ISPRS Journal of Photogrammetry and Remote Sensing 228 (2025) 505–518

Contents lists available at ScienceDirect

ISPRS Journal of Photogrammetry and Remote Sensing

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

# Leak detection using thermal imagery: Deep learning versus traditional computer vision state-of-the-art

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

Dataset link: http://doi.org/10.5281/zenodo.14287864

*Keywords:*
Anomaly detection
District heating systems
Semantic segmentation
Thermal imagery
Transformers
Unmanned aircraft systems

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.

## 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 (International Energy Agency (IEA), 2023). 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 (Arbeitsgemeinschaft Fernwärme (AGFW), 2023). 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 (Friman et al., 2014). Considering current heat-related goals in the effort to limit anthropogenic global warming (United Nations Environment Programme, Global Alliance for Buildings and Construction, 2024), 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 (El-Zahab and Zayed, 2019). To this end, a thermography-based approach has emerged, centred around Axelsson (1988)'s and Ljungberg and Rosengren (1988)'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 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 (Friman et al., 2014).

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 (Vollmer et al., 2023). 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 (Vollmer et al., 2023, 2025a). 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 (Pang et al., 2022), this application works at a finer spatial resolution and defines anomalies as clusters of warm pixels – also known as hot-spots – within

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

https://doi.org/10.1016/j.isprsjprs.2025.06.006
Received 8 January 2025; Received in revised form 22 May 2025; Accepted 5 June 2025
Available online 28 July 2025

**Acronyms**

| | |
|---|---|
| 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 |

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 (Vollmer et al., 2025a). Together with data processing and false alarm removal, anomaly detection is therefore considered one of the key steps in TIR-based DHS leak detection (Vollmer et al., 2023, 2024).

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 (Goodfellow et al., 2017). In the case of image processing, DL in particular has become highly relevant owing to its versatility and performance (Goodfellow et al., 2017). 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 (He et al., 2021). Other approaches for general pipeline inspection, such as via acoustic emission signal analysis, already implement DL to great effect (Siddique et al., 2023). Despite this, previous work on TIR-based DHSs leak detection has focused on traditional CV methods to identify anomalies, such as saliency mapping, local thresholding, and histogram-based methods (Vollmer et al., 2024). Machine learning (ML) or, seldomly, DL has so far only been used for their classification (Berg et al., 2016; Hossain et al., 2020; Vollmer et al., 2024), mainly because annotation creation is an exceptionally labour-intensive undertaking (Vollmer et al., 2023; Cheng et al., 2024).

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 (Vollmer et al., 2024).
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. (2024)'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 (Ruck et al., 2025) and code[1] will be published alongside this paper to ensure reproducibility.

To this end, the paper is structured as follows: Section 2 discusses related literature and the research gap; Section 3 describes methodology, from data processing over model selection to implementation; Section 4 includes a thorough model evaluation and simultaneous comparison to classical CV methods, while Section 5 concludes the paper with an outlook.

## 2. Related work

After Axelsson (1988) and Ljungberg and Rosengren (1988)'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 (Vollmer et al., 2024).

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 (Vollmer et al., 2024).

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. (2014) and Berg et al. (2016) 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 (Berg et al., 2016).
2. Xu et al. (2016) and Zhong et al. (2019) develop a saliency mapping (SM) approach, with Sledz and Heipke (2021) suggesting modifications. They also employ image georeferencing and masking.

[1] https://www.github.com/emvollmer/TASeg

3. Sledz et al. (2020) 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. (2019) and Hossain et al. (2020) 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. (2016)'s random forest.
5. Vollmer et al. (2023) utilise an enhanced 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. (2024) 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 (Ruck et al., 2024). Vollmer et al. (2024)

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. (2024)'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.

## 3. Methodology

### 3.1. Data preparation

To allow for comparable results, Vollmer et al. (2024)'s UAS-based TIR datasets (Ruck et al., 2024) 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 (Vollmer et al., 2023). This ensures that leaks can be clearly distinguished as thermal anomalies from their surroundings (Vollmer et al., 2023). Furthermore, flights are only performed in dry weather conditions, as rain and snowfall greatly diminish TIR image quality (Vollmer et al., 2023). 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 (Vollmer et al., 2023, 2024). This provides every image $T$ with a full corrected temperature array $T_u$ (unmasked) and one reduced to the relevant areas $T_m$ (masked) (Vollmer et al., 2024). The combination of stringent acquisition guidelines and image preprocessing mean that the temperature distributions across datasets are comparatively similar (see Appendix A). 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 (Vollmer et al., 2025a). If one wishes to use inputs of differing channel counts, an additional convolutional layer can be included to map these to the expected three (Vollmer et al., 2025a). An ablation study was conducted to identify the most suitable combination of input channels, which is summarised in Appendix B. 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 the unmasked $T_u$ provides additional context information, in particular to areas close to the masking border. Given the TIR resolution of $512 \times 640$, 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. (2024) 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. (2024) – 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 (Vollmer et al., 2023). In its general form, the algorithm draws a triangle hypotenuse from the histogram's peak to its outer right edge (Vollmer et al., 2023). 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 (Vollmer et al., 2023). 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 (Vollmer et al., 2023).

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. (2024). Table 1 provides an overview of Vollmer et al. (2024)'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 (Vollmer et al., 2024), one dataset – specifically *MU2* – is excluded from all but the test set (Vollmer et al., 2024). This allows for an unbiased assessment during model evaluation.

**Table 1**
Overview of the automatically generated and manually annotated datasets. Data based on Ruck et al. (2024).

| | 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 | | |

[2] For their AI classification model training, Hossain et al. (2020) report annotating 243,082 images by hand.

