Automation in Construction 146 (2023) 104690

Contents lists available at ScienceDirect

Automation in Construction

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

Deep learning approaches to building rooftop thermal bridge detection from aerial images

Zoe Mayer [a,*], James Kahn [a,b], Yu Hou [c], Markus Götz [a,b], Rebekka Volk [a], Frank Schultmann [a]

[a] Karlsruhe Institute of Technology (KIT), Karlsruhe, 76131, Germany
[b] Helmholtz AI, Germany
[c] Carnegie Mellon University (CMU), USA

ARTICLE INFO

Keywords:
Building analysis
Thermal bridges
Drones
Deep learning
Computer vision
Object detection

ABSTRACT

Thermal bridges are weak points of building envelopes that can lead to energy losses, collection of moisture, and formation of mould in the building fabric. To detect thermal bridges of large building stocks, drones with thermographic cameras can be used. As the manual analysis of comprehensive image datasets is very time-consuming, we investigate deep learning approaches for its automation. For this, we focus on thermal bridges on building rooftops recorded in panorama drone images from our updated dataset of Thermal Bridges on Building Rooftops (TBBRv2), containing 926 images with 6,927 annotations. The images include RGB, thermal, and height information. We compare state-of-the-art models with and without pretraining from five different neural network architectures: MaskRCNN R50, Swin-T transformer, TridentNet, FSAF, and a MaskRCNN R18 baseline. We find promising results, especially for pretrained models, scoring an Average Recall above 50 % for detecting large thermal bridges with a pretrained Swin-T Transformer model.

1. Introduction

The emissions of carbon dioxide (CO_2) from the operation of buildings have increased to their highest level yet to around 27% of total global energy-related CO_2 emissions [1]. Thermal energy is particularly pertinent as more than half of global household energy use is for space and water heating [2]. A common reason for heat losses of buildings are thermal bridges. Thermal bridges are areas of the building envelope with low thermal resistance that conduct heat faster from the warmer inside to the colder outside than adjacent areas. Reasons for this are the geometry of constructions, different thermal conductivities of used materials, or air leaks of the building envelope. Energy losses caused by thermal bridges can make up to one third of the transmission heat loss of an entire building [3]. Moreover, they may lead to dampness and mould growth, which in the long term degrades the building fabric and is associated with health concerns caused by poor indoor air quality. For buildings inhabitants, thermal bridges also can lead to uncomfortable spaces due to cold interior surfaces [4,5].

To detect thermal bridges, thermography is currently the state-of-the-art [6]. Recording thermographic images with a terrestrial camera is a method that has been used for building audits and thermal bridge detection for many years [7]. Classical terrestrial thermography, though, lacks the ability to record rooftops or other parts of high buildings inaccessible from the ground [8]. Moreover, manually recorded thermographic images are not suitable for efficiently analysing the thermal quality of multiple buildings within a short time due to the time-consuming nature of the method and property rights, which only allow the capturing of street-views without owner permissions to enter properties [9,10]. The analysis of many buildings at urban scales, however, is becoming increasingly in demand. Examples include the development of retrofit plans for whole city districts like Community Energy Strategic Planning in the USA [11], Community Energy Planning in Canada [12], Positive Energy Districts in Europe [13], and "energetische Quartierskonzepte" in Germany [14].

To use thermography in urban environments, Unmanned Aerial Vehicles (UAVs, drones)[1] can be used for the scalable and automated recording of building images [8]. In this work, we compare the ability of five popular, state-of-the-art neural network architectures to automatically detect thermal bridges in aerial panorama images obtained using drones. In doing so, we also investigate the benefits of utilising additional height map information. We focus exclusively on thermal bridges of building rooftops as they can be exceptionally well captured from the aerial perspective. To perform this investigation, we utilise open-source computer vision libraries and analyse an updated Thermal

* Corresponding author.
E-mail addresses: zoe.mayer@partner.kit.edu (Z. Mayer), james.kahn@kit.edu (J. Kahn).

[1] Sometimes referred to as Unmanned Aerial Systems (UAS) that also include a drone pilot and the controlling system.

https://doi.org/10.1016/j.autcon.2022.104690
Received 10 July 2022; Received in revised form 13 November 2022; Accepted 26 November 2022

Bridges on Building Rooftops (TBBRv2). We have made the dataset, code, and all neural network configurations used in this work publicly available on Zenodo [15] and https://github.com/Helmholtz-AI-Energy/TBBRDet.

2. Related work

Non-stationary thermography with automated thermal bridge detection software has been investigated to speed up and simplify the process of building audits for large building stocks. In 2018, Garrido et al. [16] performed a study where they placed an infrared camera on the roof of a vehicle to record images of a building facade at an angle of 45°. They used an automatic detection approach and characterised thermal bridges based on geometric properties, measured temperature differences, and the calculation of the thermophysical properties of the linear heat transfer. The proportion of false positive detected thermal bridges was 45%, the proportion of missing thermal bridges was 32%, and the dataset used to evaluate the methodology only includes three images shown in the publication. Macher et al. [17] also installed an infrared camera on a vehicle. They intended to detect windows and thermal bridges by taking geometric and thermal characteristics into account and by modelling a thermographic 3D point cloud. For identification they used an iterative histogram approach to analyse global and local temperature maxima. They were able to reliably detect thermal bridges between floors and under balconies, and most windows could also be recognised automatically. The authors stated that windows located on the ground floor or basement are difficult to extract due to the limited field of view of the camera. Windows behind plants or objects cannot be detected in this way either. No quantitative information was given on the precision of the used algorithm.

A disadvantage of thermography with terrestrial vehicles is that no rooftops and only low facades facing the street can be analysed. Drones overcome this limitation. Due to their almost unlimited mobility, the entire outer envelope of a building can be recorded. In addition, the interference due to facade covering by e.g. trees or pedestrians walking past is reduced. Therefore, research is increasingly focusing on non-stationary thermographic audits by drones. Dios and Ollero [18] attempted to automatically detect and quantify heat losses through windows after a thermographic survey of buildings with a drone helicopter. They created heat maps of thermal images and defined a temperature difference of more than 7 °C to the facade as a criterion for a thermal bridge. Thermal irregularities were then classified according to their temperature distributions. This approach was suitable for detecting thermal bridges on windows, however it lacked the precise quantitative information for evaluating the results. Furthermore, this approach is not suitable for a fully automated evaluation. Rakha et al. [19] used a drone with a thermal camera to visually identify areas of thermal anomalies on building envelopes. They worked with a manual temperature thresholding approach and automatic edge filtering to generate a 3D model of a building with its detected thermal bridges. They state the overall precision of their algorithm of about 75%. Mirzabeigi and Razkenari [20] used thermographic cameras installed on a drone to collect close-up images from building sites. They designed a drone flight path for data collection and implemented a computer vision algorithm working with a dynamic thresholding approach to identify thermal anomalies of the building envelope. The study lacks in quantitative information on the quality of the thermal anomaly detection approach.

