


Enhancing UAS-Based Multispectral Semantic Segmentation Through Feature Engineering

Elena Vollmer , Mishal Benz , James Kahn , Leon Klug, Rebekka Volk , Frank Schultmann ,
and Markus Götz , *Member, IEEE*

Abstract—Deep learning (DL) is one of the key tools for analyzing images beyond the visible light spectrum, such as thermal data, for energy-related inspection and fault detection. However, publications using multispectral data focus on developing specialized models to handle quality issues without considering the imagery itself. This article investigates how feature engineering (FE), the process of adapting raw data to serve as DL training data, can impact performance when transferring prevalent model architectures to combined red, green, blue (RGB) thermal imagery. The popular U-Net is utilized for the common task of multiclass semantic segmentation in remote sensing. A comprehensive ablation study is performed on a novel, uncrewed aircraft system-based dataset from two German cities to detect thermal urban features. Common performance metrics, training, and energy consumption statistics are compared to find the most suitable combination of platform-specific and general enhancing FE while identifying the impact of resolution, channel count, RGB, and color information. The study reveals FE to significantly influence predictive performance, where the choice of ablation parameters are found to have a 7% pt–10% pt impact. Computational resource utilization depends on image size, following a logarithmic growth curve. Importantly, the study demonstrates that in-depth FE of thermal imagery can replace the need for additional RGB data.

Index Terms—Deep learning (DE), feature engineering (FE), multispectral images, remote sensing, semantic segmentation, thermography.

I. INTRODUCTION

WHILE political and social efforts strive to limit anthropogenic global warming, the building sector remains one of the greatest contributors to climate change [1]. It is responsible for approximately 30% of the global energy demand, primarily due to operational requirements such as heating [1]. The German government has, therefore, recently passed a law to

establish a concept for country-wide heat supply, whereby municipalities are charged with creating comprehensive plans for climate-neutral heating in their communes [2]. Similar legislation already in effect in Scandinavian countries such as Denmark has led to the implementation of centralized technologies—specifically district heating systems (DHSs)—which currently provide two thirds of the population with 89% climate-neutral heat [3]. However, the efficient operation of such systems still presents a significant challenge. In Germany, for example, network losses were estimated at over 10% in 2022 [3]. Keeping in mind current and future heating-related aims, a key part of enabling sustainable cities must lie in ensuring a high level of efficiency and minimal thermal losses in all involved infrastructure.

A versatile and holistic monitoring approach can be achieved by combining modern technologies: uncrewed aircraft system with multispectral sensors gather image data to be efficiently analyzed via computer vision methods such as artificial intelligence (AI) [4]. Where images are concerned, deep learning (DL) has proven to be extremely effective for classification, detection, and segmentation tasks in numerous domains. The past decade has seen an increase in the application of DL to imagery beyond the visible light—380–700 nm wavelength—spectrum [5], [6], [7]. Thermal infrared (TIR) data, for instance, captures emitted electromagnetic waves corresponding to surrounding temperatures, and thus, can provide context information on heat sources [6], [8]. With regard to cities, its use ranges from autonomous driving over crowd counting to the maintenance of energy-related systems, including defect detection in DHS, solar technologies, or transmission lines, and building inspections [6], [8], [9], [10]. These problems are increasingly being addressed with complex, high-level semantic segmentation models as they enable pixelwise classification and more nuanced contextual perception [6]. Merging standard red, green, blue (RGB) with TIR imagery is said to enrich scene understanding, help supply missing information under complex illumination, and greatly increase segmentation accuracy [6], [7], [10].

However, transferring DL to novel data types comes with a set of challenges. While images are generally subject to noise and acquisition-dependent artifacts [11], TIRs are particularly susceptible due to the involved sensor technology. They suffer from substantially lower resolution and undesirable effects—most prominently blurring and nonuniformity [5], [12], [13]. Although data quality is known to have a great impact on DL performance [11], studies focus on model adaptation instead of

Received 6 November 2024; revised 29 December 2024; accepted 26 January 2025. Date of publication 30 January 2025; date of current version 27 February 2025. This work was supported by the European Union through the AI4EOSC project (Horizon Europe) under Grant 101058593. (*Corresponding author: Elena Vollmer.*)

Elena Vollmer, Rebekka Volk, and Frank Schultmann are with the Institute of Industrial Production, Karlsruhe Institute of Technology, 76187 Karlsruhe, Germany (e-mail: elena.vollmer@kit.edu).

Mishal Benz and Markus Götz are with the Helmholtz AI, Scientific Computing Centre, Karlsruhe Institute of Technology, 76344 Karlsruhe, Germany.

James Kahn was with Helmholtz AI, Scientific Computing Centre, Karlsruhe Institute of Technology, 76344 Karlsruhe, Germany. He is now with HAL Systems, Melbourne 3079, Australia.

Leon Klug was with the Institute of Industrial Production, Karlsruhe Institute of Technology, 76187 Karlsruhe, Germany. He is now with Product management data science and AI, Check24 Autoteile GmbH, 80339 Munich, Germany.

Digital Object Identifier 10.1109/JSTARS.2025.3537330

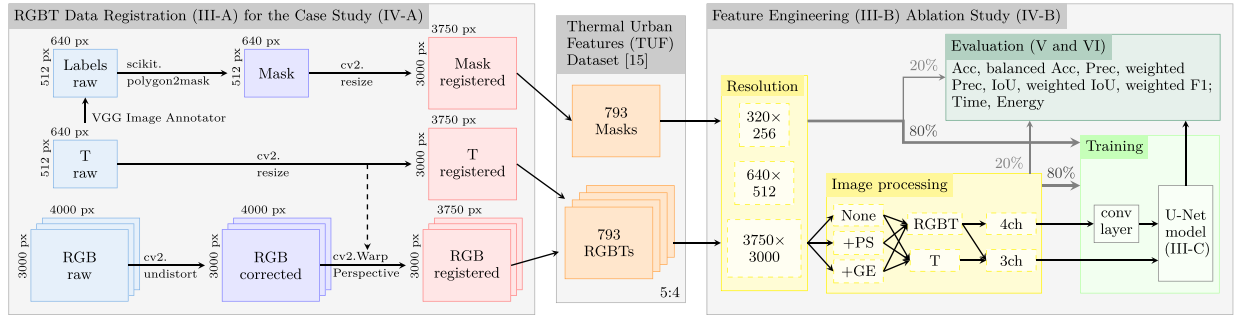


Fig. 1. Overview of the study and developed data processing pipelines: From RGB and TIR image registration for RGBT dataset creation to exhaustive FE ablation study. The numbers in brackets refer to manuscript sections containing further details on the step in question.

manual feature engineering (FE)—the process of adapting raw data for its use in DL model training—to tackle the issue.

This article, therefore, presents a thorough investigation on the effects of FE on remote sensing (RS) red, green, blue, thermal (RGBT)-based DL. The challenging and increasingly popular computer vision task of semantic segmentation is examined within the framed context of city infrastructure monitoring—specifically for identifying thermal, meaning heat-related, urban features, and anomalies. Our contributions are as follows.

- 1) We identify key factors impacting quality in uncrewed aircraft system (UAS)-based red, green, blue, thermals (RGBTs) as either platform-specific (notably vignetting) or general (specifically contrast and blurring) and find suitable algorithms for mitigation.
- 2) For increased impact, we adapt a prevalent DL model to a new, real-world case study in heat-related inspection, thus creating a novel RGBT dataset.
- 3) An extensive ablation study is conducted to analyze numerous data-related aspects—including filters, channel constellations, and image sizes—with particular emphasis on the impact of information loss. Our comprehensive evaluation compares not only a broad range of performance metrics, but also resource statistics to assess our AI in terms of sustainability [14].
- 4) We provide best-practice conclusions for multispectral remote sensing (RS) image acquisition and analysis for future sustainable cities.

In addition, following open science principles, our novel RGBT dataset [15] and code¹ are published alongside this article, thus ensuring reproducibility.

Fig. 1 visualizes the developed data processing pipelines, including registration to create an RGBT dataset forming the basis of the FE ablation study.

The rest of this article is organized as follows. After covering related work in Section II, all elements in Fig. 1 are introduced in Section III and detailed in Section IV. Results are presented in Section V and discussed in Section VI, and finally, Section VII concludes this article.

II. RELATED WORK

A review in [16] of semantic segmentation in RS found that 89% of papers compare different models, while 93%

perform architecture-based ablation studies. Similarly in multispectral data, the authors in [6], [7], and [10] list various models developed for RGBT data—such as the Penn Subterranean Thermal Network [17]—but omit FE.

Many studies, especially in RS, disregard FE owing to their use of benchmark datasets [16]. Publications that inspect data cleaning in RGB imagery find that it improves DL performance. Undesirable illumination can be mitigated with contrast-enhancing algorithms such as contrast limited adaptive histogram equalization (CLAHE) to increase segmentation accuracy [18]. In satellite imagery, atmospheric artifacts can be reduced by increasing contrast, i.e., through unsharp mask and median filtering to enhance classification accuracy and decrease computational complexity [19].