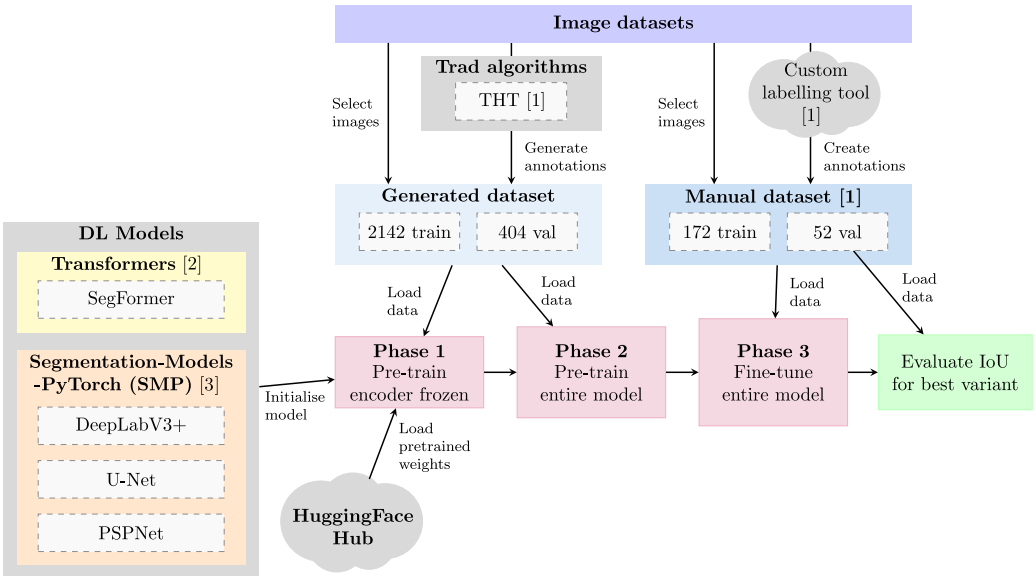


**Fig. 1.** Visualisation of the developed multi-stage DL model training procedure. Reference [1] refers to Vollmer et al. (2024), [2] to Wolf et al. (2020), and [3] to Iakubovskii (2019).

### 3.2. Model development

#### 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 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 (Vollmer et al., 2023). 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 (Pang et al., 2022), as it diverges somewhat from the classical, predominantly classification-based field of outlier detection (Pang et al., 2022). However, to remain consistent in our comparison with traditional CV methods (Vollmer et al., 2023), we will refer to it using the umbrella term anomaly detection.

A comparatively new research field, semantic image segmentation builds upon Long et al. (2015)'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 (Chen et al., 2017). Numerous architecture adaptations have led to significant performance improvements and established models like the U-Net (Ronneberger et al., 2015) (encoder–decoder with skip connections), PSPNet (Zhao et al., 2017) (pyramid pooling), and DeepLabV3+ (Chen et al., 2018) (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 (Liu et al., 2021).

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 (Vaswani et al., 2017). 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. (2021)'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 (Liu et al., 2021).

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 (Ronneberger et al., 2015), PSPNet (Zhao et al., 2017), and DeepLabV3+ (Chen et al., 2018) – each representing a different NN type. The results of these experiments are presented in Section 3.2.4, after discussing the details of the training procedure and configuration.

#### 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 (Safonova et al., 2023). The developed methodology, as visualised in Fig. 1, is model independent, easily transferable, and can be used across architectures to improve performance when facing similarly challenging use cases.

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 pre-trained ones from training on public databases. While most available

**Table 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 | $L_J(y,\hat{y}) = 1 - \frac{(y\cdot\hat{y})+\varepsilon}{(y+\hat{y}-y\cdot\hat{y})+\varepsilon}$ | 67.6 | 79.6 | 79.0 | 82.5 |
| Dice | $L_D(y,\hat{y}) = 1 - \frac{(2y\cdot\hat{y})+\varepsilon}{(y+\hat{y})+\varepsilon}$ | **69.1** | 80.7 | 80.0 | **83.6** |
| Tversky ($\alpha = 0.3, \beta = 0.7$) | $L_T(y,\hat{y}) = 1 - \frac{(y\cdot\hat{y})+\varepsilon}{(y\cdot\hat{y})+\alpha\cdot FN+\beta\cdot FP+\varepsilon}$ | 68.1 | **81.5** | **81.9** | 80.1 |

weights are based on RGB datasets, studies such as Li et al. (2021) 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 (Zhou et al., 2019) 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. 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 (Cheng et al., 2024). 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 (Goodfellow et al., 2017). 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 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.

### 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 (Cheng et al., 2024; Nogueira et al., 2024). 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 (Nogueira et al., 2024). Highly skewed data are problematic for all manner of DL, including semantic segmentation, as they bias a model towards the majority class (Johnson and Khoshgoftaar, 2019). The selection of a suitable loss function helps counteract this unwanted effect by emphasising the importance of the minority class during training (Johnson and Khoshgoftaar, 2019; Jadon, 2020).

Table 2 summarises an ablation study using four such functions. Performance is measured based on the common semantic segmentation metrics IoU, R, P, and $F_\beta$ score[4] recall (R) and precision (P). 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 (Jadon, 2020), 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 $\varepsilon$ prevents a division by zero (Duque-Arias et al., 2021). Dice loss is derived in analogous fashion from the dice coefficient (Sudre et al., 2017). Lastly, Tversky generalises dice loss for refined control over weighting of FP (via $\alpha$) and FN (via $\beta$) (Salehi et al., 2017). With $\beta > \alpha$ and $\alpha + \beta = 1$, we penalise FN more than FP and increase R, as is desired for leak detection.