In all the aforementioned non-stationary thermography studies, thresholding and histogram approaches were applied. While they are applicable to close-up images, they encounter problems in panorama settings, which record multiple buildings and infrastructure in between with varying angles. There is a high likelihood of falsely identify thermal bridges coming from thermal anomalies in the background of buildings or to miss true thermal bridges with irregular shapes due to the varying recording perspectives. To detect thermal bridges only on buildings or specific building parts, like rooftops, a segmentation step to extract the building or building parts from the rest of the image is usually required, which is computationally expensive for images covering large areas such as a city district.

Supervised learning methods can aid in improving thresholding approaches. Utilising manually annotated training data, they are able to generalise and automatically annotate thermal bridges in previously unseen aerial images. Recently, Barahona et al. [21] used a camera on a car vehicle to detect thermal anomalies on building envelopes, such as thermal bridges, trained on 2000 labelled infrared images. They achieved a precision score of 89.2% and recall of 75.6% on a test dataset of 1184 infrared images. They used supervised learning with a linear model for panorama images to identify those containing anomalies, but segmented the anomalies and particular building components manually in a second, non-trained step to complement their results. Kim et al. [22] focused on terrestrial thermographic images and employed a neural network approach to detect thermal bridges. The study used a multi-step method including thermal anomaly area clustering, feature extraction, and an artificial-neural-network for thermal bridge detection. The average precision and recall of the detected thermal bridges for eight test images was 89% and 87%, respectively. However, the images used in the study are also close-ups of buildings and not panorama images.

Studies using deep learning approaches to detect thermal bridges on aerial thermographic panorama images are not known to the authors. In this study, we present deep learning neural network based approaches for detecting thermal bridges on panorama drone images on rooftops without building part segmentation. In doing so, we build on a previous publication [23] in which the authors presented the first results of the AI-based detection of thermal bridges on rooftops. To maximise the quality of our automated thermal bridge detection results, we use and compare multiple neural network architectures with and without pretraining on an open access dataset.

3. Methods and materials

3.1. Dataset

The dataset used in this study is an updated version of Thermal Bridges on Building Rooftops (TBBRv2) [15], consisting of five channels which are combined RGB[2] and thermal panorama drone images with a height map. Fig. 1 shows the RGB, thermal, and height map channels of an example image. The raw images for the dataset were recorded early in the morning in March 2019 in the inner city of Karlsruhe, Germany. All images are panorama images that, in addition to the actual objects of interest (buildings), also show the surrounding environment and infrastructure, such as streets, people, cars, trams, and trees.

The recorded area contains six large city perimeter blocks of roughly 20 buildings per block. Each building appears in the dataset around 20 times from different angles due to a high overlap rate during the recording process. The images were recorded with a normal (RGB) and a FLIR-XT2 (thermal) camera on a DJI M600 drone and are converted to a constant format of 2680 × 3370 pixels. Each image contains GPS information and flight altitudes (between 60–80 m above ground).

TBBRv2 contains 926 panorama images and annotations of 6927 thermal bridges on rooftops, split into train and test subsets. The training subset covers five building blocks recorded on 723 images with 5614 annotations, the test subset covers one building block recorded on 203 images with 1313 annotations. The updated TBBRv2 dataset provides more precise annotations due to better overlaps of the five information channels. These annotations only include thermal bridges that are visually clearly identifiable by experts, and thus also include thermal bridges that are not annotated due to being unclear. Because of

[2] RGB (Red Green Blue) images contain an information channel for each colour.

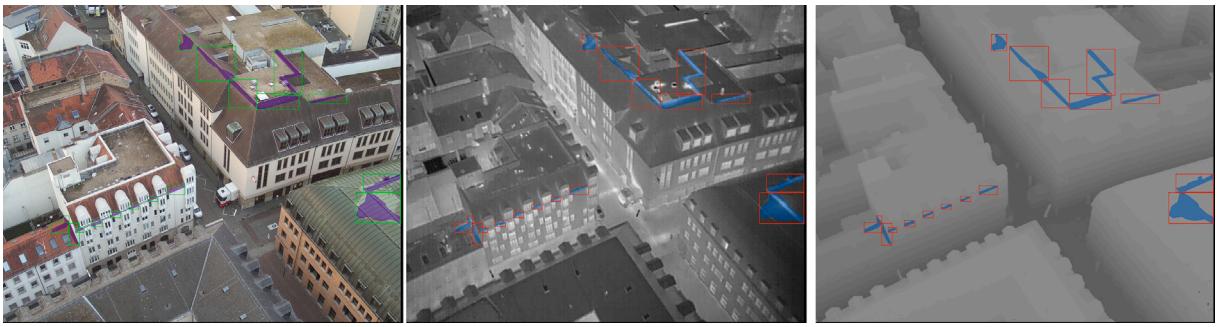

Fig. 1. Example annotations from the TBBRv2 dataset for RGB (left), thermal (centre), and height map (right).

image overlap, each thermal bridge is annotated on average about 20 times from different angles. The original TBBR dataset was published on Zenodo [24], full details of the image recording and dataset creation procedure can be found in Mayer et al. [23].

3.2. Object detection libraries

For the experiments in this paper, two popular computer vision libraries were used: Facebook AI Research's Detectron2 (v0.6) [25] and OpenMMLab's MMDetection (v2.21.0) [26]. Both of these libraries offer a framework within which object detection neural networks can be implemented, evaluated, and visualised. Our intention is to utilise popular, open-source libraries which offer ready-to-use, state-of-the-art (SOTA) object detection neural network architecture implementations. Given the significantly larger choice of object detection model implementations available in MMDetection, this library was used as the main implementation platform for the performed experiments in this work. Detectron2 was used only for comparing the results of this study to former results achieved with the TBBRv1 dataset [23]. There is otherwise no significant difference between the two libraries' capabilities.

3.3. Neural network architectures

While the specific implementations of object detection architectures varies, they predominantly follow the same procedure of first extracting meaningful features from the input image, and then translating these into task-specific predictions. Specifically, one can divide the components into a *backbone* to extract meaningful representations (feature maps) from the image pixels, a *neck* which is commonly used to further extract features for handling objects of different sizes/scales within the image (feature pyramid), and a *head* which uses the extracted features to make the output predictions [27]. In addition, there is generally a region proposal mechanism, which selects specific regions of interest within an image for the head to focus its predictions on. How these components are arranged and implemented in practice we will refer to as the framework.

For the first experiments in this work, performed using Detectron2, a MaskRCNN framework [28] with a ResNet-18 (R18) [29] backbone is used.[3] This is chosen for a direct comparison with that of Mayer et al. [23].