Fewer studies look into FE in TIR-based DL models. In spite of the technology’s numerous uses, He et al. [5] find that current implementations are more akin to laboratory than industrial applications. This may allow for less artifact-ridden data and explain the model-driven focus. TIR quality improvement is only described as increasing resolution by using specifically developed DL architectures [5]. While Chaverot et al. [20] observe that DL commonly replaces image processing, they find TIR quality enhancement, i.e., via deblurring to significantly improve object detection. Data cleaning of TIRs for DL is described by Herrmann et al. [21], who apply various filters to mimic RGB appearance, and find that these improve the performance of RGB-pretrained DL.

Although FE improves RGB- and TIR-based DL performance, there exists—to the knowledge of the authors—no study exploring the effects on combined RGBT-based semantic segmentation. This article, therefore, utilizes the well-known U-Net model for an ablation study focused on the central task of multiclass thermal urban feature segmentation. By identifying common and anomalous heat sources in urban environments, DL can help detect DHS leakages conservatively by removing false alarms [9].

III. METHODS

A. RGBT Data Registration

RGBT imagery is made up of four channels: three color and one temperature-dependent one. Recording these data requires both a visible light and infrared sensor. The most common and cheapest thermal sensors are uncooled microbolometers, which

¹[Online]. Available: <https://github.com/emvollmer/TUFSeg>

capture long-wavelength infrared (LWIR) radiation emitted by all objects of common temperature on Earth (-83°C to 727°C). The resulting outputs are in grayscale, with lighter pixels denoting higher temperatures [5], [8], [12].

Typically, two separate sensors are used for acquisition [7], placed side by side to match the fields of view [17]. Hybrid cameras can facilitate the process by incorporating both technologies [7], [22]. RS acquisition, especially by UAS, is simplified through merged dual camera and gimbal systems [23]. In all cases, the raw images require alignment to compensate differing fields of view, resolution, and aspect ratios [22], [24]. TIRs (around 640×512 or less [5]) generally have a considerably lower resolution than RGBs (around 1920×1080 [7] or up to 4000×3000 [24]).

Aligning RGBs and TIRs typically consists of two steps: distortion correction and image registration [17], [23]. Depending on the used camera, images typically suffer from two types of distortions: radial (barrel or pincushion effects) or tangential (lack of lens alignment with the image plane) [25]. To correct these, a camera's intrinsic and extrinsic parameters need to be estimated. Most simply, this is achieved by recording reference images of a standardized pattern, such as checkerboard of black and white squares, with which calibration functions from computer vision libraries such as OpenCV [26] can estimate a camera's intrinsic matrix and distortion coefficients. When applied, these will correct sensor-specific distortion in the given data [17], [23], [27], [28].

After distortion correction, the two sensor outputs are aligned on pixel-level to compensate any offset between camera views (existing even for dual cameras). A simple registration approach estimates the homography matrix that describes the cameras' relationship with coordinate pairs from matching key points. Both TIR and RGB are manually sampled to identify a list of corresponding pixel locations for matrix estimation. Applying this transformation warps the images to the same resolution [23], [28].

B. Feature Engineering (FE)

Raw imagery can be affected by noise, lack of contrast, problematic lighting conditions, blur, and other artifacts. How severely these are expressed depends on the utilized technology and acquisition conditions. For instance, capturing images by popular UAS [5], [7] elicits pronounced nonuniformity in TIRs [12]. Nighttime recordings are most useful for TIR information due to low thermal reflectance, but RGBs suffer from reduced luminance.

Observations from previous studies and existing conditions help focus on two key aspects: platform-specific and general image enhancement FE. Platform-specific speaks to the compensation of effects induced by the utilized RS acquisition method—in this case UAS. General enhancement (GE) mitigates quality issues unrelated to the form of acquisition and instead pertaining to the sensors themselves. These are centered around contrast and blurring.

1) *Platform-Specific FE*: UAS image acquisition has several advantages, including high flexibility, easy operability, and high spatial resolution through low flight heights [29]. Uncrewed

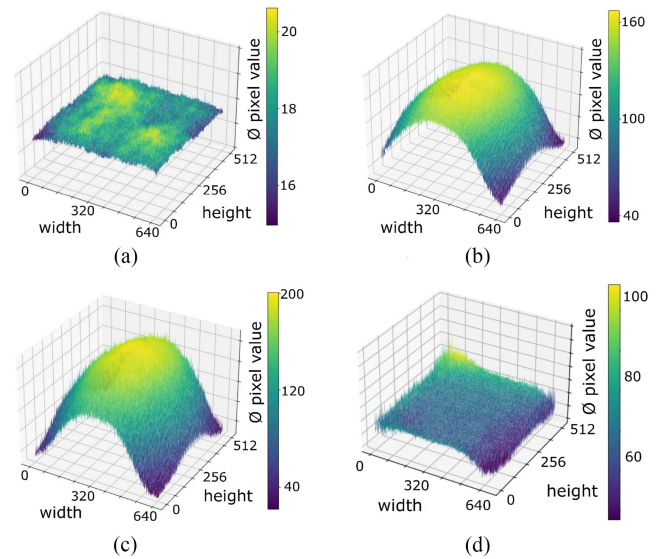


Fig. 2. Location-dependent pixel distributions, colored according to pixel values from blue (low) to yellow (high). (a) Averages over all RGBs. (b) Averages over all TIRs. (c) Averages over one select TIR dataset. (d) Corrected averages over one select TIR dataset.

aircrafts (UAs) are utilized with dual RGB and TIR sensors for crop [29], building [24], power line [30] monitoring, and many more. However, the prevalent conditions during flights can particularly affect the thermal camera, as its sensor, lens, and housing temperatures are affected by heat from the gimbal motor and coolness from propellers' slipstream [12]. The resulting nonuniformity manifests as a cooling of image corners and edges [12]. Although microbolometers have an integrated correction to compensate for fixed pattern noise, they cannot adequately offset this so-called “vignetting” or “halo” effect [12].

Fig. 2 visualizes the average pixel values of all thermal versus visual images. Ideally, the averages are close to equal for a uniform distribution, as is true for RGBs [see Fig. 2(a)]. The TIR channel, however, displays an extreme radial deviation [see Fig. 2(b)]. This effect manifests differently depending on individual flight conditions [see Fig. 2(c)].

Various vignetting correction (VC) approaches exist, but they do not always negate the entire effect and often require a reference image [12], [31]. Therefore, a method utilizing radial polynomial functions as described in [32] is implemented. Pixels are grouped based on radial distance to the image center and bin averages used to model a vignetting function. The function is approximated for each image so that magnitude variations between images or datasets do not pose a problem. Here, 100 bins and a tenth degree polynomial function are found to be optimal. Fig. 3 exemplifies how the algorithm corrects substantial vignetting in a TIR from Fig. 2(c) dataset. Fig. 2(d) shows the corrected pixel distribution for that specific dataset.

2) *General Enhancement (GE)*: Image enhancement refers to the improvement of quality, specifically contrast and blur, regardless of the acquisition platform.

a) *Contrast enhancement*: Undesirable lighting conditions can greatly degrade image quality. Phenomena such as

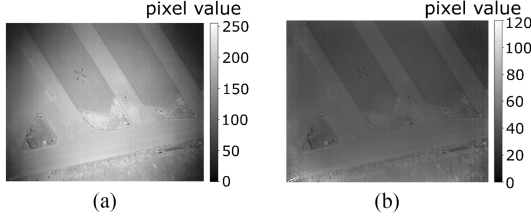


Fig. 3. Visualization of the VC algorithm. (a) TIR image. (b) VC TIR.

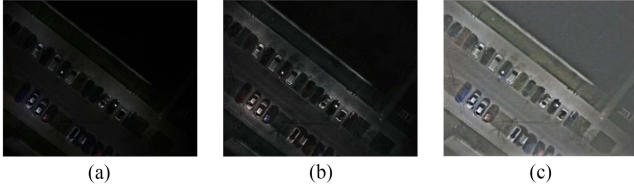


Fig. 4. Visualization of the contrast-enhancing algorithms CLAHE [see Fig. 4(b)] and Retinex [see Fig. 4(c)]. (a) RGB image. (b) CLAHE RGB. (c) Retinex RGB.

brightness, shadows, over-, or underexposure induce noise and obscure object features, especially outdoors. Contrast enhancement restores images to compensate such effects and increase quality. Illumination is particularly problematic in RGBs when images are acquired at night [18].

Several algorithms can perform this task, most prominently the histogram-based CLAHE [18] and Retinex [33] methods. Both are designed to compensate for spatially varying nonuniformity resulting from varying brightness (i.e., enhance local contrast), but can also suppress noise in RGBs and TIRs [13], [18], [34], [35].