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 (Xie et al., 2021), 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. (2021), an AdamW optimiser (Loshchilov and Hutter, 2017) 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 image amounts. In the context of image processing, common techniques include mirroring, enlargement, section rotations, or a combination thereof Goodfellow et al. (2017). 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.

[3] For the exact values, see Appendix C and Section 3.2.4.

[4] Here $F_2$, as R takes precedence over P in leak detection (Vollmer et al., 2024).

[5] This prevents individual training steps from inciting drastic changes which impede the finding of local optima (Goodfellow et al., 2017).

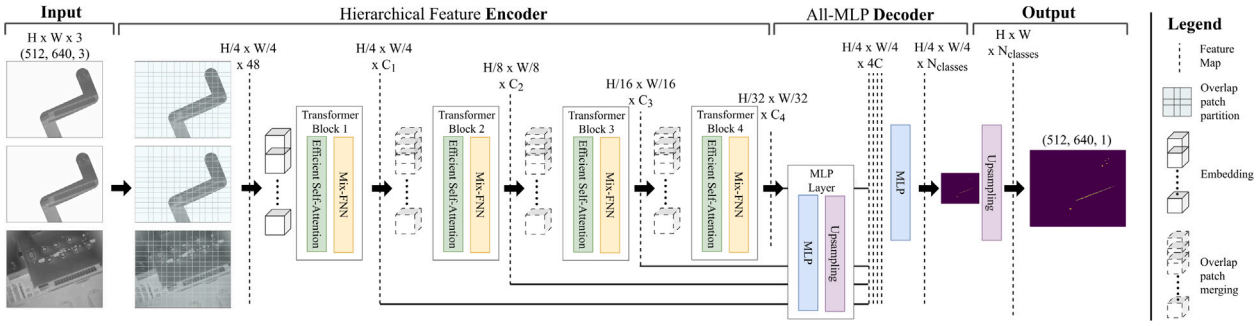


**Fig. 2.** SegFormer architecture adapted to the TIR anomaly detection problem, with every transformer block containing a FFN and the decoder including MLP layers. Image based on Xie et al. (2021).

**Table 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 |

#### 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 3.2.1 and analogous to Xie et al. (2021), results for the conventional semantic segmentation CNN DeepLabV3+ (Chen et al., 2018) with a ResNet101 encoder pre-trained on ImageNet (Deng et al., 2009) is included. Both U-Net (Ronneberger et al., 2015) and PSPNet (Zhao et al., 2017) 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 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 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. (2021) in two significant ways. Firstly, the fact that SegFormers B2 and B3 achieve the highest IoUs contradicts literature such as Xie et al. (2021), 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. (2021) report that their similarly configured DeepLabV3+ scores a 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 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 (Vollmer et al., 2024). 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. 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 $4 \times 4$ 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 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.

### 3.3. Implementation

All DL models are implemented via the 'PyTorch' (Paszke et al., 2019) and 'PyTorch-Lightning' (Falcon and The PyTorch Lightning team, 2019) libraries. To ensure access to a wide range of model architectures, the 'Segmentation-Models-PyTorch (SMP)' (Iakubovskii, 2019) and 'Hugging-Face-Transformers' (Wolf et al., 2020) 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. (2020)'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

[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 4**
Quantitative results of the SegFormer compared to traditional CV algorithms from Vollmer et al. (2024), evaluated on the manually annotated validation and test sets (see Table 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 |

(NVIDIA Corporation, 2022) 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 3.2.4 took approximately 72 min to train. Energy requirements, measured via the 'perun' package (Gutiérrez Hermosillo Muriedas et al., 2023), 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]

## 4. Evaluation and comparison

The SegFormer model described in Section 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. (2024).

### 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. (2024). The last two are custom metrics, defining the proportion of anomalies identified out of all and those larger than 30 pixels respectively (Vollmer et al., 2024).

Table 4 shows the results for the SegFormer applied to the validation and test splits of the manual dataset, coined evaluation dataset in Vollmer et al. (2024). 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 (Vollmer et al., 2024), the binarisation should strive to minimise FNs. This means the selection must tend towards mid- and low-range values.[8] A similar observation is made by Alkan and Karasaka (2023) 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.

To allow for a direct comparison, Table 4 also includes results of the best performing variants among the classical algorithms THT, SM, and LT (Vollmer et al., 2024). 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.

### 4.2. Qualitative evaluation

A qualitative evaluation, shown in Fig. 3, is performed on the same exemplary images used by Vollmer et al. (2024). 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 (Vollmer et al., 2024).

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. 3.1), the model identifies all hot-spots without classifying as large a pixel area as THT. In scenes where a uniform threshold does not allow for sufficiently accurate differentiation and THT struggles (Figs. 3.2 and 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.

### 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 (Holzinger et al., 2022). Of the large variety of existing xAI methods, explanations are most often created through visualisation techniques, commonly CAM-based methods (Islam et al., 2022). 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 (Selvaraju et al., 2020). Although most techniques for explaining image-based AI models focus on classification tasks (Gipiškis et al., 2024), this method was adapted to semantic segmentation via implementations such as the Seg-Grad-CAM (Vinogradova et al., 2020).

[7] For reference, Gowda et al. (2024) 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.

[8] While high thresholds may ensure low FP rates, this comes at the detriment of FN.