For the second experiments using MMDetection, of the implemented frameworks and backbones available, we consider only those with available pretrained models, as is required in our experiments. We first select a MaskRCNN with a ResNet-50 (R50) backbone for our baseline. The MaskRCNN R50 is a standard baseline comparison in object detection tasks. We then selected the following for comparison with the baseline: Swin-T Transformer [30], TridentNet [31], and Feature Selective Anchor-Free (FSAF) [32], with the explanation for each choice detailed in the following.

Transformer-based computer vision networks have outperformed popular object detection and instance segmentation benchmarks in recent years. In particular, the Shifted Windows (Swin) Transformer [30] and its variants, such as the Swin-V2 [33] and DINO [34], have dominated the popular Common Objects in Context (COCO) [35] object detection and instance segmentation benchmarks. In this work, the Swin-T transformer is tested as an alternative backbone for the MaskRCNN, which is roughly equivalent in size to a ResNet-50. An illustration of a Swin Transformer architecture is shown in Fig. 2.

Given the angled view of building rooftops in TBBRv2, the dataset contains different sized instances of same thermal bridges across multiple images. The TridentNet [31] architecture attempted to adapt the standard ResNet backbone of the Faster-RCNN framework [36] to be scale-aware.[4] We hypothesise that this will offer an advantage over the regular convolutions used in the baseline model.

The FSAF [32] model is a near-SOTA, single-shot, anchor-free framework, which unlike the Mask/FasterRCNN-based approaches, does not separate the region proposal and feature extraction stages. This has the advantage of removing the dependence on anchor boxes, whose predefined sizes will determine which objects in an image are processed at which scale (which feature map they are associated with). As with the scale-awareness issue that TridentNet attempts to address, this anchor dependence causes the same thermal bridge object, captured at different distances across multiple images, to be redundantly processed by different feature maps within the network. FSAF instead allows the model to dynamically learn the most appropriate feature map.

4. Experimental procedure

The experiments in this work are divided into two parts: first we demonstrate, using Detectron2, the performance improvements due to the updated TBBRv2 dataset and investigate the benefits of the height map inputs, then, using MMDetection, we explore the various object detection frameworks outlined in Section 3.3 to determine the optimal model and performance. An example of the experiment workflow for the pretrained MaskRCNN R50 baseline from MMDetection is shown in Fig. 3. In line with Mayer et al. [23], Average Recall (AR) scores averaged over the intersection over union (IoU) range $0.5\,to\,0.95$ on the test set are used to assess model performance. The AR score is defined as the ratio of correctly identified thermal bridges to all present thermal bridges. The IoU ranges which define what is considered an identified thermal bridge follow those of the commonly used COCO benchmark for object detection [35]. The reported AR score in all cases is that with

[3] The 18 in ResNet-18 here indicates the number of convolutional layers within the neural network.

[4] Scale-awareness means that a model is able to recognise the same object at different scales (sizes) in an image as being the same object, rather than learning each size as its own individual object.

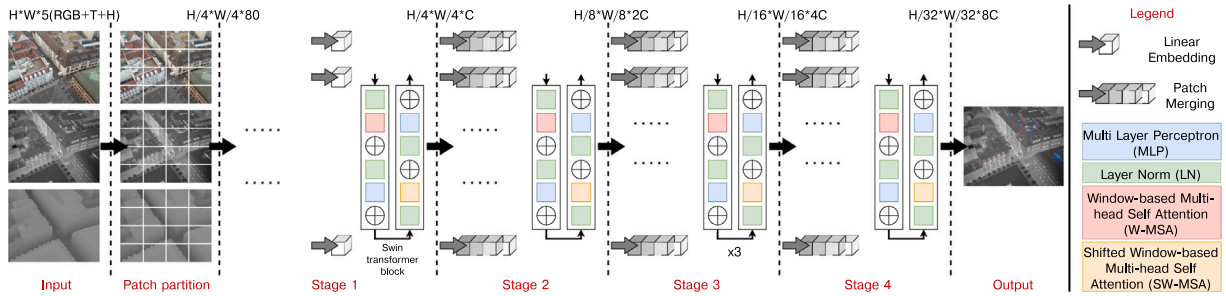

Fig. 2. Swin architecture overview. The input image is divided in 4 × 4 patches, which are then projected into a linear embedding and passed through successive Swin transformer blocks and patch merging layers until the final output representation of the image which is used to produce thermal bridge predictions. Image adapted from Liu et al. [30], where full details of each component can be found.

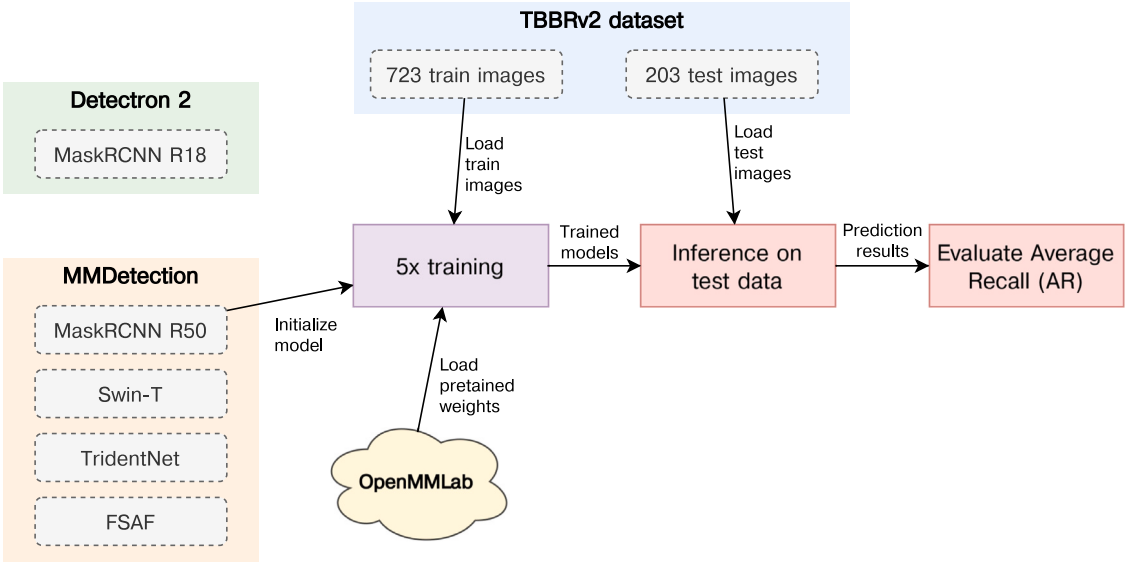

Fig. 3. Example experiment workflow for a pretrained MaskRCNN R50.

the highest AR for the top 100 detections per image (`AR@100`) across all training epochs. As the Average Precision (AP) penalises finding unannotated but correct thermal bridges, this metric is unsuitable for the TBBR dataset and not used here.