Histogram equalization enhances image contrast by stretching existing pixel values across all possible values. Although effective, this technique can introduce noise and reduce information content in images with high contrast. An improved variant is CLAHE, a method that first divides the image into small areas before applying histogram equalization to limit excessive contrast and increase noise robustness in low-contrast areas [18], [36].

Retinex imitates the behavior of photoreceptors and ganglion cells in the human eye (retina) and processing structures of the mind (cortex). Images are interpreted as a multiplication of illumination (overall scene lighting) and reflectance (intrinsic scene properties). By separating the two components and applying local logarithmic luminance compression and spectral brightening, varied illumination can be compensated and quality improved [33], [34].

While Retinex can be applied directly to three-channel data, CLAHE only works with single channels. It is common practice to convert RGBs into $L \times A \times B$ color space (where L represents luminance and A and B color distributions) and enhance the L channel [37]. Fig. 4 compares both algorithms. Retinex is capable of extracting significantly more information from darker image regions while preserving the structure of illuminated ones. It is, therefore, implemented for contrast increase here.

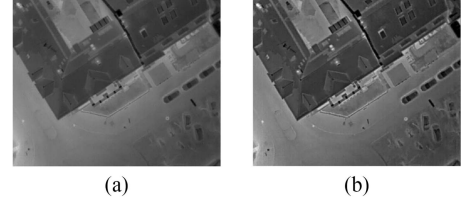


Fig. 5. Visualization of the UM algorithm. (a) TIR image. (b) UM TIR.

As the VC from Section III-B-1 can noticeably darken TIRs, the algorithm is also applied to increase their luminance.

b) Deblurring: TIRs generally suffer from blurring, which complicates the detection of smaller or farther objects. This can be mitigated with the unsharp masking (UM) algorithm, as visualized in Fig. 5. The technique involves blurring an image I with a noise-reducing filter (typically Gaussian) G and concatenating the output with I according to (1). λ is a positive integer defining the effect's strength with a default of 1, used here to prevent additional, unwanted noise artifacts [20], [38]

$$I_{\text{corr}} = I + \lambda(I - G(I)). \quad (1)$$

UM is also commonly applied to RGBs. In satellite imagery, it has been shown to increase DL classification accuracy by reducing artifacts and enhancing edges [19], and therefore, is used on our remotely sensed RGBs as well.

C. Model Architecture

Despite the ongoing development of specialized models (see Section II), the U-Net architecture [39] is still the most prevalent in RS [40]. The model enhances a classical fully convolutional network by including skip connections between encoder and decoder [39]. This improves segmentation when working with limited training data, as is common for real-world implementations [39]. The model excels in various fields, is among the most popular for urban feature segmentation [16], [41], [42], and known to proficiently analyze multispectral satellite data [43]. Its continued relevance makes the U-Net a good choice, as it allows for more generalizable, and thus, significant results.

Architecture-related details are based on survey of urban feature segmentation in RS images in [16]. The most common backbone (also for RGBT data [7]) is the ResNet, which is why the ResNet-152 is selected as an encoder [16]. Chiefly, cross entropy (CE) functions are used to determine loss [16]. An adapted variant—focal CE—targets difficult-to-learn instances and helps manage the common issue of class imbalance [44], [45]. As this works well for training a U-Net on multispectral satellite data [44], the sigmoid focal CE function [45] is implemented here.

Transfer learning can balance limited datasets [46]. Initializing a model with weights from other trainings helps compensate for a lack of data with knowledge from a related task [46]. While public weights are based on RGBs, ablation studies such as [24] have shown DL performance to significantly improve when used with RGBT data. Therefore, the U-Net encoder backbone is loaded with weights trained on the popular ImageNet

dataset [47], as done in [46]. Because of this, the model only accepts inputs with the same channel count. Two options exist to adapt to RGBT images: concatenate the data into three channels or map the fourth channel to the given three with a preceding convolutional layer. Both are compared here.

IV. EXPERIMENT

The overall data handling procedure developed for this case study experiment is visualized in Fig. 1. The following subsections address relevant aspects of both pipelines.

A. Case Study Description and Data Registration

The case study comprises images from various suburban regions around Munich (Germany), captured from 8 PM to 6 AM in December 2019 with temperatures ranging from -5°C to 2°C . Additional data from urban areas in a second German city, Karlsruhe, help diversify the study. These recordings were acquired in January and March 2022 with temperatures of 0°C – 3°C .

Acquisition took place via Matrice 600 Pro [48] and 300 RTK [49] UA. The flight was almost entirely automated to follow a lawnmower pattern at 60-m altitude in nadir. RGBT imagery was captured simultaneously with DJI's Zenmuse XT2 [50], a combined gimbal and camera incorporating a 4k RGB and thermal sensor by FLIR. The RGBs have a size of 4000×3000 pixels and the TIRs of 640×512 .

Of 8452 combined images, 793 are selected for annotation—700 from Munich and 93 from Karlsruhe. Owing to an 88% overlap, only every ninth image depicts an entirely new scene and is, therefore, worth considering for annotation. Of these, trained experts select the most suitable by avoiding duplicate areas and motion blur due to turns during UAS flight. Within the TIRs, we identify nine classes of common thermal urban features and label a total of 8010 polygons using Visual Geometry Group's VGG Image Annotator [51]. The annotation masks required for semantic segmentation are generated from these polygons by assigning each pixel a number per its defined class and all unlabeled pixels to the background. The class distribution in Appendix A highlights an annotation and pixel imbalance common to these types of tasks.

The raw images are processed according to Section III-A, as shown in the left part of Fig. 1. This case study's RGBs suffer from a distinct fisheye distortion, which is corrected using 129 key points before image registration with Python's OpenCV [26] and scikit [52]. A result of annotating the TIRs is an aspect ratio of 5 : 4. The data are scaled to 3000×3750 to match RGB resolution.

B. Ablation Study

Table I shows how the ablation study investigates manual FE for DL. Aside from image filters, we examine the effect of information loss—specifically color when reducing channel count and content when varying resolution—to identify attributes influencing model performance.

Based on Section III-B, three key FE options are defined: none, platform-specific (PS) processing (meaning vignetting

TABLE I
OVERVIEW OF PARAMETER COMBINATIONS AND CHANNEL INPUT DEFINITIONS

Parameters			Model channel inputs			
proc	ch	data	ch1	ch2	ch3	ch4
none	3	T	T	T	T	-
+PS	3	T	vc T	vc T	vc T	-
+GE	3	T	vc T	ret T	um T	-
none	3	RGBT	gray RGB	T	T	-
+PS	3	RGBT	gray RGB	vc T	vc T	-
+GE	3	RGBT	ret RGB	ret T	um T	-
none	4	RGBT	R	G	B	T
+PS	4	RGBT	R	G	B	vc T
+GE	4	RGBT	ret RGB	ret T	um RGB	um T

removal), and GE (meaning additional Retinex and UM algorithms). These are applied in parallel (individually per channel) as this has been shown to yield better results in TIR-based DL than consecutive preprocessing [21].

As discussed in Section III-C, both the concatenation into three-channel and model adaptation to four-channel inputs are tested. RGBs are converted into single-channel grayscales, meaning the relevance of color can be investigated. With a thermography-centered objective, we can additionally investigate the sole use of TIR inputs as a baseline. This alleviates higher processing costs of combining RGBs and TIRs and potential registration discrepancies [6].

In all Table I configurations, the input data are scaled to high (3750×3000), mid (640×512), and low (320×256) resolution, as larger files require more significant hardware capacities. We can thereby investigate what impact image size has on RGBT or TIR model performance and whether an inverted U-shaped relationship exists here. To allow for a simple, statistical analysis [24], each configuration is trained with four seeds (see Appendix B for more details).

The resulting 108 models are evaluated using a wide range of semantic segmentation metrics: overall accuracy, balanced accuracy, precision, weighted precision, weighted F1-score, mean intersection over union (IoU), and weighted mean IoU. These are chosen based on the most common metrics in RS urban feature detection [16] and RGBT segmentation [7]. The balanced and weighted variants consider class imbalance by calculating classwise scores and determining (weighted) averages. These metrics, alongside their standard equivalents, are particularly helpful for a comprehensive estimation of model performance here.

Measured resource metrics include time used for FE and model training as well as energy consumption in accordance with sustainable AI principles [14]. The energy used is calculated with Perun [53], which outputs both kW·h and kgCO₂eq.

V. RESULTS

Table II summarizes the ablation study results, divided into performance- and resource-related metrics. To compensate fluctuations due to seed initialization, these are presented as “mean \pm standard deviation (SD),” calculated across the four selected seeds (see Appendix B).

