

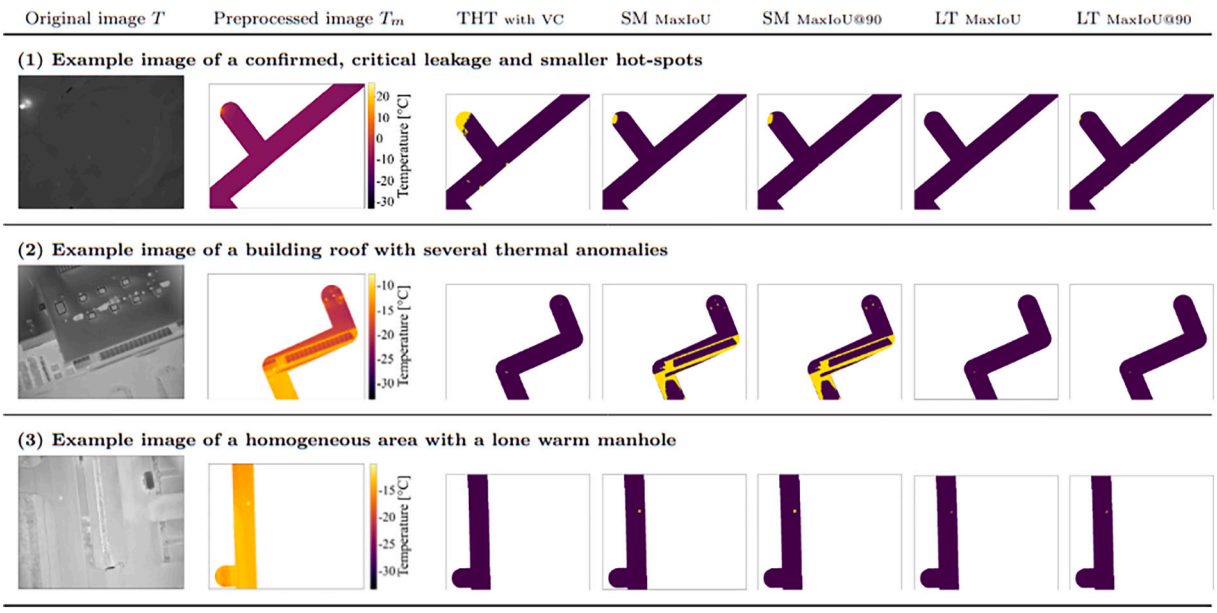


Fig. 8. Visualisation of segmentation mask results of the different configurations for three example scenarios.

parameters are selected. As expected, configurations that achieve a higher R or DR often score lower in P and IoU.

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.

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. 8 shows examples of common scenarios in their original form, preprocessed according to Section 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. 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. 8.2. Similar behaviour cannot be observed in any other method. Fig. 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. 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.

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 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 8 summarises the results after anomaly clustering (see Section 2.3.1) and classification by temperature difference ΔT into four categories (see Section 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 *MU*2, these are found to be leakages and manholes; for *KA*1, 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

Table 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

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.

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. [6]'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. [22] 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. [23].

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 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.

Funding

This work is supported by funding from the European Union through the AI4EOSC project (Horizon Europe) under Grant number 101058593.

CRediT authorship contribution statement

Elena Vollmer: Conceptualization, Methodology, Investigation, Data curation, Visualization, Writing – original draft, Writing – review & editing. **Julian Ruck:** Methodology, Investigation, Data curation, Software, Formal analysis, Visualization, Writing – review & editing. **Rebekka Volk:** Investigation, Writing – review & editing, Supervision. **Frank Schultmann:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Elena Vollmer reports financial support was provided by European Union. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

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].

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.

Appendix A. Grid search performance statistics

Table A.1 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 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 A.1
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 A.2 depicts mean and standard deviations of the leakage detection algorithm variants applied to validation and test datasets, as described in Section 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 A.3 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 A.1 owing to the more focused parameter choice.

Table A.2
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 A.3
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)
	$\lvert \ominus \rvert$	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	–

References

[1] United Nations Environment Programme, Global Alliance for Buildings and Construction, Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector, in: Technical Report, 2024, https://doi.org/10.59117/20.500.11822/45095. URL: https://wedocs.unep.org/20.500.11822/45095. Accessed 28 June 2024.

[2] The German Federal Ministry for Housing, Urban Development and Building, Gesetz für die Wärmeplanung und zur Dekarbonisierung der Wärmenetze [Law for heat planning and decarbonization of heat networks], URL, https://www.bmwsb.bund.de/SharedDocs/gesetzgebungsverfahren/Webs/BMWSB/DE/kommunale-waermeplanung.html, 2023. Enacted on 17 November 2023, effective from 1 January 2024. Accessed 28 June 2024.

[3] Arbeitsgemeinschaft Fernwärme (AGFW), J. Dornberger, Hauptbericht 2022 [Main Report 2022], Technical Report. AGFW, 2023. URL, https://www.agfw.