Five trainings are performed for each architecture tested, and the results are used to produce a mean and standard deviation.[5] We set the following five (randomly chosen) seeds to initialise the neural network weights, listed here for reproducibility: 3000, 10 117, 10 001, 20 770 001, 1 008 111. The same five seeds are used for all architecture trainings. The `deterministic` flag of MMDetection is also enabled in all experiments to maximise reproducibility. In all cases, the same pixel mean and standard deviation input normalisations are used as in Mayer et al. [23].

All trainings are performed on a single node of the HoreKa super computing system, located at Karlsruhe Institute of Technology (KIT), with four NVIDIA A100 40 GB GPUs in a data-distributed [37] manner. The nodes are reserved exclusively for each individual training and we report the total computing time and energy consumption [38] of the nodes used during training. Full details of all training configurations, along with the code used for training and evaluation, can be found at https://github.com/Helmholtz-AI-Energy/TBBRDet. Node hardware specifications are shown in Table A.3.

4.1. Detectron2 experiments

The experiments begin with an investigation into the improvements given by the updated alignments in TBBRv2. For this, the MaskRCNN R18 is configured according to Mayer et al. [23], running with the random seed used in that work (56689614). We follow this up with an ablation study in which we remove the height map data from the inputs. All other hyperparameters[6] are kept fixed, and the five random seeds described above used to estimate the variance in performance. Our aim is to investigate the benefit of height map information in ensuring predicted thermal bridges are located only on building rooftops and not on street level.

4.2. MMDetection experiments

In these experiments, a baseline model is trained using the MMDetection library. For this, the MaskRCNN framework with a ResNet-50 backbone is used. The baseline is trained both from scratch and using pretrained models from the MMDetection model zoo, trained on the popular computer vision benchmark Common Objects in Context (COCO 2017) [35]. The COCO dataset contains scenes with everyday objects in their regular context. We use it as model pretraining for two

[5] While five trainings is not enough for a statistically significant standard deviation, this does provide a useful insight into the fluctuation in performance due to the random seed.

[6] Hyperparameters are all parameters used to configure the architecture and training procedure that are not derived during the training itself. For example the number and type of layers in the network architecture, the random initialisation seeds, etc.

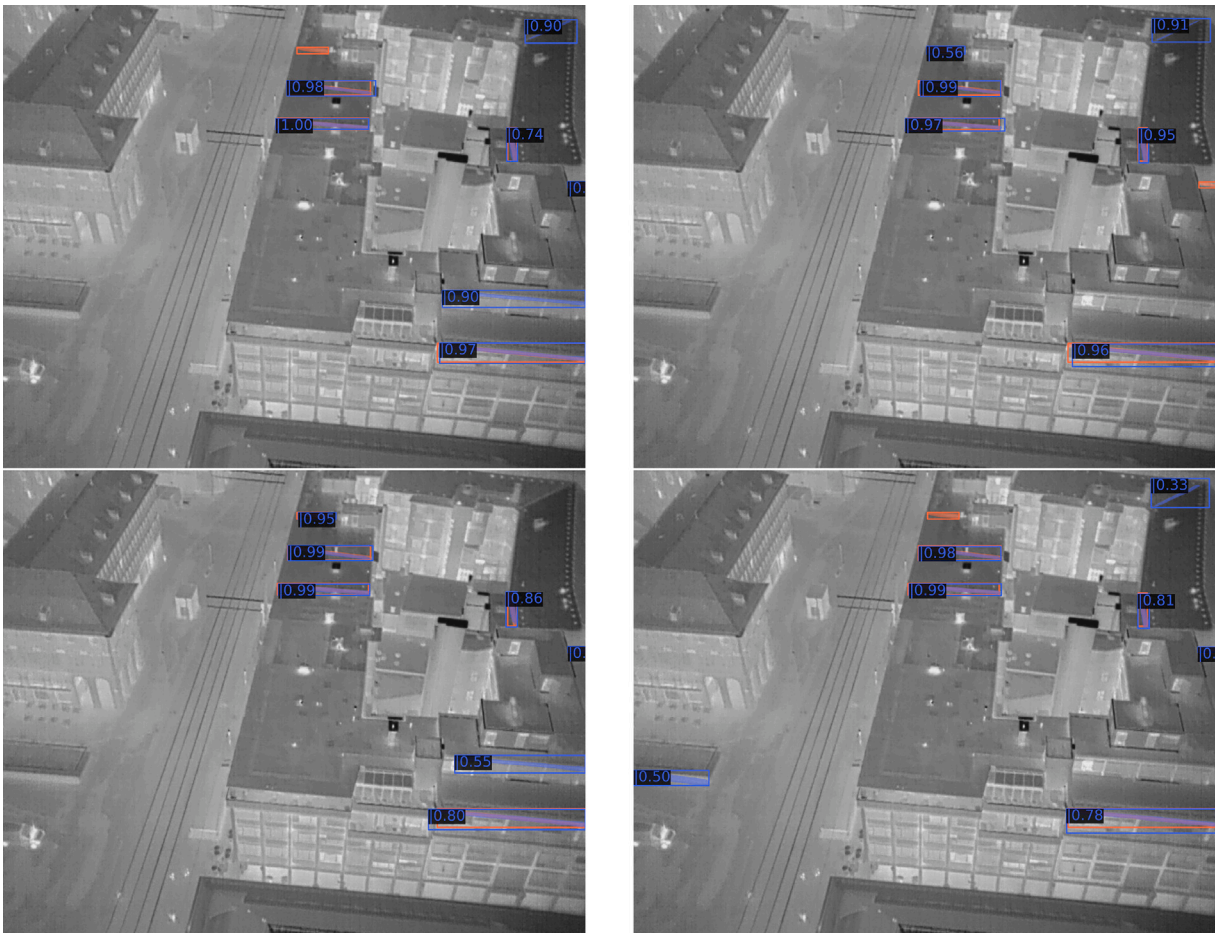

Fig. 4. Example predictions from MaskRCNN R50 baselines, numbers show model prediction scores. Left predictions are based on trainings from scratch and right predictions are based on trainings with pretraining. The top row shows predictions based on trainings with the full RGB + Thermal + Height inputs and bottom based on trainings of the RGB + Thermal ablation study.

reasons: first, OpenMMLab provides COCO pretrained versions of all models implemented in MMDetection, enabling accessible reproduction of our work, and second, the large corpus of everyday objects requires identifying edges, colour changes, etc., which, when finetuned on TBBRv2, we expect will transfer well to identifying thermal bridges on objects.

Using the baseline, the height map ablation study performed with Detectron2 is repeated. Based on both of these experiments' results, we determine the utility of the height map inputs, and proceed with training several other MMDetection architectures for comparison with the baseline.