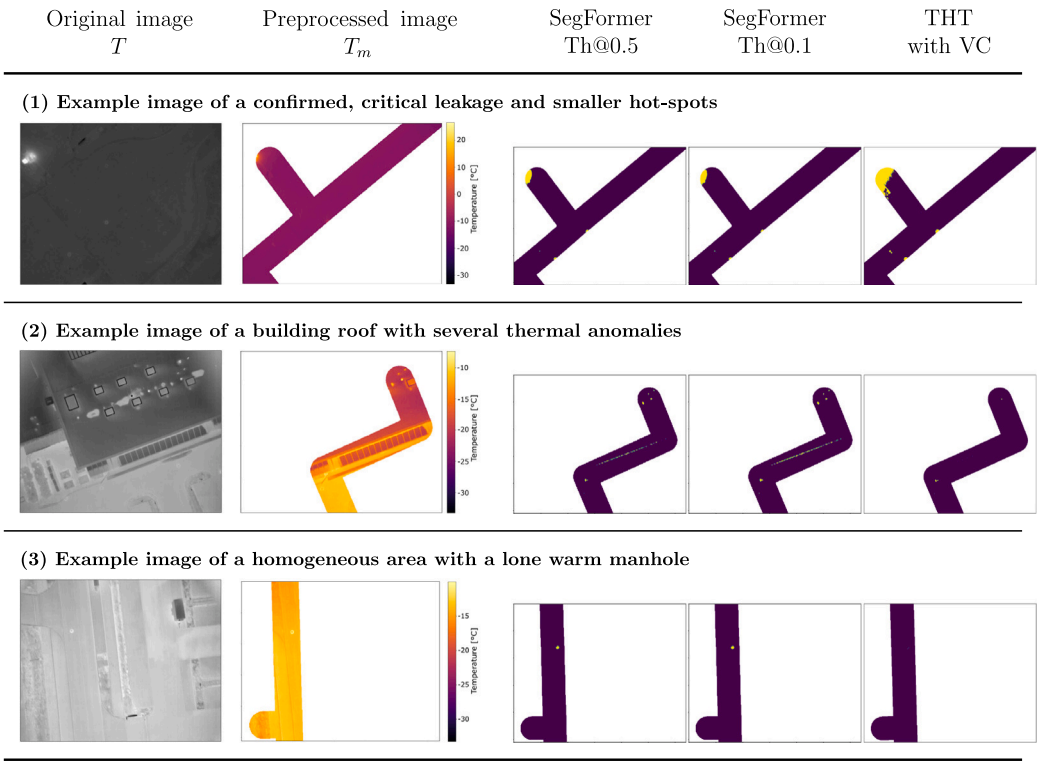


**Fig. 3.** Segmentation masks predicted by the SegFormer for three example scenarios, with (Vollmer et al., 2024)'s winning THT for comparison.

While both Sections 4.1 and 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 (2021)'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. 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.

The first image 4.1 features an example in which prediction and annotation masks are equivalent. The segmentation Grad-CAM outputs in 4.1.e) and f) showcase how the choice of three-channel input constellations, consisting of duplicated 4.1.a) and single 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.

The second example 4.2 exhibits an output 4.2.d) that differs from the given ground truth 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 (Vollmer et al., 2025a).

### 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 $\leq \Delta T$) (Vollmer et al., 2024). As in Vollmer et al. (2024), 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 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 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

[9] To exemplify, the 0.5 threshold is utilised to balance anomaly amounts while ensuring high quantitative 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'.

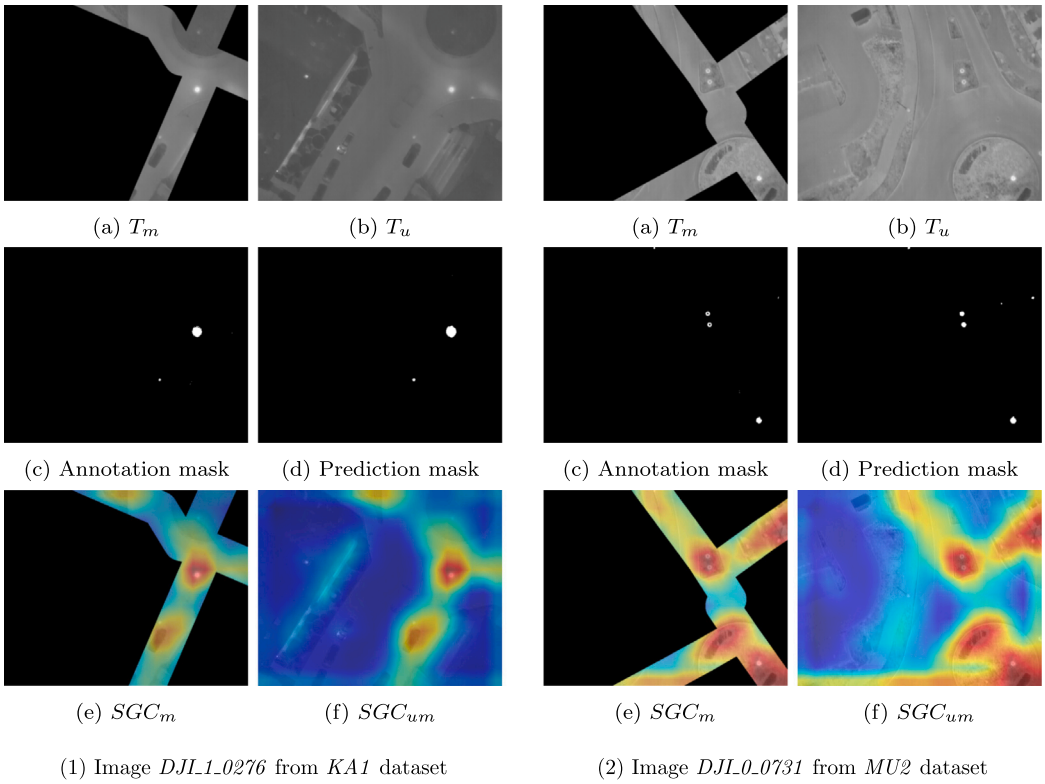


**Fig. 4.** Seg-Grad-CAM explanations for exemplary TIR input images.

**Table 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 |
| Classifiedby $\Delta T$ | 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 |
| Classifiedby type(manually) | leakage | 19 | 20 | 0 | 0 |
| | manhole | 61 | 82 | 52 | 51 |
| | car | 90 | 46 | 10 | 10 |
| | other | | | 14 | 19 |

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 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 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 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

[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.