de/zahlen-und-statistiken/agfw-hauptbericht. Accessed 28 June 2024.
[4] International Energy Agency (IEA), World Energy Outlook 2023, Technical Report. IEA, 2023. URL, https://www.iea.org/reports/world-energy-outlook-2023. Accessed 28 June 2024.
[5] S. El-Zahab, T. Zayed, Leak detection in water distribution networks: an introductory overview, Smart Water 4 (2019) 1–23, https://doi.org/10.1186/s40713-019-0017-x.
[6] O. Friman, P. Follo, J. Ahlberg, S. Sjokvist, Methods for large-scale monitoring of district heating systems using airborne thermography, IEEE Trans. Geosci. Remote Sens. 52 (2014) 5175–5182, https://doi.org/10.1109/TGRS.2013.2287238.
[7] S.A. Ljungberg, M. Rosengren, 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, International Society for Optics and Photonics, SPIE, Orlando, United States, 1987, pp. 257–343, https://doi.org/10.1117/12.940525.
[8] S. Axelsson, Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography, IEEE Trans. Geosci. Remote Sens. 26 (1988) 686–692, https://doi.org/10.1109/36.7695.
[9] B.N. Coelho, UAVs and their role in future cities and industries, in: Smart and Digital Cities: From Computational Intelligence to Applied Social Sciences, Springer International Publishing, Cham, 2019, pp. 275–285, https://doi.org/10.1007/978-3-030-12255-3_17.
[10] E. Vollmer, R. Volk, F. Schultmann, Automatic analysis of UAS-based thermal images to detect leakages in district heating systems, Int. J. Remote Sens. 44 (2023) 7263–7293, https://doi.org/10.1080/01431161.2023.2242586.
[11] A. Berg, J. Ahlberg, M. Felsberg, Enhanced analysis of thermographic images for monitoring of district heat pipe networks, Pattern Recogn. Lett. 83 (2016) 215–223, https://doi.org/10.1016/j.patrec.2016.07.002.
[12] A. Sledz, J. Unger, C. Heipke, 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, 2020, pp. 499–505, https://doi.org/10.5194/isprs-archives-XLIII-B1-2020-499-2020.
[13] Y. Xu, X. Wang, Y. Zhong, L. Zhang, Thermal anomaly detection based on saliency computation for district heating system, in: 2016 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), Beijing, China, 2016, pp. 681–684, https://doi.org/10.1109/IGARSS.2016.7729171.
[14] Y. Zhong, Y. Xu, X. Wang, T. Jia, G. Xia, A. Ma, L. Zhang, Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions, ISPRS J. Photogramm. Remote Sens. 151 (2019) 207–222, https://doi.org/10.1016/j.isprsjprs.2019.02.021.
[15] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259, https://doi.org/10.1109/34.730558.
[16] A. Sledz, C. Heipke, Thermal Anomaly Detection Based on Saliency Analysis from Multimodal Imaging Sources. ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2021, pp. 55–64, https://doi.org/10.5194/isprs-annals-V-1-2021-55-2021.
[17] A.P. Dempster, A generalization of Bayesian inference, Journal of the Royal Statistical Society Series B (Statistical Methodology) 30 (1968) 205–247, https://doi.org/10.1007/978-3-540-44792-4_4.
[18] G. Shafer, A Mathematical Theory of Evidence, Princeton University Press, 1976, https://doi.org/10.2307/j.ctv10vm1qb.1. isbn:978-0691100425.
[19] K. Hossain, F. Villebro, S. Forchhammer, Leakage detection in district heating systems using UAV IR images: comparing convolutional neural network and ML classifiers, in: Proceedings of 27th European Signal Processing Conference, European Association for Signal Processing (EURASIP), A Coruña, Spain, 2019, https://doi.org/10.23919/EUSIPCO45326.2019.
[20] K. Hossain, F. Villebro, S. Forchhammer, UAV image analysis for leakage detection in district heating systems using machine learning, Pattern Recogn. Lett. 140 (2020) 158–164, https://doi.org/10.1016/j.patrec.2020.05.024.
[21] G.W. Zack, W.E. Rogers, S.A. Latt, Automatic measurement of sister chromatid exchange frequency, J. Histochem. Cytochem. 25 (1977) 741–753, https://doi.org/10.1177/25.7.70454.
[22] E. Vollmer, R. Volk, M. Vogl, Automatic analysis of UAS-based thermal images to detect leakages in district heating systems: source code and exemplary dataset, Zenodo (2023), https://doi.org/10.5281/zenodo.7851726.
[23] J. Ruck, E. Vollmer, M. Vogl, R. Volk, Finding district heating leakages in thermal imagery: a comparison of anomaly detection methods - source code and datasets, Zenodo (2024), https://doi.org/10.5281/zenodo.11085776.
[24] W. Yuan, W. Hua, A case study of Vignetting nonuniformity in UAV-based uncooled thermal cameras, Drones 6 (2022) 394, https://doi.org/10.3390/drones6120394.
[25] R. Cong, J. Lei, H. Fu, M.M. Cheng, W. Lin, Q. Huang, Review of visual saliency detection with comprehensive information, IEEE Trans. Circuits Syst. Video Technol. 29 (2019) 2941–2959, https://doi.org/10.1109/tcsvt.2018.2870832.
[26] J. Kapur, P. Sahoo, A. Wong, A new method for gray-level picture thresholding using the entropy of the histogram, Computer Vision, Graphics, and Image Processing 29 (1985) 273–285, https://doi.org/10.1016/0734-189x(85)90125-2.
[27] SZ DJI Technology Co. Ltd, Matrice 300 RTK, URL, https://enterprise.dji.com/matrice-300/specs, 2020. Accessed 28 June 2024.
[28] SZ DJI Technology Co. Ltd, Matrice 600 Pro, URL, https://www.dji.com/matrice600-pro, 2018. Accessed 28 June 2024.
[29] SZ DJI Technology Co. Ltd, Zenmuse XT 2: User Manual, URL, https://www.dji.com/downloads/products/zenmuse-xt2, 2018. Accessed 28 June 2024.
[30] S. Jadon, A survey of loss functions for semantic segmentation, in: 2020 IEEE Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB), IEEE, Via del Mar, Chile, 2020, pp. 1–7, https://doi.org/10.1109/cibcb48159.2020.9277638.
[31] D. Chicco, Ten quick tips for machine learning in computational biology, BioData Mining 10 (2017) 35, https://doi.org/10.1186/s13040-017-0155-3.