Given the lack of a validation subset in TBBRv2, for all MMDetection-based experiments we forego hyperparameter optimisation, instead using the model configurations as-is wherever reasonably possible, making changes only to accommodate for the image sizes and extra input channels of our dataset. In particular, for multi-scale trainings, i.e. where inputs images are randomly resized as a form of data augmentation, we adjust the scales to their equivalent from our image sizes. All models are trained for a maximum of 36 epochs. In cases when the memory consumption exceeds that of the GPUs used (e.g. for the Swin Transformer), we use FP16 half precision[7] (floating point 16) for the neural network weights. According to the benchmarks of mixed precision trainings provided by MMDetection [26], this has a negligible impact on overall performance.

[7] IEEE 754-2019 [39] compliant binary representation of floating numbers using 16 bits, 1 for the sign, 5 for the exponent and 11 for the significant.

5. Results and discussion

In the following, we present the results of all experiments. We note here that these results are qualitative, in that they allow one to deduce thermal bridge locations and sizes across a large area. Due to high distances (>20 m) and varying recording angles of the drone relative to the buildings, a precise quantitative measurement of the thermal bridges cannot be made [40]. The interpretation of detected thermal bridges, e.g. for characterising them in terms of their risk of mould formation, energy losses, retrofit costs, or retrofit benefits, must be performed in a further step, for example using the methods presented by Mayer et al. [41].

Test results for the bounding box and segmentation AR scores, total node energy consumption in Megajoules (MJ) [38], and computing time in minutes are shown in Tables 1 and 2. We report the AR scores according to the standard object detection COCO benchmark [35]. The total AR is averaged across an Intersection-over-Union (IoU) between the predicted and ground truth thermal bridges of $0.5\,to\,0.95$, at different numbers of top-N (by prediction confidence) predictions: 1, 10, and 100 predictions. An additional score separation into medium (`AR_m`) and large (`AR_l`) objects is also given, for objects with an area between 32^2 and 96^2 pixels and greater than 96^2 pixels, respectively. The scores for small detection regions (less than 32^2 pixels) are not shown as they contain no thermal bridges. In the following subsections interpretations are given for the results of each architecture.

Table 1
Energy usage and bounding box Average Recall scores for each model's training. The ablation column indicates whether height information was excluded from the input. The MaskRCNN R18 architectures were trained using Detectron2. MaskRCNN R18* indicates the model initialised with random seed 56689614, used by Mayer et al. [23]. The best results are marked in bold.

Architecture	Pretrained	Ablation	Energy (MJ)	Time (min)	AR@1	AR@10	AR@100	AR_m@100	AR_l@100
MaskRCNN R18*			20.5	205.5	0.060	0.169	0.169	0.119	0.250
MaskRCNN R18			20.00 ± 0.20	205.3 ± 0.5	0.061 ± 0.002	0.165 ± 0.007	0.166 ± 0.006	0.129 ± 0.007	0.227 ± 0.010
		✓	19.42 ± 0.10	199.7 ± 0.6	0.060 ± 0.005	0.170 ± 0.010	0.170 ± 0.010	0.130 ± 0.020	0.230 ± 0.010
MaskRCNN R50			3.00 ± 0.03	39.5 ± 0.6	0.072 ± 0.008	0.270 ± 0.020	0.308 ± 0.008	0.270 ± 0.020	0.380 ± 0.010
	✓		2.83 ± 0.01	35.6 ± 0.4	0.076 ± 0.008	0.310 ± 0.020	0.370 ± 0.010	0.350 ± 0.020	0.420 ± 0.010
		✓	2.91 ± 0.03	38.1 ± 0.4	0.060 ± 0.010	0.260 ± 0.040	0.304 ± 0.007	0.280 ± 0.020	0.350 ± 0.020
	✓	✓	$\mathbf{2.74 \pm 0.01}$	$\mathbf{34.5 \pm 0.4}$	0.068 ± 0.004	0.290 ± 0.030	0.360 ± 0.020	0.350 ± 0.020	0.400 ± 0.020
Swin-T			7.90 ± 0.10	125.3 ± 1.3	0.069 ± 0.003	0.239 ± 0.007	0.318 ± 0.004	0.290 ± 0.010	0.370 ± 0.010
	✓		7.09 ± 0.03	107.3 ± 1.9	$\mathbf{0.089 \pm 0.006}$	$\mathbf{0.380 \pm 0.020}$	$\mathbf{0.454 \pm 0.007}$	$\mathbf{0.430 \pm 0.010}$	$\mathbf{0.507 \pm 0.007}$
TridentNet			4.92 ± 0.08	57.7 ± 1.0	0.031 ± 0.003	0.140 ± 0.010	0.215 ± 0.007	0.160 ± 0.010	0.311 ± 0.010
	✓		4.70 ± 0.10	51.9 ± 0.8	0.060 ± 0.010	0.210 ± 0.040	0.300 ± 0.050	0.220 ± 0.050	0.420 ± 0.070
FSAF			10.20 ± 0.09	103.7 ± 0.3	0.049 ± 0.008	0.150 ± 0.020	0.248 ± 0.008	0.223 ± 0.006	0.300 ± 0.010
	✓		10.00 ± 0.10	102.2 ± 0.3	0.070 ± 0.010	0.270 ± 0.020	0.380 ± 0.010	0.370 ± 0.020	0.410 ± 0.020

Table 2
Segmentation Average Recall scores for each model's training. As both FSAF and TridentNet are object detection architectures only and do not perform instance segmentation, they have no scores to report. Note that the FSAF and TridentNet are object detection frameworks and hence only predict bounding boxes. MaskRCNN R18* indicates the model initialised using the seed 56689614 used by Mayer et al. [23]. The best results are marked in bold.

Architecture	Pretrained	Ablation	AR@1	AR@10	AR@100	AR_m@100	AR_l@100
MaskRCNN R18*			0.040	0.094	0.094	0.069	0.134
MaskRCNN R18			0.037 ± 0.003	0.086 ± 0.002	0.086 ± 0.002	0.067 ± 0.004	0.119 ± 0.006
		✓	0.036 ± 0.001	0.089 ± 0.003	0.090 ± 0.003	0.073 ± 0.008	0.118 ± 0.004
MaskRCNN R50			0.047 ± 0.005	0.179 ± 0.008	0.201 ± 0.009	0.190 ± 0.010	0.225 ± 0.008
	✓		0.047 ± 0.005	0.190 ± 0.020	0.219 ± 0.008	0.217 ± 0.006	0.230 ± 0.020
		✓	0.041 ± 0.009	0.160 ± 0.020	0.191 ± 0.009	0.190 ± 0.010	0.210 ± 0.020
	✓	✓	0.040 ± 0.003	0.180 ± 0.030	0.220 ± 0.020	0.230 ± 0.020	0.220 ± 0.030
Swin-T			0.046 ± 0.002	0.153 ± 0.005	0.206 ± 0.004	0.203 ± 0.006	0.220 ± 0.007
	✓		$\mathbf{0.054 \pm 0.004}$	$\mathbf{0.230 \pm 0.020}$	$\mathbf{0.280 \pm 0.010}$	$\mathbf{0.280 \pm 0.010}$	$\mathbf{0.280 \pm 0.020}$

5.1. Detectron2 experiments

Comparing the bounding box average recall (AR@100) of the MaskRCNN R18* from Table 1 with that reported in Mayer et al. [23] of 9.4% (14.4% for large regions), we see an almost doubling of the performance. Given all else was equal, we can attribute this improvement to the improved annotations in TBBRv2 alone. Looking then at the training of the same model with the five random seeds, we see that they are relatively consistent with the MaskRCNN R18* result. We also observe that the ablation study results without height information are in agreement with those using the full RGB + Thermal + Height inputs, though we do note several falsely predicted thermal bridges on ground-level in the ablation trained model. Given there are only a small number of ground-level predictions, and that there appears to be no significant changes to the overall AR scores, we would therefore expect such false predictions to disappear given a larger training dataset. This is important, as the height map creation procedure used [42] is non-trivial, and therefore an obstacle to the ease of use of the preprocessing during the training procedure.