TABLE II
ABLATION STUDY RESULTS

ch	data	proc	res	Performance Metrics							Resource Metrics			
				Accuracy	balanced Acc	Precision	weighted P	IoU	weighted IoU	weighted F1	t_{data} [min]	t_{train} [min]	energy [kWh]	energy [kgCO ₂ e]
3	T	none	high	0.915 ± 0.067	0.556 ± 0.043	0.570 ± 0.033	0.946 ± 0.017	0.453 ± 0.056	0.870 ± 0.081	0.918 ± 0.054	1.310 ± 0.088	15.12 ± 1.154	0.180 ± 0.016	0.076 ± 0.007
			mid	0.892 ± 0.137	0.566 ± 0.041	0.562 ± 0.097	0.948 ± 0.032	0.449 ± 0.108	0.848 ± 0.163	0.899 ± 0.117	0.458 ± 0.299	14.46 ± 0.084	0.167 ± 0.004	0.070 ± 0.002
			low	0.904 ± 0.095	0.514 ± 0.079	0.542 ± 0.062	0.943 ± 0.031	0.417 ± 0.086	0.860 ± 0.116	0.909 ± 0.084	0.062 ± 0.045	4.720 ± 0.020	0.051 ± 0.000	0.021 ± 0.000
		+PS	high	0.933 ± 0.042	0.561 ± 0.025	0.574 ± 0.045	0.954 ± 0.012	0.460 ± 0.043	0.895 ± 0.050	0.936 ± 0.034	2.982 ± 0.491	15.13 ± 0.895	0.192 ± 0.007	0.080 ± 0.003
			mid	0.945 ± 0.013	0.507 ± 0.063	0.573 ± 0.052	0.950 ± 0.013	0.435 ± 0.053	0.904 ± 0.023	0.941 ± 0.017	2.423 ± 0.288	14.50 ± 0.104	0.180 ± 0.003	0.076 ± 0.001
			low	0.951 ± 0.004	0.531 ± 0.024	0.572 ± 0.006	0.956 ± 0.003	0.455 ± 0.015	0.916 ± 0.006	0.950 ± 0.004	0.418 ± 0.032	4.715 ± 0.017	0.053 ± 0.001	0.022 ± 0.001
	RGBT	+GE	high	0.949 ± 0.009	0.538 ± 0.037	0.573 ± 0.056	0.954 ± 0.006	0.457 ± 0.034	0.910 ± 0.015	0.944 ± 0.011	5.208 ± 0.456	14.54 ± 0.121	0.202 ± 0.011	0.084 ± 0.004
			mid	0.954 ± 0.001	0.575 ± 0.004	0.584 ± 0.019	0.960 ± 0.002	0.477 ± 0.010	0.921 ± 0.002	0.953 ± 0.002	4.792 ± 0.270	14.37 ± 0.120	0.193 ± 0.003	0.080 ± 0.001
			low	0.946 ± 0.017	0.542 ± 0.042	0.531 ± 0.045	0.955 ± 0.008	0.435 ± 0.035	0.909 ± 0.022	0.945 ± 0.014	1.245 ± 0.048	4.718 ± 0.039	0.059 ± 0.001	0.024 ± 0.001
		none	high	0.948 ± 0.013	0.585 ± 0.031	0.580 ± 0.036	0.955 ± 0.008	0.478 ± 0.029	0.911 ± 0.018	0.947 ± 0.011	1.632 ± 0.555	14.54 ± 0.147	0.172 ± 0.005	0.072 ± 0.002
			mid	0.949 ± 0.004	0.563 ± 0.031	0.580 ± 0.025	0.956 ± 0.005	0.464 ± 0.025	0.913 ± 0.009	0.948 ± 0.007	0.510 ± 0.262	14.50 ± 0.057	0.168 ± 0.003	0.070 ± 0.001
			low	0.943 ± 0.011	0.537 ± 0.032	0.565 ± 0.021	0.952 ± 0.007	0.444 ± 0.034	0.904 ± 0.017	0.941 ± 0.011	0.062 ± 0.019	4.690 ± 0.043	0.051 ± 0.001	0.022 ± 0.001
4	T	+PS	high	0.947 ± 0.007	0.522 ± 0.040	0.545 ± 0.051	0.955 ± 0.005	0.434 ± 0.041	0.911 ± 0.010	0.946 ± 0.006	2.940 ± 0.423	15.05 ± 1.193	0.186 ± 0.019	0.078 ± 0.008
			mid	0.916 ± 0.073	0.515 ± 0.076	0.526 ± 0.111	0.924 ± 0.068	0.407 ± 0.110	0.861 ± 0.113	0.902 ± 0.096	2.420 ± 0.313	14.40 ± 0.079	0.178 ± 0.005	0.074 ± 0.002
			low	0.948 ± 0.010	0.509 ± 0.041	0.552 ± 0.016	0.952 ± 0.008	0.434 ± 0.020	0.909 ± 0.016	0.945 ± 0.010	0.390 ± 0.000	4.748 ± 0.096	0.054 ± 0.002	0.022 ± 0.001
		+GE	high	0.878 ± 0.158	0.516 ± 0.094	0.546 ± 0.075	0.944 ± 0.035	0.417 ± 0.109	0.834 ± 0.183	0.888 ± 0.135	5.300 ± 0.402	14.42 ± 0.295	0.200 ± 0.013	0.084 ± 0.005
			mid	0.951 ± 0.008	0.501 ± 0.017	0.543 ± 0.049	0.955 ± 0.006	0.430 ± 0.026	0.915 ± 0.012	0.948 ± 0.009	4.730 ± 0.306	14.55 ± 0.034	0.196 ± 0.005	0.082 ± 0.002
			low	0.950 ± 0.006	0.505 ± 0.036	0.527 ± 0.031	0.955 ± 0.004	0.429 ± 0.018	0.914 ± 0.008	0.948 ± 0.006	1.235 ± 0.031	4.775 ± 0.147	0.060 ± 0.002	0.025 ± 0.001
	RGBT	none	high	0.937 ± 0.030	0.574 ± 0.035	0.563 ± 0.065	0.949 ± 0.019	0.458 ± 0.066	0.896 ± 0.043	0.937 ± 0.027	1.955 ± 0.590	14.81 ± 0.356	0.180 ± 0.008	0.075 ± 0.003
			mid	0.931 ± 0.042	0.555 ± 0.073	0.563 ± 0.062	0.953 ± 0.015	0.458 ± 0.071	0.892 ± 0.055	0.934 ± 0.036	0.600 ± 0.232	15.31 ± 1.361	0.179 ± 0.026	0.075 ± 0.011
			low	0.953 ± 0.005	0.498 ± 0.022	0.557 ± 0.022	0.956 ± 0.004	0.439 ± 0.018	0.918 ± 0.009	0.950 ± 0.006	0.335 ± 0.359	4.930 ± 0.315	0.057 ± 0.010	0.024 ± 0.004
		+PS	high	0.956 ± 0.004	0.534 ± 0.018	0.576 ± 0.006	0.960 ± 0.003	0.452 ± 0.010	0.923 ± 0.006	0.954 ± 0.003	3.192 ± 0.599	14.77 ± 0.316	0.187 ± 0.007	0.078 ± 0.003
			mid	0.915 ± 0.054	0.513 ± 0.048	0.525 ± 0.059	0.939 ± 0.020	0.402 ± 0.062	0.868 ± 0.067	0.917 ± 0.044	2.462 ± 0.318	14.56 ± 0.122	0.180 ± 0.004	0.075 ± 0.002
			low	0.919 ± 0.070	0.482 ± 0.047	0.527 ± 0.081	0.943 ± 0.028	0.395 ± 0.084	0.878 ± 0.082	0.921 ± 0.060	0.432 ± 0.078	4.762 ± 0.046	0.054 ± 0.001	0.022 ± 0.001
	RGBT	+GE	high	0.936 ± 0.036	0.558 ± 0.025	0.556 ± 0.047	0.951 ± 0.017	0.449 ± 0.051	0.897 ± 0.048	0.938 ± 0.031	5.580 ± 0.632	15.06 ± 0.988	0.209 ± 0.009	0.087 ± 0.004
			mid	0.952 ± 0.005	0.563 ± 0.047	0.551 ± 0.026	0.958 ± 0.002	0.460 ± 0.021	0.919 ± 0.006	0.951 ± 0.004	5.020 ± 0.256	14.57 ± 0.074	0.199 ± 0.006	0.083 ± 0.003
			low	0.946 ± 0.010	0.504 ± 0.027	0.539 ± 0.046	0.952 ± 0.006	0.425 ± 0.030	0.908 ± 0.014	0.944 ± 0.009	1.268 ± 0.028	4.765 ± 0.026	0.060 ± 0.001	0.025 ± 0.000
		none	high	0.878	0.482	0.525	0.924	0.395	0.834	0.888	0.063	4.690	0.051	0.021
			mid	0.956	0.585	0.584	0.960	0.478	0.923	0.954	5.580	15.31	0.209	0.087
			low	0.078	0.102	0.059	0.036	0.083	0.089	0.065	5.518	10.62	0.158	0.066