**Table 6**
Pipeline run times exemplified on *KA1*, with 496 images and anomaly amounts listed in Table 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 |

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 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. (2021) 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.

## 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. (2024). 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 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.

[14] This value is given for the SegFormer-B2, for single-scale inference and a batch size of 16 on the ADE20K dataset (Xie et al., 2021).

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 (Safonova et al., 2023), 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. (2025b), 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 (Bayomi and Fernandez, 2023). Coupled with code optimisation to enable faster run times and potentially real-time implementation (Xie et al., 2021), 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.

## CRediT authorship contribution statement

**Elena Vollmer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Julian Ruck:** Writing – review & editing, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Rebekka Volk:** Writing – review & editing, Supervision, Funding acquisition. **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 through the AI4EOSC project (Horizon Europe) under Grant number 101058593. 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.

## Acknowledgements

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.

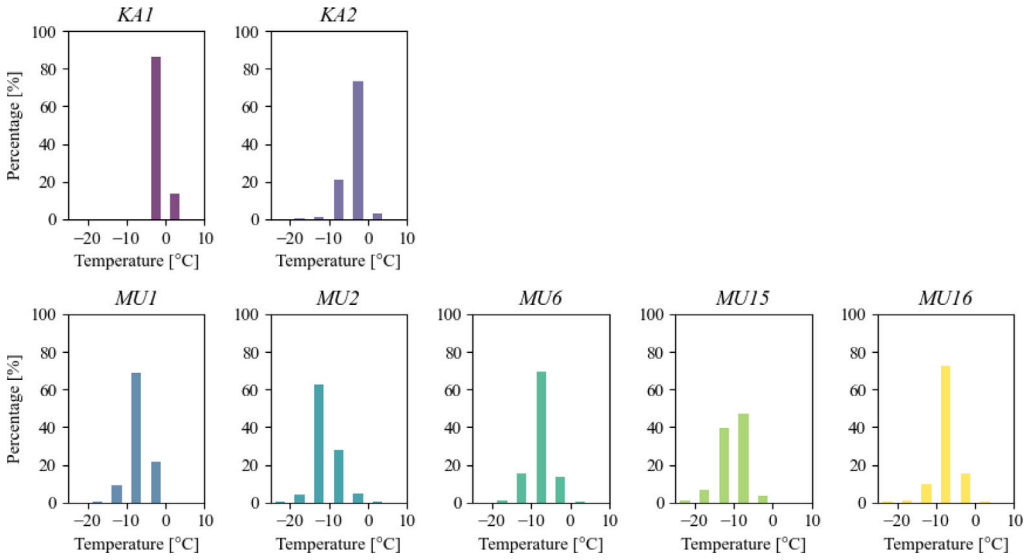


**Fig. A.5.** Histograms of temperature distributions per individual dataset.

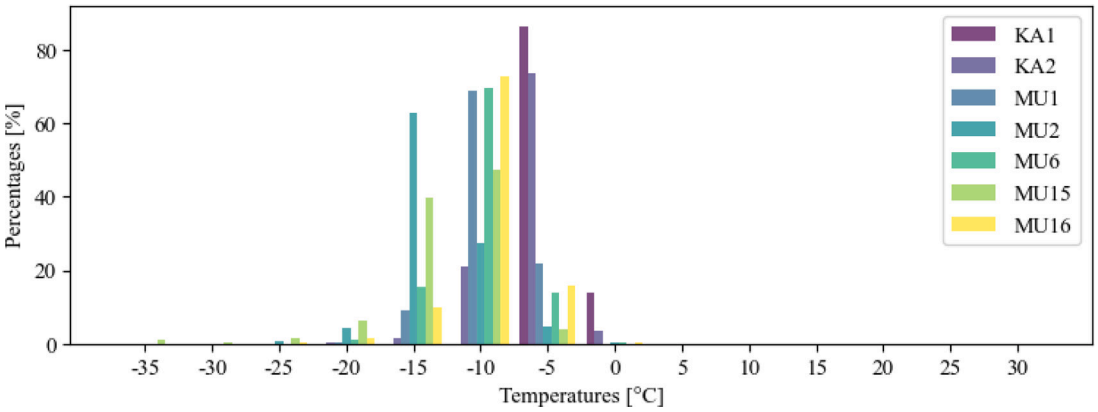


**Fig. A.6.** Histograms of temperature distributions per individual dataset, grouped in one plot.

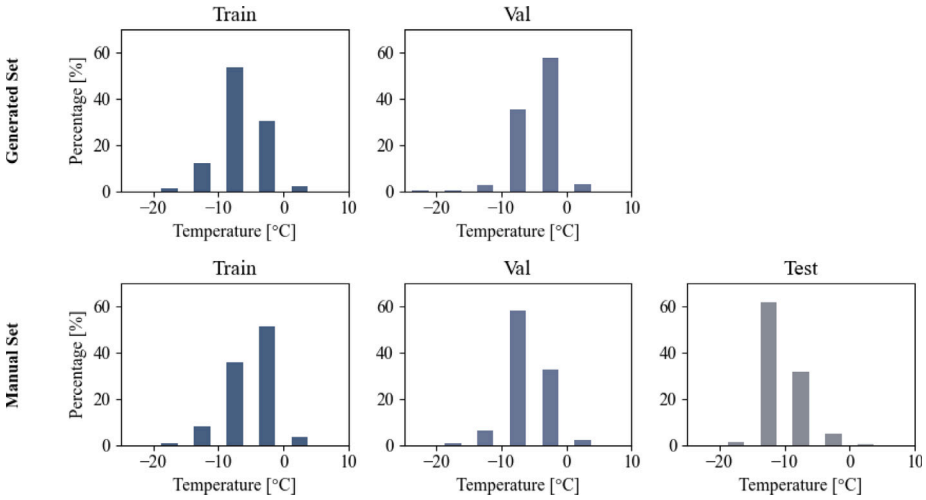


**Fig. A.7.** Histograms of temperature distributions in train, validation, and test splits for model training.

## Appendix A. 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. A.5, Fig. A.6 highlights how the majority of the data is similarly positioned between −15 °C to −5 °C.