An interesting finding in the trainings using Detectron2 is the high energy consumption used during training. This was primarily due to extensive training times, which we found difficult to reduce, even when leveraging many dataloader processes to minimise data-loading times. While further expert optimisations are certainly possible to bring this down, we regard this result as a rather significant point against the ease-of-use factor when considering the Detectron2 library.

5.2. MaskRCNN R50 baseline

In all metrics, we observe agreement between the full and ablation trainings without height information. This holds for both trainings from scratch and those using a pretrained model. Fig. 4 shows an example of the predictions on a sample image from the test dataset for all training scenarios. Similar to the Detectron2 trained MaskRCNN R18, we observe several predictions on ground level in the ablation trained models. For this reason, we proceed with the remainder of experiments in this work using the full RGB + Thermal + Height input. However, as the ground-level predictions only detect 16 unique objects appearing across 23 images within the entire test dataset for the ablation training from scratch, we again believe that a larger labelled training dataset would resolve this and allow the RGB + Thermal information to be sufficient. We also observe a significant improvement in performance given by the pretraining, including a 5%*to*7% higher score for all AR@100 metrics.

The AR scores significantly outperform the MaskRCNN R18, almost doubling the AR@100. While this is likely due to the increased model size, i.e. more layers, we also note the drastically lower energy consumption due to a significant speedup observed in model training time. This demonstrates an excellent overall out-of-the-box performance of the MMDetection library.

5.3. Comparison with baseline

The pretrained Swin-T transformer achieves the highest AR by a significant margin. Interestingly, the from-scratch training (without

pretraining) only scores as well as the baseline. Transformer-based models are notoriously memory-hungry, and we see this reflected in a longer overall training time and hence larger energy consumption.[8]

The TridentNet architecture performs worse than the baseline on all metrics. We found that the pretrained model was especially unstable during training, regularly suffering from exploding gradients (and hence loss values), only stabilising when the learning rate was turned down an order of magnitude from the default 0.02 to 0.002.

FSAF also performs worse than the baseline in the from-scratch training, but the same when pretrained. Due to its ResNet-101 backbone, the training times were significantly longer, something we see reflected in the fact that it has the largest energy consumption of all MMDetection trainings.

Overall, the use of COCO [35] pretrained model weights proved to be an advantage regardless of the architecture. We therefore recommend this as an essential component in thermal bridge detection when utilising learned object detection approaches, and suggest it as an avenue of investigation for further improving detection performance.

6. Conclusion and outlook

The detection of thermal bridges on building rooftops can be automated by using deep learning approaches on thermographic images. For aerial panorama images, the main advantage of neural networks instead of computer vision approaches working with temperature thresholds is the ability to learn identifying building parts of interest and to include changing shapes of thermal bridges due to different recording angles.

In this study, the best results were achieved with the MMDetection library using a pretrained Swin-T Transformer model, scoring an Average Recall of 50.7% for large thermal bridges. Overall, we find consistently better results for pretrained models than for models without pretraining. Moreover, this work showed the ability of neural networks to propose predictions of thermal bridges only on rooftops by using height information to the input images. While this work has demonstrated promising results in identifying individual thermal bridges from drone images, we believe there is still significant potential for improvement with a larger annotated dataset. A larger dataset would allow for the allocation of a validation subset, enabling tuning of hyperparameters to improve training performance.

While no existing works target the detection of thermal bridges on aerial panorama images with deep learning approaches, we can compare our results with other existing thermal bridge detection procedures. Barahona et al. [21] achieved an Average Recall of 75% for the binary classification of images containing thermal anomalies, however they segmented the anomalies as a subsequent manual step, something that our approach automates entirely. Kim et al. [22] used a multi-step procedure on close-up images to achieve an Average Recall of 87%, which does not deal with the presence of multiple buildings and non-building objects within images. However, both of these works, the latter in particular, represent an optimal benchmark that may be achieved by our approach with the larger dataset proposed above.

Our scoring for this work has only considered the raw Average Recall score across each individual image, yet the images are not independent and instead contain significant overlap. We therefore propose in future to consider the scores across all instances of the same thermal bridge. For this, it is possible to track instances across all images containing the same thermal bridge or to set a threshold for requiring at least two detections of a thermal bridge to count it [43]. Identifying which thermal bridges are matching instances, however, would result in additional effort when creating the dataset.

Further improvements can also be made in the pretraining procedure, which has already proven successful in improving performance. Performing additional pretraining on existing UAV datasets, such as UAVDT [44] or iSAID [45], is one example that would closer align the pretrained models with the TBBR images. Self-supervised pretraining, used with great success in BERT [46], performed on the larger set of unannotated TBBR images presents another avenue for investigation.

Despite these limitations of our study, we believe we have provided important insights into the benefits of deep learning for automated building analysis in an urban context, which is becoming increasingly important in building and district management. In future, our approach could also be transferred to the analysis of other thermal anomalies on panorama drone images, such as the detection of district heating pipe leakages at ground level.

CRediT authorship contribution statement

Zoe Mayer: Conceptualization, Methodology, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Project administration. **James Kahn:** Conceptualization, Methodology, Software, Formal analysis, Investigation, Data Curation, Writing - Original Draft, Writing - Review & Editing, Visualization. **Yu Hou:** Methodology, Software, Writing - Original Draft, Writing - Review & Editing, Visualization. **Markus Götz:** Methodology, Writing - Review & Editing, Supervision. **Rebekka Volk:** Conceptualization, Data Curation, Writing - Review & Editing, Supervision, Project administration. **Frank Schultmann:** Writing - Review & Editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data, code, and configurations used in this work have been made publicly available online.

Acknowledgements

This work is supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant and the HAICORE@KIT partition. We thank Marinus Vogl and the Air Bavarian GmbH for their support with equipment and service for the recording of images. We also thank Tobias Beiersdörfer for support in the development of the TBBR dataset. All authors approved the version of the manuscript to be published.