Legend: ch = channel, proc = processing, res = resolution

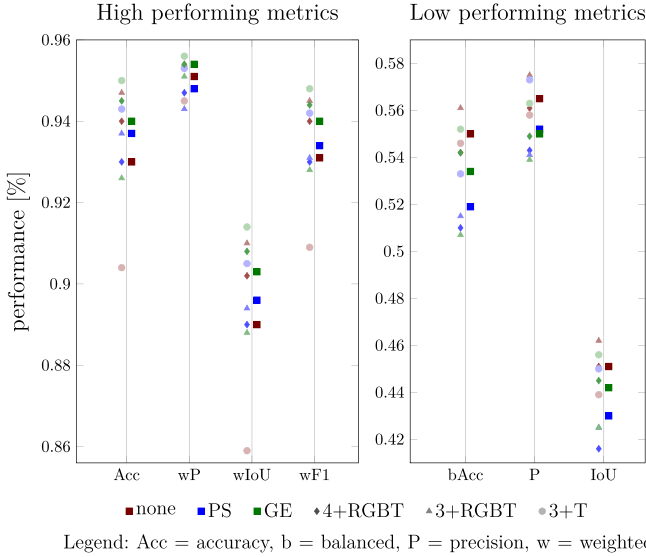


Fig. 6. Performance metrics for filter and channel input combinations, averaged over all image sizes.

VI. DISCUSSION

A. Performance

1) *Quantitative Evaluation*: A characteristic trend of overall high- and low-scoring metrics confirms the presence of a strong class imbalance. A generally high mean accuracy (88%–96%) and low balanced accuracy (48%–59%) signifies that dominant classes are predicted much more accurately than underrepresented ones. Like balanced accuracy, precision considers each class equally and lies in a comparable range. With a weighting factor defined by the number of true class instances, weighted precision balances class size discrepancies and scores very high (92%–96%). The averaged mean IoU scores are lowest

(39%–48%), while the class-weighted variant reaches 83%–92%. Even higher scores are obtained by the weighted F1 metric. In total, the choice of parameter values accounts for a 7%pt–10%pt difference across the various performance metrics.

SDs are only 3.7% on average, but a 2.74% median and 10%–18% peak values signify outliers—most prominently the three-channel TIR midresolution none, RGBT midsize PS, and high-resolution GE. These stem from individual seeds curtailing performance—here numbers 1000, 1 234 567, and 1000, respectively. While subsequent analyzes use mean values, these outliers can influence general conclusions.

The simplest data processing without filters using three-channel RGBT high-resolution images scores highest for the lower metrics, specifically balanced accuracy, mean IoU, and second for precision. The higher performing metrics (accuracy, weighted precision, weighted IoU, and weighted F1) almost all peak for PS FE with the highest resolution four-channel inputs. Despite this, the overall best results are found for GE three-channel midsize TIR data, which scores best in both precision-related metrics and second in all others. This is an interesting observation, as it indicates that supplementary RGBs are not necessarily as necessary when the TIRs are feature engineered. Table II reveals a clear pattern for near to all metrics regarding the sole use of thermal imagery: PS and GE improve upon performance (up to 5%pt averaged). To allow further study, filters and channels are analyzed separately from image sizes.

Fig. 6 compares averages over all image sizes. Overall, raw data yield the best results for lower metrics, while GE (followed by PS) scores highest in the higher metrics. As the latter measure representative performance despite class imbalance, this indicates that FE helps overall performance, but does not aid with underrepresented classes.

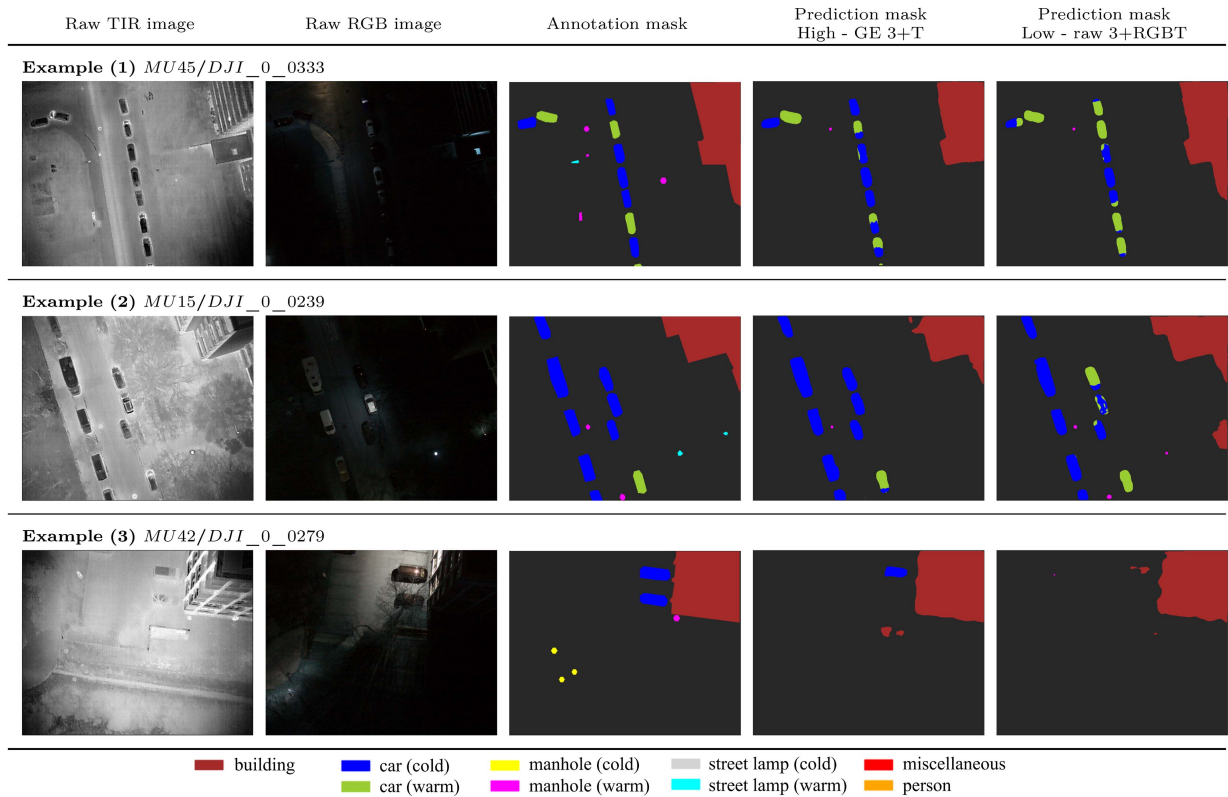


Fig. 7. Qualitative comparison of high (3+T) versus low (3+RGBT) performing metrics winners, exemplified via the model variant of resolution 640×512 and seed 42.

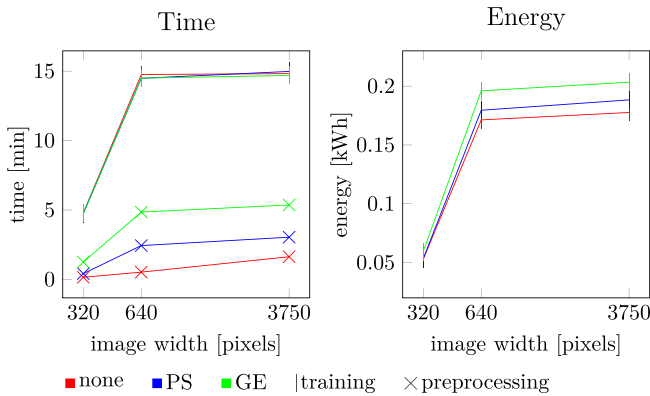


Fig. 8. Resource resolution relationship.

Preferable channel combinations also vary. For the raw data, 3+RGBT inputs are best, closely followed by 4+RGBT. Using only TIRs yields conspicuously low results (up to 5%pt decrease), indicating that RGBs should be included when utilizing raw data. Contrarily, the closeness of 3+ and 4+RGBT scores shows that the loss of color caused by the greyscale transformation has little impact.

For PS and GE, the best results are achieved with 3+T inputs, meaning the sole use of TIRs is beneficial. This could imply that RGB FE is less helpful than initially hypothesized, but individual

score comparisons contradict this. The mentioned outliers may be causing the discrepancy. Compared to raw 3+T inputs, both PS processing and added RGB information individually improve results, yet their combination in PS 3+RGBT performs worse than either. This may indicate an issue with data registration.

An evaluation of how image size influences performance finds the highest resolution (3750×3000) to provide the best results most often, followed closely by mid-sized images. Only in few instances does the smallest resolution yield the best models. Interestingly, for GE, mid-resolution always produces the highest scores. Analyzing averages for each image size shows that highest resolution images are particularly helpful in increasing performance of the low-performing metrics by up to 4%pt. For higher scoring metrics, total averages vary only slightly (around or less than 1%pt), meaning differences have a smaller impact.