Once combined into train, validation, and test splits for model training, the distributions become very similar, as highlighted in Fig. A.7.

**Table B.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 |

**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 |

**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 |

## Appendix B. 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 (Vollmer et al., 2025a). 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 B.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 ($T_m, T_m, T_u$).

## Appendix C. Impact of the multi-phase training procedure

Different experiments were conducted to assess the impact of various aspects of the developed multi-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 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.

## Data availability

All data, code, and configurations used in this work will be made available with this publication via Zenodo (Ruck et al., 2025) (http://doi.org/10.5281/zenodo.14287864) and GitHub (https://www.github.com/emvollmer/TASeg).

## References

Alkan, D., Karasaka, L., 2023. Segmentation of landsat-8 images for burned area detection with deep learning. Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLVIII-M-1-2023, 455–461. http://dx.doi.org/10.5194/isprs-archives-XLVIII-M-1-2023-455-2023.

Arbeitsgemeinschaft Fernwärme (AGFW), 2023. Consortium for District Heating, J. Dornberger, Hauptbericht 2022. Main Report 2022, Technical Report, AGFW, URL: https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht, (Accessed 19 December 2024).

Axelsson, S., 1988. Thermal modeling for the estimation of energy losses from municipal heating networks using infrared thermography. IEEE Trans. Geosci. Remote Sens. 26 (5), 686–692. http://dx.doi.org/10.1109/36.7695.

Bayomi, N., Fernandez, J.E., 2023. Eyes in the sky: Drones applications in the built environment under climate change challenges. Drones 7 (10), 637. http://dx.doi.org/10.3390/drones7100637.

Berg, A., Ahlberg, J., Felsberg, M., 2016. Enhanced analysis of thermographic images for monitoring of district heat pipe networks. Pattern Recognit. 83, 215–223. http://dx.doi.org/10.1016/j.patrec.2016.07.002.

Buslaev, A., Iglovikov, V.I., Khvedchenya, E., Parinov, A., Druzhinin, M., Kalinin, A.A., 2020. Albumentations: Fast and flexible image augmentations. Information 11 (2), http://dx.doi.org/10.3390/info11020125.

Chen, L.-C., Papandreou, G., Schroff, F., Adam, H., 2017. Rethinking atrous convolution for semantic image segmentation. http://dx.doi.org/10.48550/ARXIV.1706.05587.

Chen, L.-C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation. http://dx.doi.org/10.48550/arXiv.1802.02611.

Cheng, J., Deng, C., Su, Y., An, Z., Wang, Q., 2024. Methods and datasets on semantic segmentation for unmanned aerial vehicle remote sensing images: A review. ISPRS J. Photogramm. Remote Sens. 211, 1–34. http://dx.doi.org/10.1016/j.isprsjprs.2024.03.012.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., Fei-Fei, L., 2009. ImageNet: A large-scale hierarchical image database. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. IEEE, pp. 248–255. http://dx.doi.org/10.1109/cvpr.2009.5206848.

Duque-Arias, D., Velasco-Forero, S., Deschaud, J.-E., Goulette, F., Serna, A., Decencière, E., Marcotegui, B., 2021. On power Jaccard losses for semantic segmentation. In: Proc. Int. Jt. Conf. Comput. Vis. Imaging Comput. Graph. Theory Appl.. SCITEPRESS - Science and Technology Publications, pp. 561–568. http://dx.doi.org/10.5220/0010304005610568.

El-Zahab, S., Zayed, T., 2019. Leak detection in water distribution networks: An introductory overview. Smart Water 4 (1), 1–23. http://dx.doi.org/10.1186/s40713-019-0017-x.

Falcon, W., The PyTorch Lightning team, 2019. PyTorch Lightning. Zenodo, http://dx.doi.org/10.5281/zenodo.7469930.

Friman, O., Follo, P., Ahlberg, J., Sjokvist, S., 2014. Methods for large-scale monitoring of district heating systems using airborne thermography. IEEE Trans. Geosci. Remote Sens. 52 (8), 5175–5182. http://dx.doi.org/10.1109/TGRS.2013.2287238.

Gildenblat, J., contributors, 2021. PyTorch Library for CAM Methods. GitHub, https://github.com/jacobgil/pytorch-grad-cam.

Gipiškis, R., Tsai, C.-W., Kurasova, O., 2024. Explainable AI (XAI) in image segmentation in medicine, industry, and beyond: A survey. ICT Express 10 (6), 1331–1354. http://dx.doi.org/10.1016/j.icte.2024.09.008.

Goodfellow, I., Bengio, Y., Courville, A., Bach, F., 2017. Deep Learning. MIT Press, https://www.deeplearningbook.org/.

Gowda, S.N., Hao, X., Li, G., Gowda, S.N., Jin, X., Sevilla-Lara, L., 2024. Watt for what: Rethinking deep learning's energy-performance relationship. http://dx.doi.org/10.48550/arXiv.2310.06522.