Funding

All of the sources of funding for the work described in this publication are acknowledged below:

This work is supported by the Helmholtz Association Initiative and Networking Fund under the Helmholtz AI platform grant and the HAICORE@KIT partition.

Appendix. Training hardware details

See Table A.3.

[8] Memory itself is has a substantial power draw.

Table A.3
Hardware details for nodes used in all model trainings.

CPUs	Intel Xeon Platinum 8368
CPU Sockets per node	2
CPU Cores per node	76
CPU Threads per node	152
Cache L1	64k (per core)
Cache L2	1 MB (per core)
Cache L3	57 MB (shared, per CPU)
Main memory	512 GB
Accelerators	4x NVIDIA A100-40
Memory per accelerator	40 GB
Local discs	960 GB NVMe SSD
Interconnect	InfiniBand

References

[1] 2021 Global Status Report for Buildings and Construction: Towards a Zero-Emission, Efficient and Resilient Buildings and Construction Sector, Technical Report, United Nations Environment Programme (2021), United Nations, Nairobi, 2021, URL https://globalabc.org/resources/publications/2021-global-status-report-buildings-and-construction, accessed 2022-04-22.

[2] Energy Technology Perspectives 2014, International Energy Agency, 2014, http://dx.doi.org/10.1787/energy_tech-2014-en.

[3] T.G. Theodosiou, A.M. Papadopoulos, The impact of thermal bridges on the energy demand of buildings with double brick wall constructions, Energy Build. 40 (2008) 2083–2089, http://dx.doi.org/10.1016/j.enbuild.2008.06.006.

[4] P. Schmidt, S. Windhausen, Bauphysik-Lehrbuch: Wärmeschutz - Energieeinsparung - Feuchte- Und Tauwasserschutz - Schallschutz - Raumakustik, Reguvis Fachmedien GmbH, ISBN: 978-3-8462-0407-8, 2017.

[5] A. Alhawari, P. Mukhopadhyaya, Thermal bridges in building envelopes – an overview of impacts and solutions, Int. Rev. Appl. Sci. Eng. 9 (2018) 31–40, http://dx.doi.org/10.1556/1848.2018.9.1.5.

[6] I. Garrido, M. Solla, S. Lagüela, M. Rasol, Review of InfraRed thermography and ground-penetrating radar applications for building assessment, Adv. Civ. Eng. 2022 (2022) e5229911, http://dx.doi.org/10.1155/2022/5229911.

[7] E. Lucchi, Applications of the infrared thermography in the energy audit of buildings: A review, Renew. Sustain. Energy Rev. 82 (2018) 3077–3090, http://dx.doi.org/10.1016/j.rser.2017.10.031.

[8] B. Tejedor, E. Lucchi, I. Nardi, Application of qualitative and quantitative infrared thermography at urban level: Potential and limitations, in: D. Bienvenido-Huertas, J. Moyano-Campos (Eds.), New Technologies in Building and Construction: Towards Sustainable Development, in: Lecture Notes in Civil Engineering, Springer Nature, Singapore, 2022, pp. 3–19, http://dx.doi.org/10.1007/978-981-19-1894-0_1.

[9] M. Previtali, L. Barazzetti, R. Brumana, F. Roncoroni, Thermographic analysis from uav platforms for energy efficiency retrofit applications, J. Mob. Multimedia (2013) 066–082, URL https://journals.riverpublishers.com/index.php/JMM/index, accessed 2022-11-08.

[10] G. Bitelli, P. Conte, T. Csoknyai, F. Franci, V.A. Girelli, E. Mandanici, Aerial thermography for energetic modelling of cities, Remote Sens. 7 (2015) 2152–2170, http://dx.doi.org/10.3390/rs70202152.

[11] U.S. Department of Energy (DOE), Guide To Community Energy Strategic Planning, Technical Report, U.S. Department of Energy (DOE), 2013, URL https://www.energy.gov/eere/slsc/guide-community-energy-strategic-planning, accessed 2022-11-08.

[12] Dale Littlejohn, Richard Laszlo, National Report on Community Energy Plan Implementation, Technical Report, Quality Urban Energy Systems of Tomorrow (QUEST), 2015, URL https://questcanada.org/national-report-on-community-energy-plan-implementation/, accessed 2022-11-08.

[13] JPI Urban Europe, SET plan action 3.2, in: White Paper on PED Reference Framework for Positive Energy Districts and Neighbourhoods, Technical Report, JPI Urban Europe, 2020, URL https://jpi-urbaneurope.eu/ped/, accessed 2022-11-08.

[14] Federal Ministry of the Interior, Building and Community (BMI), Energy-Efficient Urban Redevelopment: A Funding Programme for Climate Protection at the Neighbourhood Level, Technical Report, Federal Ministry of the Interior, Building and Community (BMI), 2020, URL https://www.bmi.bund.de/SharedDocs/downloads/EN/publikationen/building/energie-efficient-urban-redevelopment.html.

[15] Z. Mayer, J. Kahn, Y. Hou, T. Beiersdörfer, M. Götz, R. Volk, Thermal bridges on building rooftops - hyperspectral (RGB thermal height) drone images of Karlsruhe, Germany, with thermal bridge annotations, 2022, http://dx.doi.org/10.5281/zenodo.6517768.

[16] I. Garrido, S. Lagüela, P. Arias, J. Balado, Thermal-based analysis for the automatic detection and characterization of thermal bridges in buildings, Energy Build. 158 (2018) 1358–1367, http://dx.doi.org/10.1016/j.enbuild.2017.11.031.

[17] H. Macher, T. Landes, P. Grussenmeyer, Automation of thermal point clouds analysis for the extraction of windows and thermal bridges of building facades, in: The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume XLIII-B2-2020, Copernicus GmbH, 2020, pp. 287–292, http://dx.doi.org/10.5194/isprs-archives-XLIII-B2-2020-287-2020.

[18] J.R.M.-D. Dios, A. Ollero, Automatic detection of windows thermal heat losses in buildings using UAVs, in: 2006 World Automation Congress, 2006, pp. 1–6, http://dx.doi.org/10.1109/WAC.2006.375998.

[19] T. Rakha, A. Liberty, A. Gorodetsky, B. Kakillioglu, S. Velipasalar, Heat mapping drones: An autonomous computer-vision-based procedure for building envelope inspection using unmanned aerial systems (UAS), Technol.|Archit. + Des. 2 (2018) 30–44, http://dx.doi.org/10.1080/24751448.2018.1420963.

[20] S. Mirzabeigi, M. Razkenari, Automated vision-based building inspection using drone thermography, in: 20th Annual New York State Green Building Conference, American Society of Civil Engineers, 2022, pp. 737–746, http://dx.doi.org/10.1061/9780784483961.077.

[21] B. Barahona, R. Buck, O. Okaya, P. Schuetz, Detection of thermal anomalies on building façades using infrared thermography and supervised learning, J. Phys. Conf. Ser. 2042 (2021) 012013, http://dx.doi.org/10.1088/1742-6596/2042/1/012013.