2) *Qualitative Evaluation:* The qualitative comparison in Fig. 7 visualizes the impact of FE and RGB information loss and its influence on model behavior. The generated predictions allow conclusions to be drawn about how variants trained on differently processed data react to representative scenarios. In line with Fig. 6, a representative model is selected for low (raw 3+RGBT) and high (GE 3+T) performing metrics each—specifically 640×512 resolution and seed number 42, since these are closest to their variants' average performance.

Fig. 7 clearly underlines the previous differentiation between high- and low-performing metrics. Common classes (i.e., buildings and cars) are overlooked far less often than the less prevalent

ones (i.e., street lamps) because of annotated instance and pixel amounts (see Appendix A).

A closer look at example (1) in Fig. 7 shows the 3+T model trained on FE data to be capable of more nuanced distinctions between common objects, such as warm and cold cars. It specifically associates only the warm pixels with said class, actually surpassing the annotation mask in detail. The 3+RGBT model trained on raw data displays a tendency to overclassify, in this instance warm cars. This might be explained with the fact that the combined FE implemented in this study is able to compensate thermal halos such as those around the cold vehicles in the raw TIR image. In contrast, this may also be what allows the raw 3+RGBT variant to be somewhat more sensitive to highly underrepresented classes, as shown in example (2). Although not entirely correct, the raw 3+RGBT model's prediction is closer to the annotation mask in terms of warm manholes and street lamps, illustrating the reason for it scoring highest among the low-performing metrics.

While related work generally sets the expectation that including additional RGB information will bolster segmentation performance [6], example (3) in Fig. 7 contradicts this assumption. Although the RGB visually aids human observers in identifying the cars in the top image half, the results show the 3+RGBT model utilizing the raw imagery as incapable of doing so. The other variant, relying solely on FE thermal imagery, correctly classifies the nonobscured car. This, again, supports an important conclusion derived from the quantitative analysis, specifically that carefully selected FE can compensate or even outperform a lack of additional RGB data.

B. Resource Utilization

As expected, the consumption of both time and energy for model training is mostly dependent on image resolution. Fig. 8 exemplifies the correlation using averages for each image size. Interestingly, the relationship seems to follow a logarithmic growth curve. Although high-resolution images are $6\times$ the size of mid ones, the difference in resource requirements is negligible—especially compared to small images half their size. Significant time and energy savings can only be achieved with small images, or conversely, higher resolutions are not problematic in either regard after a certain saturation threshold has been met.

The impact of FE on resource utilization, as highlighted in Fig. 8, is secondary in comparison to resolution. FE is negligible with respect to the training duration, as Fig. 8 shows that the training times are only seconds apart. Where energy consumption is concerned, training with FE instead of raw imagery requires somewhat more kW·h. These numbers increase only slightly for higher resolutions, from ca. 12% to 14%. However, these are one-time costs, as the model requires training only once before it can be utilized. In terms of data preprocessing, GE unsurprisingly takes the longest, followed by PS. This also means that more time must be anticipated for inference, although it only amounts to 0.4 s (GE) versus 0.12 s (raw) per image for the highest resolution.

VII. CONCLUSION

This study analyzed the influence of FE—specifically PS (VC) and GE (contrast increase and deblurring)—on a novel RGBT dataset for the task of multiclass thermal urban feature segmentation. We find such manual FE to account for a 7%pt – 10%pt difference in performance and can thus discredit the common assumption that DL will directly infer engineered features itself. While supplementary RGB information is beneficial when working with raw imagery, the overall best results are achieved by applying in-depth GE FE and using only TIR data. This has a significant implication for thermography-focused implementations as it means a less expensive acquisition with simple thermal sensors can counteract a lack of RGB information if FE is performed. In the context of managing more sustainable cities, this may have wide-reaching implications for future designs of smart city monitoring applications, where the problem of leakage detection in pervasive DHS technology can be combined with other, solely thermography-based applications such as building and solar panel inspections by cheaper means.

Regarding image size, high-resolution data not only yield the best results most often, but also particularly improves those metrics performing less well due to class imbalance. Owing to the analysis of time and energy consumption, we now know that this costs surprisingly little additional resources owing to their logarithmic relationship. In contrast to findings of previous work, high-quality TIR sensors are crucial in collecting data that will improve performance of economical, heat-related DL models.

This study is subject to some limitations. Only four seeds were used for initialization, lessening the statistical significance of the calculated SD [24]. Data annotation and registration are subject to human error. In future, FE studies—especially regarding RGBs and data assimilation—can provide further insights. Additional experiments can help quantify the contribution of each implemented filter to DL model performance using standard, spectral, and multispectral data. This includes analyses using explainable AI to help further understand and characterize the models trained on differently preprocessed data. Implementing the presented FE in combination with other models will help to assess the general applicability of the derived conclusions.

APPENDIX A

An overview of class distributions is given in Table III.

TABLE III
OVERVIEW OF CLASS DISTRIBUTIONS

Class	Polygon count	Pixel count
background	-	205,357,517
building	1,404	48,111,260
car (cold)	2,532	3,804,713
car (warm)	1,036	1,993,912
manhole (cold)	520	92,415
manhole (warm)	1,379	244,538
miscellaneous	81	50,762
person	275	38,901
street lamp (cold)	100	22,822
street lamp (warm)	683	133,400

APPENDIX B

Due to case study size, the data are simply split into 80% training and 20% test sets. While the split is randomized, the following three conditions are ensured:

- 1) all classes are represented in both sets;
- 2) both sets contain images from both cities;
- 3) the annotation distribution is close to 80–20.

Each configuration is trained with four seeds, arbitrarily chosen to initialize model weights unspecified by transfer learning. These are: 42, 1000, 1 234 567, and 10 110 110. All variants are trained on the bwUniCluster2.0 high-performance computing system using a NVIDIA A100-PCI GPU for 35 epochs and a batch size of 8. We use Python 3.8 with OpenCV 4.6.0.66 and scikit-image 0.19.3 for processing, segmentation_models 1.0.1 (tensorflow 2.10.0) for training [54], tensorflow-addons 0.20.0 for loss definition, and scikit-learn 1.3.2 for evaluation.

Data availability

The utilized, novel RGBT dataset will be made available with this publication via Zenodo [15]. The code is available at <https://github.com/emvollmer/TUFSeg>.

Authors' contributions

Elena Vollmer: Conceptualization, methodology, data curation, investigation, software development, formal analysis, visualization, writing—original draft preparation, and writing—review and editing; Mishal Benz: Conceptualization, methodology, and writing—review and editing; James Kahn: Data curation and software development; Leon Klug: Methodology, software development, and visualization. Rebekka Volk: Writing—review and editing, and supervision; Frank Schultmann: Writing—review and editing, and supervision; Markus Götz: Writing—review and editing, and supervision. All authors have read and agreed to the published version of the manuscript.

ACKNOWLEDGMENT

The datasets were acquired in collaboration with the Air Bavarian GmbH and Munich's and Karlsruhe's municipal utilities companies. The authors acknowledge support by the state of Baden-Württemberg through bwHPC.