Gutiérrez Hermosillo Muriedas, J.P., Flügel, K., Debus, C., Obermaier, H., Streit, A., Götz, M., 2023. Perun: Benchmarking energy consumption of high-performance computing applications. In: Cano, J., Dikaiakos, M.D., Papadopoulos, G.A., Pericàs, M., Sakellariou, R. (Eds.), Euro-Par 2023: Parallel Processing. Springer Nature Switzerland, Cham, pp. 17–31. http://dx.doi.org/10.1007/978-3-031-39698-4_2.

He, Y., Deng, B., Wang, H., Cheng, L., 2021. Infrared machine vision and infrared thermography with deep learning: A review. Infrared Phys. Technol. 116, 103754. http://dx.doi.org/10.1016/j.infrared.2021.103754.

Holzinger, A., Saranti, A., Molnar, C., Biecek, P., Samek, W., 2022. Explainable AI methods - a brief overview. In: Holzinger, A., Goebel, R., Fong, R., Moon, T., Müller, K.-R., Samek, W. (Eds.), XxAI - beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers. Springer International Publishing, Cham, pp. 13–38. http://dx.doi.org/10.1007/978-3-031-04083-2_2.

Hossain, K., Villebro, F., Forchhammer, S., 2019. Leakage detection in district heating systems using UAV IR images: Comparing convolutional neural network and ML classifiers. In: Proc. Eur. Signal Process. Conf.. Eur. Assoc. Signal Process. (EURASIP), https://orbit.dtu.dk/en/publications/leakage-detection-in-district-heating-systems-using-uav-ir-images.

Hossain, K., Villebro, F., Forchhammer, S., 2020. UAV image analysis for leakage detection in district heating systems using machine learning. Pattern Recognit. 140, 158–164. http://dx.doi.org/10.1016/j.patrec.2020.05.024.

Iakubovskii, P., 2019. Segmentation Models Pytorch. GitHub, (Accessed 20 August 2024), https://github.com/qubvel/segmentation_models.pytorch.

International Energy Agency (IEA), 2023. World Energy Outlook 2023. Technical Report, IEA, URL: https://www.iea.org/reports/world-energy-outlook-2023, Accessed 19 December 2024.

Islam, M.R., Ahmed, M.U., Barua, S., Begum, S., 2022. A systematic review of explainable artificial intelligence in terms of different application domains and tasks. Appl. Sci. 12 (3), 1353. http://dx.doi.org/10.3390/app12031353.

Jadon, S., 2020. A survey of loss functions for semantic segmentation. In: Proc. IEEE Conf. Comput. Intell. Bioinf. Comput. Biol.. CIBCB, IEEE, pp. 1–7. http://dx.doi.org/10.1109/cibcb48159.2020.9277638.

Johnson, J.M., Khoshgoftaar, T.M., 2019. Survey on deep learning with class imbalance. J. Big Data 6 (1), 27. http://dx.doi.org/10.1186/s40537-019-0192-5.

Li, C., Xia, W., Yan, Y., Luo, B., Tang, J., 2021. Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation. IEEE Trans. Neural Netw. Learn. Syst. 32 (7), 3069–3082. http://dx.doi.org/10.1109/TNNLS.2020.3009373.

Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B., 2021. Swin transformer: Hierarchical vision transformer using shifted windows. In: Proc. IEEE/CVF Int. Conf. Comput. Vis.. ICCV, IEEE, pp. 9992–10002. http://dx.doi.org/10.1109/iccv48922.2021.00986.

Ljungberg, S.-A., Rosengren, M., 1988. Aerial and mobile thermography to assess damages and energy losses from buildings and district heating networks - operational advantages and limitations. Int. Arch. Photogramm. Remote Sens. XXVII-B7, 348–359, https://www.isprs.org/proceedings/XXVII/congress/part7/348_XXVII-part7.pdf.

Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, IEEE, pp. 3431–3440. http://dx.doi.org/10.1109/cvpr.2015.7298965.

Loshchilov, I., Hutter, F., 2017. Decoupled weight decay regularization. http://dx.doi.org/10.48550/ARXIV.1711.05101.

Nogueira, K., Faita-Pinheiro, M.M., Marques Ramos, A.P., Gonçalves, W.N., Junior, J.M., Dos Santos, J.A., 2024. Prototypical contrastive network for imbalanced aerial image segmentation. In: 2024 IEEE/CVF Winter Conference on Applications of Computer Vision. WACV, IEEE, Waikoloa, HI, USA, pp. 8351–8361. http://dx.doi.org/10.1109/WACV57701.2024.00818.

NVIDIA Corporation, 2022. A100 tensor core GPU. URL: https://www.nvidia.com/en-us/data-center/a100/, Accessed 19 December 2024.

Pang, G., Shen, C., Cao, L., Hengel, A.V.D., 2022. Deep learning for anomaly detection: A review. ACM Comput. Surv. 54 (2), 1–38. http://dx.doi.org/10.1145/3439950.

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., Chintala, S., 2019. PyTorch: An imperative style, high-performance deep learning library. In: Adv. Neural Inf. Process. Syst.. Curran Associates, Inc., pp. 8024–8035. http://dx.doi.org/10.48550/arXiv.1912.01703.

Ronneberger, O., Fischer, P., Brox, T., 2015. U-Net: Convolutional networks for biomedical image segmentation. In: Lecture Notes in Computer Science. Springer International Publishing, pp. 234–241. http://dx.doi.org/10.1007/978-3-319-24574-4_28.

Ruck, J., Vollmer, E., Volk, R., Vogl, M., 2024. Detecting District Heating Leaks in Thermal Imagery: Comparison of Anomaly Detection Method - Source Code and Datasets. Zenodo, http://dx.doi.org/10.5281/zenodo.11085776.