[22] C. Kim, J.-S. Choi, H. Jang, E.-J. Kim, Automatic detection of linear thermal bridges from infrared thermal images using neural network, Appl. Sci. 11 (2021) 931, http://dx.doi.org/10.3390/app11030931.

[23] Z. Mayer, J. Kahn, Y. Hou, R. Volk, AI-based thermal bridge detection of building rooftops on district scale using aerial images, in: Jimmy Abualdenien, André Borrmann, Lucian-Constantin Ungureanu, Timo Hartmann (Eds.), EG-ICE 2021 Workshop on Intelligent Computing in Engineering, Universitätsverlag der TU Berlin, 2021, pp. 497–507, http://dx.doi.org/10.5445/IR/1000136256.

[24] Z. Mayer, Y. Hou, J. Kahn, T. Beiersdörfer, R. Volk, Thermal bridges on building rooftops - hyperspectral (RGB thermal height) drone images of Karlsruhe, Germany, with thermal bridge annotations, 2021, http://dx.doi.org/10.5281/zenodo.4767772.

[25] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, 2019, https://github.com/facebookresearch/detectron2, Accessed 2022-11-08.

[26] K. Chen, J. Wang, J. Pang, Y. Cao, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Xu, Z. Zhang, D. Cheng, C. Zhu, T. Cheng, Q. Zhao, B. Li, X. Lu, R. Zhu, Y. Wu, J. Dai, J. Wang, J. Shi, W. Ouyang, C.C. Loy, D. Lin, MMDetection: Open mmlab detection toolbox and benchmark, 2019, arXiv preprint arXiv:1906.07155.

[27] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, 2016, http://dx.doi.org/10.48550/ARXIV.1612.03144, URL https://arxiv.org/abs/1612.03144.

[28] K. He, G. Gkioxari, P. Dollár, R. Girshick, Mask R-CNN, in: 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980–2988, http://dx.doi.org/10.1109/ICCV.2017.322.

[29] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778, http://dx.doi.org/10.1109/CVPR.2016.90.

[30] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: 2021 IEEE/CVF International Conference on Computer Vision, 2021, pp. 9992–10002, http://dx.doi.org/10.1109/ICCV48922.2021.00986.

[31] Y. Li, Y. Chen, N. Wang, Z.-X. Zhang, Scale-aware trident networks for object detection, in: 2019 IEEE/CVF International Conference on Computer Vision, 2019, pp. 6053–6062, http://dx.doi.org/10.1109/ICCV.2019.00615.

[32] C. Zhu, Y. He, M. Savvides, Feature selective anchor-free module for single-shot object detection, in: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, IEEE Computer Society, 2019, pp. 840–849, http://dx.doi.org/10.1109/CVPR.2019.00093.

[33] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, F. Wei, B. Guo, Swin transformer V2: Scaling up capacity and resolution, in: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 11999–12009, http://dx.doi.org/10.1109/CVPR52688.2022.01170.

[34] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L.M. Ni, H.-Y. Shum, DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection, 2022, arXiv:2203.03605 [cs].

[35] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, C.L. Zitnick, Microsoft COCO: Common objects in context, in: D. Fleet, T. Pajdla, B. Schiele, T. Tuytelaars (Eds.), Computer Vision – European Conference on Computer Vision (ECCV) 2014, Springer International Publishing, Cham, 2014, pp. 740–755, http://dx.doi.org/10.1007/978-3-319-10602-1_48.

[36] S. Ren, K. He, R. Girshick, J. Sun, Faster R-CNN: Towards real-time object detection with region proposal networks, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2017) 1137–1149, http://dx.doi.org/10.1109/TPAMI.2016.2577031.

[37] D. Coquelin, C. Debus, M. Götz, F. von der Lehr, J. Kahn, M. Siggel, A. Streit, Accelerating neural network training with distributed asynchronous and selective optimization (DASO), J. Big Data 9 (2022) 14, http://dx.doi.org/10.1186/s40537-021-00556-1, arXiv:2104.05588.

[38] René Caspart, Sebastian Ziegler, Arvid Weyrauch, Holger Obermaier, Simon Raffeiner, Leon Pascal Schuhmacher, Jan Scholtyssek, Darya Trofimova, Marco Nolden, Ines Reinartz, Fabian Isensee, Markus Götz, Charlotte Debus, Precise Energy Consumption Measurements of Heterogeneous Artificial Intelligence Workloads, 2022, http://dx.doi.org/10.48550/arXiv.2212.01698, arXiv:2212.01698.

[39] IEEE Standard for Floating-Point Arithmetic, IEEE Std 754-2019 (Revision of IEEE 754-2008), IEEE Computer Society, 2019, pp. 1–84, http://dx.doi.org/10.1109/IEEESTD.2019.8766229.

[40] N. Fouad, T. Richter, Leitfaden Thermografie Im Bauwesen, Theorie, Anwendungsgebiete, Praktische Umsetzung, 2007.

[41] Z. Mayer, J. Heuer, R. Volk, F. Schultmann, Aerial thermographic image-based assessment of thermal bridges using representative classifications and calculations, Energies 14 (2021) 7360, http://dx.doi.org/10.3390/en14217360.

[42] Y. Hou, R. Volk, M. Chen, 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, Autom. Constr. 124 (2021) 103554, http://dx.doi.org/10.1016/j.autcon.2021.103554.

[43] L. Yang, Y. Fan, N. Xu, Video instance segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 5188–5197, URL https://openaccess.thecvf.com/content_ICCV_2019/html/Yang_Video_Instance_Segmentation_ICCV_2019_paper.html, accessed 2022-11-08.

[44] D. Du, Y. Qi, H. Yu, Y. Yang, K. Duan, G. Li, W. Zhang, Q. Huang, Q. Tian, The unmanned aerial vehicle benchmark: Object detection and tracking, in: V. Ferrari, M. Hebert, C. Sminchisescu, Y. Weiss (Eds.), Computer Vision – European Conference on Computer Vision (ECCV) 2018, in: Lecture Notes in Computer Science, Springer International Publishing, Cham, 2018, pp. 375–391, http://dx.doi.org/10.1007/978-3-030-01249-6_23.

[45] S.W. Zamir, A. Arora, A. Gupta, S. Khan, G. Sun, F.S. Khan, F.Z. 0001, L.S. 0001, G.-S. Xia, X. Bai, iSAID: A large-scale dataset for instance segmentation in aerial images, in: IEEE Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 28–37, URL http://openaccess.thecvf.com/content_CVPRW_2019/html/DOAI/Zamir_iSAID_A_Large-scale_Dataset_for_Instance_Segmentation_in_Aerial_Images_CVPRW_2019_paper.html, accessed 2022-11-08.

[46] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186, http://dx.doi.org/10.18653/v1/N19-1423.