REFERENCES

- [1] United Nations Environment Programme and 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," 2024, Accessed: Jun. 28, 2024. [Online]. Available: <https://wedocs.unep.org/20.500.11822/45095>
- [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]," 2023, Accessed: Jun. 28, 2024. [Online]. Available: <https://www.bmwsb.bund.de/SharedDocs/gesetzgebungsverfahren/Webs/BMWSB/DE/kommunale-waermeplanung.html>
- [3] A. Fernwärme and J. Dornberger, "Hauptbericht 2022 [main report 2022]," AGFW, 2023. Accessed: Jun. 28, 2024. [Online]. Available: <https://www.agfw.de/zahlen-und-statistiken/agfw-hauptbericht>
- [4] N. Bayomi and J. E. Fernandez, "Eyes in the sky: Drones applications in the built environment under climate change challenges," *Drones*, vol. 7, no. 10, 2023, Art. no. 637, doi: [10.3390/drones7100637](https://doi.org/10.3390/drones7100637).
- [5] Y. He, B. Deng, H. Wang, and L. Cheng, "Infrared machine vision and infrared thermography with deep learning: A review," *Infrared Phys. Technol.*, vol. 116, 2021, Art. no. 103754, doi: [10.1016/j.infrared.2021.103754](https://doi.org/10.1016/j.infrared.2021.103754).
- [6] Z. Küttük and G. Algan, "Semantic segmentation for thermal images: A comparative survey," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 285–294, doi: [10.1109/CVPRW56347.2022.00043](https://doi.org/10.1109/CVPRW56347.2022.00043).
- [7] K. Song, Y. Zhao, L. Huang, Y. Yan, and Q. Meng, "RGB-T image analysis technology and application: A survey," *Eng. Appl. Artif. Intell.*, vol. 120, 2023, Art. no. 105919, doi: [10.1016/j.engappai.2023.105919](https://doi.org/10.1016/j.engappai.2023.105919).
- [8] R. Gade and T. B. Moeslund, "Thermal cameras and applications: A survey," *Mach. Vis. Appl.*, vol. 25, no. 1, pp. 245–262, 2014, doi: [10.1007/s00138-013-0570-5](https://doi.org/10.1007/s00138-013-0570-5).
- [9] E. Vollmer, R. Volk, and F. Schultmann, "Automatic analysis of UAS-based thermal images to detect leakages in district heating systems," *Int. J. Remote Sens.*, vol. 44, pp. 7263–7293, 2023, doi: [10.1080/01431161.2023.2242586](https://doi.org/10.1080/01431161.2023.2242586).
- [10] Y. Zhang, D. Sidibé, O. Morel, and F. Mériaudeau, "Deep multimodal fusion for semantic image segmentation: A survey," *Image Vis. Comput.*, vol. 105, 2021, Art. no. 104042, doi: [10.1016/j.imavis.2020.104042](https://doi.org/10.1016/j.imavis.2020.104042).
- [11] K. Maharana, S. Mondal, and B. Nemade, "A review: Data pre-processing and data augmentation techniques," *Glob. Transitions Proc.*, vol. 3, no. 1, pp. 91–99, 2022, doi: [10.1016/j.gltp.2022.04.020](https://doi.org/10.1016/j.gltp.2022.04.020).
- [12] W. Yuan and W. Hua, "A case study of vignetting nonuniformity in UAV-Based uncooled thermal cameras," *Drones*, vol. 6, no. 12, 2022, Art. no. 394, doi: [10.3390/drones6120394](https://doi.org/10.3390/drones6120394).
- [13] X. Zeng, J. Xu, and X. Gao, "A potential method for the nonuniformity correction and noise removal of infrared thermal image," *Acta Physica Polonica A*, vol. 137, pp. 1055–1060, 2020, doi: [10.12693/APhysPolA.137.1055](https://doi.org/10.12693/APhysPolA.137.1055).
- [14] C. Debus, M. Piraud, A. Streit, and M. Götz, "Reporting electricity consumption is essential for sustainable AI," *Nature Mach. Intell.*, vol. 5, pp. 1176–1178, 2023, doi: [10.1038/s42256-023-00750-1](https://doi.org/10.1038/s42256-023-00750-1).
- [15] E. Vollmer et al., "Thermal urban feature segmentation—Multispectral (RGB + Thermal) UAS-based images from Germany with annotations," Zenodo, 2025, doi: [10.5281/zenodo.10814413](https://doi.org/10.5281/zenodo.10814413).
- [16] B. Neupane, T. Horanont, and J. Aryal, "Deep learning-based semantic segmentation of urban features in satellite images: A review and meta-analysis," *Remote Sens.*, vol. 13, no. 4, 2021, Art. no. 808, doi: [10.3390/rs13040808](https://doi.org/10.3390/rs13040808).
- [17] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-Thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 1909, pp. 9441–9447, doi: [10.1109/ICRA40945.2020.9196831](https://doi.org/10.1109/ICRA40945.2020.9196831).
- [18] T.-S. Wang, G. T. Kim, M. Kim, and J. Jang, "Contrast enhancement-based preprocessing process to improve deep learning object task performance and results," *Appl. Sci.*, vol. 13, no. 19, 2023, Art. no. 10760, doi: [10.3390/app131910760](https://doi.org/10.3390/app131910760).
- [19] Y. H. Robinson, S. Vimal, M. Khari, F. C. L. Hernández, and R. G. Crespo, "Tree-based convolutional neural networks for object classification in segmented satellite images," *Int. J. High Perform. Comput. Appl.*, 2020, Art. no. 1094342020945026, doi: [10.1177/1094342020945026](https://doi.org/10.1177/1094342020945026).
- [20] M. Chaverot, M. Carré, M. Jourlin, A. Bensrhair, and R. Grisel, "Improvement of small objects detection in thermal images," *Integr. Comput.-Aided Eng.*, vol. 30, no. 4, pp. 1875–8835, 2023, doi: [10.3233/ICA-230715](https://doi.org/10.3233/ICA-230715).
- [21] C. Herrmann, M. Ruf, and J. Beyerer, "CNN-based thermal infrared person detection by domain adaptation," *Proc. SPIE*, vol. 10643, 2018, Art. no. 1064308, doi: [10.1117/12.2304400](https://doi.org/10.1117/12.2304400).
- [22] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "MFNet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst.*, 2017, pp. 5108–5115, doi: [10.1109/IROS.2017.8206396](https://doi.org/10.1109/IROS.2017.8206396).
- [23] Y. Hou, R. Volk, M. Chen, and L. Soibelman, "Fusing tie points' RGB and thermal information for mapping large areas based on aerial images: A study of fusion performance under different flight configurations and experimental conditions," *Automat. Construction*, vol. 124, 2021, Art. no. 103554, doi: [10.1016/j.autcon.2021.103554](https://doi.org/10.1016/j.autcon.2021.103554).
- [24] Z. Mayer, J. Kahn, Y. Hou, M. Götz, R. Volk, and F. Schultmann, "Deep learning approaches to building rooftop thermal bridge detection from aerial images," *Automat. Construction*, vol. 146, 2023, Art. no. 104690, doi: [10.1016/j.autcon.2022.104690](https://doi.org/10.1016/j.autcon.2022.104690).