Ruck, J., Vollmer, E., Volk, R., Vogl, M., 2025. Thermal Anomaly Segmentation Dataset – Thermal UAS-based Images from Germany with Annotations for Semantic Segmentation Model Training. Zenodo, http://dx.doi.org/10.5281/zenodo.14287864.

Safonova, A., Ghazaryan, G., Stiller, S., Main-Knorn, M., Nendel, C., Ryo, M., 2023. Ten deep learning techniques to address small data problems with remote sensing. Int. J. Appl. Earth Obs. Geoinf. 125, 103569. http://dx.doi.org/10.1016/j.jag.2023.103569.

Salehi, S.S.M., Erdogmus, D., 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. http://dx.doi.org/10.1007/978-3-319-67389-9_44.

Selvaraju, R.R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D., 2020. Grad-CAM: Visual explanations from deep networks via gradient-based localization. Int. J. Comput. Vis. 128 (2), 336–359. http://dx.doi.org/10.1007/s11263-019-01228-7.

Siddique, M.F., Ahmad, Z., and, J.-M.K., 2023. Pipeline leak diagnosis based on leak-augmented scalograms and deep learning. Eng. Appl. Comput. Fluid Mech. 17 (1), 2225577. http://dx.doi.org/10.1080/19942060.2023.2225577.

Sledz, A., Heipke, C., 2021. Thermal anomaly detection based on saliency analysis from multimodal imaging sources. ISPRS Ann. Photogramm. Remote. Sens. Spat. Inf. Sci. 55–64. http://dx.doi.org/10.5194/isprs-annals-V-1-2021-55-2021.

Sledz, A., Unger, J., Heipke, C., 2020. UAV-based thermal anomaly detection for distributed heating networks. ISPRS Int. Arch. Photogramm. Remote. Sens. Spat. Inf. Sci. XLIII-B1-2020, 499–505. http://dx.doi.org/10.5194/isprs-archives-XLIII-B1-2020-499-2020.

Sudre, C.H., Li, W., Vercauteren, T., Ourselin, S., 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. http://dx.doi.org/10.1007/978-3-319-67558-9_28.

United Nations Environment Programme, Global Alliance for Buildings and Construction, 2024. Global Status Report for Buildings and Construction - Beyond foundations: Mainstreaming sustainable solutions to cut emissions from the buildings sector. Technical Report, http://dx.doi.org/10.59117/20.500.11822/45095.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I., 2017. Attention is all you need. In: Adv. Neural Inf. Process. Syst.. NeurIPS, http://dx.doi.org/10.48550/ARXIV.1706.03762.

Vinogradova, K., Dibrov, A., Myers, G., 2020. Towards interpretable semantic segmentation via gradient-weighted class activation mapping. Proc. AAAI Conf. Artif. Intell. 34 (10), 13943–13944. http://dx.doi.org/10.1609/aaai.v34i10.7244, cs.

Vollmer, E., Benz, M., Kahn, J., Klug, L., Volk, R., Schultmann, F., Götz, M., 2025a. Enhancing UAS-based multispectral semantic segmentation through feature engineering. IEEE J. Sel. Top. Appl. Earth Obs. Remote. Sens. 18, 6206–6216. http://dx.doi.org/10.1109/JSTARS.2025.3537330.

Vollmer, E., König, S., Horstmann, V., Klug, L., Kahn, J., Volk, R., Vogl, M., 2025b. Thermal Urban Feature Segmentation - Multispectral (RGB + Thermal) UAS-based images from Germany with annotations. Zenodo, http://dx.doi.org/10.5281/zenodo.10814413.

Vollmer, E., Ruck, J., Volk, R., Schultmann, F., 2024. Detecting district heating leaks in thermal imagery: Comparison of anomaly detection methods. Autom. Constr. 168, 105709. http://dx.doi.org/10.1016/j.autcon.2024.105709.

Vollmer, E., Volk, R., Schultmann, F., 2023. Automatic analysis of UAS-based thermal images to detect leakages in district heating systems. Int. J. Remote Sens. 44 (23), 7263–7293. http://dx.doi.org/10.1080/01431161.2023.2242586.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M., 2020. Transformers: State-of-the-art natural language processing. In: Proc. 2020 Conf. Empir. Methods Nat. Lang. Process. Syst. Demonstr.. Association for Computational Linguistics, pp. 38–45. http://dx.doi.org/10.48550/arXiv.1910.03771.

Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. SegFormer: Simple and efficient design for semantic segmentation with transformers. In: Adv. Neural Inf. Process. Syst., vol. 34, Curran Associates, Inc., pp. 12077–12090. http://dx.doi.org/10.48550/arXiv.2105.15203.

Xu, Y., Wang, X., Zhong, Y., Zhang, L., 2016. Thermal anomaly detection based on saliency computation for district heating system. In: Proc. 2016 IEEE Int. Geosci. Remote Sens. Symp.. IGARSS, pp. 681–684. http://dx.doi.org/10.1109/IGARSS.2016.7729171.

Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit.. CVPR, IEEE, pp. 6230–6239. http://dx.doi.org/10.1109/cvpr.2017.660.

Zhong, Y., Xu, Y., Wang, X., Jia, T., Xia, G., Ma, A., Zhang, L., 2019. Pipeline leakage detection for district heating systems using multisource data in mid- and high-latitude regions. ISPRS J. Photogramm. Remote Sens. 151, 207–222. http://dx.doi.org/10.1016/j.isprsjprs.2019.02.021.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2019. Semantic understanding of scenes through the ADE20K dataset. Int. J. Comput. Vis. 127, 302–321. http://dx.doi.org/10.1007/s11263-018-1140-0.