- [25] D. Mehta, A. Bagubali, A. N. Joseph, V. Kumar, V. Karar, and S. Poddar, "Radial distortion estimation using analytical technique," in *Proc. 2nd IEEE Int. Conf. Recent Trends Electron., Inf. Commun. Technol.*, 2017, pp. 1587–1591, doi: [10.1109/RTEICT.2017.8256866](https://doi.org/10.1109/RTEICT.2017.8256866).
- [26] G. Bradski, "The OpenCV library," *Dr Dobbs's J. Softw. Tools*, vol. 22, no. 11, pp. 120–125, 2000.
- [27] K. Adrian and B. Gary, *Learning OpenCV 3: Computer Vision in C With the OpenCV Library*, 1st ed. Newton, MA, USA: O'Reilly Media, 2016.
- [28] Z. Mayer et al., "Thermal bridges on building rooftops," *Sci. Data*, vol. 10, no. 1, 2023, Art. no. 268, doi: [10.1038/s41597-023-02140-z](https://doi.org/10.1038/s41597-023-02140-z).
- [29] J. Jiang et al., "Analysis and evaluation of the image preprocessing process of a six-band multispectral camera mounted on an unmanned aerial vehicle for winter wheat monitoring," *Sensors*, vol. 19, no. 3, 2019, Art. no. 747, doi: [10.3390/s19030747](https://doi.org/10.3390/s19030747).
- [30] H. Choi, J. P. Yun, B. J. Kim, H. Jang, and S. W. Kim, "Attention-based multimodal image feature fusion module for transmission line detection," *IEEE Trans. Ind. Inform.*, vol. 18, no. 11, pp. 7686–7695, Nov. 2022, doi: [10.1109/TII.2022.3147833](https://doi.org/10.1109/TII.2022.3147833).
- [31] A. Kordecki, H. Palus, and A. Bal, "Practical vignetting correction method for digital camera with measurement of surface luminance distribution," *Signal, Image Video Process.*, vol. 10, pp. 1417–1424, 2016, doi: [10.1007/s11760-016-0941-2](https://doi.org/10.1007/s11760-016-0941-2).
- [32] A. Bal and H. Palus, "Image vignetting correction using a deformable radial polynomial model," *Sensors*, vol. 23, no. 3, 2023, Art. no. 1157.
- [33] A. S. Parihar and K. Singh, "A study on retinex based method for image enhancement," in *Proc. 2nd Int. Conf. Inventive Syst. Control*, 2018, pp. 619–624, doi: [10.1109/ICISC.2018.8398874](https://doi.org/10.1109/ICISC.2018.8398874).
- [34] S. Strat, A. Benoit, and P. Lambert, "Retina enhanced bag of words descriptors for video classification," in *Proc. 22nd Eur. Signal Process. Conf.*, 2014, pp. 1307–1311.
- [35] J. Liu et al., "Illumination and contrast balancing for remote sensing images," *Remote Sens.*, vol. 6, no. 2, pp. 1102–1123, 2014, doi: [10.3390/rs6021102](https://doi.org/10.3390/rs6021102).
- [36] Y. Yoshimi et al., "Image preprocessing with contrast-limited adaptive histogram equalization improves the segmentation performance of deep learning for the articular disk of the temporomandibular joint on magnetic resonance images," *Oral Surg., Oral Med., Oral Pathol. Oral Radiol.*, vol. 138, pp. 128–141, 2023, doi: [10.1016/j.oooo.2023.01.016](https://doi.org/10.1016/j.oooo.2023.01.016).
- [37] M. Zhou, K. Jin, S. Wang, J. Ye, and D. Qian, "Color retinal image enhancement based on luminosity and contrast adjustment," *IEEE Trans. Biomed. Eng.*, vol. 65, no. 3, pp. 521–527, Mar. 2018, doi: [10.1109/TBME.2017.2700627](https://doi.org/10.1109/TBME.2017.2700627).
- [38] G. Deng, "A generalized unsharp masking algorithm," *IEEE Trans. Image Process.*, vol. 20, no. 5, pp. 1249–1261, May 2011, doi: [10.1109/TIP.2010.2092441](https://doi.org/10.1109/TIP.2010.2092441).
- [39] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput. Assisted Interv.*, 2015, pp. 234–241, doi: [10.1007/978-3-319-24574-4_28](https://doi.org/10.1007/978-3-319-24574-4_28).
- [40] J. Lv, Q. Shen, M. Lv, Y. Li, L. Shi, and P. Zhang, "Deep learning-based semantic segmentation of remote sensing images: A review," *Front. Ecol. Evol.*, vol. 11, 2023, Art. no. 1201125, doi: [10.3389/fevo.2023.1201125](https://doi.org/10.3389/fevo.2023.1201125).
- [41] A. Majidizadeh, H. Hasani, and M. Jafari, "Semantic segmentation of UAV images based on U-NET in urban area," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 10, pp. 451–457, 2023, doi: [10.5194/isprs-annals-X-4-W1-2022-451-2023](https://doi.org/10.5194/isprs-annals-X-4-W1-2022-451-2023).
- [42] I. Ulku, P. Barmpoutis, T. Stathaki, and E. Akagunduz, "Comparison of single channel indices for U-Net based segmentation of vegetation in satellite images," *Proc. SPIE*, vol. 11433, 2020, Art. no. 1143319, doi: [10.1117/12.2556374](https://doi.org/10.1117/12.2556374).
- [43] V. Iglovikov, S. Mushinskiy, and V. Osin, "Satellite imagery feature detection using deep convolutional neural network: A Kaggle competition," 2017, doi: [10.48550/arXiv.1706.06169](https://doi.org/10.48550/arXiv.1706.06169).
- [44] R. Dong, X. Pan, and F. Li, "DenseU-Net-Based semantic segmentation of small objects in urban remote sensing images," *IEEE Access*, vol. 7, pp. 65347–65356, 2019, doi: [10.1109/ACCESS.2019.2917952](https://doi.org/10.1109/ACCESS.2019.2917952).
- [45] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2999–3007, doi: [10.1109/ICCV.2017.324](https://doi.org/10.1109/ICCV.2017.324).
- [46] A. Adiba, H. Hajji, and M. Maatouk, "Transfer learning and U-Net for buildings segmentation," in *Proc. New Challenges Data Sci.: Acts 2nd Conf. Moroccan Classification Soc.*, 2019, pp. 1–6, doi: [10.1145/3314074.3314088](https://doi.org/10.1145/3314074.3314088).
- [47] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.
- [48] SZ DJI Technology Co. Ltd., "Matrice 600 pro," 2018. Accessed: Jun. 28, 2024. [Online]. Available: <https://www.dji.com/matrice600-pro>
- [49] SZ DJI Technology Co. Ltd., "Matrice 300 RTK," 2020. Accessed: Jun. 28, 2024. [Online]. Available: <https://enterprise.dji.com/matrice-300/specs>
- [50] Zenmuse XT 2: User Manual, SZ DJI Technology Co. Ltd., 2018. Accessed: Jun. 28, 2024. [Online]. Available: <https://www.dji.com/downloads/products/zenmuse-xt2>
- [51] A. Dutta and A. Zisserman, "The VIA annotation software for images, audio and video," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 2276–2279, doi: [10.1145/3343031.3350535](https://doi.org/10.1145/3343031.3350535).
- [52] S. Van der Walt et al., "Scikit-image: Image processing in Python," *PeerJ*, vol. 2, 2014, Art. no. e453, doi: [10.7717/peerj.453](https://doi.org/10.7717/peerj.453).
- [53] J. P. G. H. Muriedas, K. Flügel, C. Debus, H. Obermaier, A. Streit, and M. Götz, "perun: Benchmarking energy consumption of high-performance computing applications," in *Euro-Par 2023: Parallel Processing*. Cham, Switzerland: Springer, 2023, pp. 17–31.
- [54] P. Iakubovskii, "Segmentation models," 2019. Accessed: Jul. 10, 2024. [Online]. Available: https://github.com/qubvel/segmentation_models,



Elena Vollmer received the B.Sc. and M.Sc. degrees in mechanical engineering in 2018 and 2021, respectively, from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany, with a semester at Trinity College Dublin, Dublin, Ireland. She is currently working toward the Ph.D. degree in engineering with the KIT.

Her research as a Ph.D. student and an Academic Associate focuses on heuristic and AI-based anomaly detection in thermal imagery, specifically to aid in the monitoring of energy-related systems for leak detection.



Mishal Benz received the B.E.E.E. degree in electrical engineering with specialization in electronics from Air University, Islamabad, Pakistan, in 2012, and the Ph.D. degree in electrical engineering from Sabanci University, Istanbul, Türkiye, in 2017.

She currently works with the Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany, as a Helmholtz AI Consultant. Her current work focuses on collaborating with other scientists and providing her expertise to resolve research issues in the field of energy using deep learning models.



James Kahn received the Ph.D. degree in physics (experimental particle physics) from the Ludwig Maximilian University of Munich, Munich, Germany, in 2019, as a member of the Excellence Cluster ORIGINS network.

He has worked as a Postdoctoral Researcher with the Scientific Computing Centre, Karlsruhe Institute of Technology, Karlsruhe, Germany for the Belle II experiment in 2019–2020 as a Scientific Computing Specialist, and for the Helmholtz AI consultants team as an Applied Artificial Intelligence (AI) Consultant

in 2020–2022. He is currently the Principal Software Engineer with HAL Systems Pty Ltd., Ivanhoe, Australia, where he leads research and development of AI-enabled predictive control and emissions reduction measures for commercial building energy consumption.



Leon Klug received the B.Sc. and M.Sc. degrees in industrial engineering from the Karlsruhe Institute of Technology, Karlsruhe, Germany, in 2020 and 2023, respectively, with a semester with the Zurich University of Applied Sciences, Winterthur, Switzerland.

During his studies, he conducted experiments on testing and optimizing neural networks for image segmentation. He currently works as a Data Scientist in Munich, Germany.



Frank Schultmann received the diploma, doctorate and habilitation degrees in industrial engineering, specifically management science, from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany (formerly University of Karlsruhe) in 1993, 1998 and 2003, respectively. He is a Chair Professor of business administration, production, and operations management with the KIT, as well as a Professor of complex project management with the University of Adelaide, Adelaide, Australia, and the Director with the Institute for Industrial Production, KIT, and the KIT's French-German Institute for Environmental Research (DFIU). His research interests include sustainable manufacturing and logistics, decision support, supply chain management and optimization, project management, technology assessment, construction management, and information and communication technologies.



Rebekka Volk received the diploma, doctorate and habilitation degrees in industrial engineering, specifically business economics, from the Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany in 2011, 2016 and 2022, respectively.

She had study and research stays with Universidad Politécnica de Madrid, Spain, University of Adelaide, Australia, and Southern University of California, Los Angeles, USA. She is currently a Postdoctoral Researcher with the Institute for Industrial Production, KIT, as the Head of research group "Resource man-

agement in the built environment." In line with this, her work revolves around developing and applying new methods to improve resource management in the built environment. Her research interests include sustainability assessment of technical, industrial, and urban systems, circular economy, and sustainable district development.



Markus Götz (Member, IEEE) received the B.Sc. and M.Sc. degrees in IT-system engineering from the University of Potsdam, Potsdam, Germany, in 2010 and 2014, respectively, and the Ph.D. degree in computational engineering from the University of Iceland, Reykjavik, Iceland, in conjunction with the Juelich Supercomputing Centre, Jülich, Germany, in 2017.

He had intermediate stays with the Blekinge Tekniska Högskola, Sweden and CERN, Switzerland.

He is currently a Postdoctoral Researcher with the Scientific Computing Centre, Karlsruhe Institute of Technology, Karlsruhe, Germany, as the Project Manager for the Helmholtz Analytics Framework and the Head of the Helmholtz AI Consultants Team. In line with his work, he focuses on applied artificial intelligence and data analysis on high-performance cluster systems to work on the grand challenges in the natural sciences. His research interests include machine learning, global optimization, as well as parallel algorithm engineering.