Thermal-SAM: Adversarial Prompt-Based Unsupervised Building Segmentation in Thermal Aerial Imagery – A Case Study in Turin

,

Shuai Niu [1 2] **Hongshan Guo** [1] **Maria Ferrara** [3] **Sebastiano Anselmo** [4]

Abstract

Thermal image building segmentation is essential for monitoring energy consumption and supporting environmental protection. Current segmentation methods are predominantly designed for RGB images, posing challenges for thermal images, especially when segmenting buildings of varied shapes from aerial views, due to their lower resolution, lack of detailed features, and channel differences. To address these challenges, we propose an unsupervised segmentation method Thermal-SAM, specifically for a new aerial thermal dataset from Turin, Italy. We enhance this method by incorporating color aerial images from the same region as an auxiliary modality to generate pseudo labels for unsupervised training. Our approach introduces an adversarial prompt-based pseudo-label generation method, utilizing several vision-language models, along with positive and negative prompts. Extensive experiments demonstrate that Thermal-SAM, surpasses state-of-the-art methods by more than 10%.

1. Introduction

Global warming intensifies environmental challenges such as hurricanes and floods. Reducing energy consumption is crucial to mitigating these effects (Rogelj et al., 2013). Buildings account for 40% of global energy use and emissions due to construction, heating, cooling, etc, (Nejat et al., 2015). Monitoring building temperatures with Unmanned Aerial Vehicle (UAV) infrared cameras is a promising approach to estimate energy use intensity (EUI) (Ham & Golparvar-Fard, 2013). Thus, efficient building segmentation is vital. However, accurately segmenting buildings in UAV-captured thermal images is difficult since current segmentation methods are mainly designed for RGB images (Kirillov et al., 2023; Chen et al., 2020).

Figure 1. Images captured by aircraft in both spectrum.

Visible cameras capture light in the RGB spectrum, providing detailed object information. In contrast, thermal images, captured by infrared cameras, consist of a single channel representing infrared intensity (Gade & Moeslund, 2014). Detailed visual data is crucial for object segmentation to classify objects and identify pixel-level edges, making RGB images more suitable for these tasks (He et al., 2017; Kirillov et al., 2023). Figure 1 shows the difference: buildings and cars are easily segmented in visible images, but difficult to discern in infrared images due to the lack of color information. The lower intensity of infrared radiation limits thermal image detail and resolution, posing challenges for accurate building segmentation and precise EUI prediction.

Current challenges for thermal image segmentation include a lack of thermal-based datasets and well-pretrained segmen-

[1]Department of Architecture, University of Hong Kong, Hong Kong, China [2]Department of Computer Science, Hong Kong Baptist University, Hong Kong, China [3]Department of Energy, Polytechnic Institute of Turin, Turin, Italy [4]DIST, Polytechnic Institute of Turin, Turin, Italy. Correspondence to: Hongshan Guo <hongshan@hku.hk>.

Accepted at ICML 2025 CO-BUILD Workshop on Computational Optimization of Buildings., Vancouver, Canada. PMLR 267, 2025.
Copyright 2025 by the author(s).

tation models. Most existing image segmentation datasets focus on RGB images from visible cameras (Zhou et al., 2017; Lin et al., 2014), leading to the development of powerful pretrained encoders for semantic and instance segmentation across various categories (Strudel et al., 2021; Kirillov et al., 2023). In contrast, thermal image segmentation mainly targets pedestrians (Wang & Bai, 2019; Altay & Velipasalar, 2022) and vehicles (Yang & Park, 2015; Masouleh & Shah-Hosseini, 2019), as these are easier to annotate. Segmenting buildings in thermal images presents challenges, including low resolution, variable shapes, and unclear boundaries, resulting in a lack of annotated datasets and well-pretrained models. Therefore, these challenges highlight the need for innovative, unsupervised segmentation methods tailored to this domain.

Based on this intuition, we propose Thermal-SAM, an unsupervised building thermal image segmentation model using a new constructed aerial thermal image dataset from Turin city. This model leverages adversarial prompts to enhance segmentation. This work focuses on *Step 1*: extracting pixel-accurate thermal building footprints from mid-wave IR mosaics. Step 2—using those masks as inputs to predict building-level energy-use intensity (EUI)—is treated in a separate manuscript, currently under review. By isolating the segmentation stage here, we provide a stand-alone, reproducible baseline that downstream energy models can directly adopt.

Contribution scope. This paper tackles *Step 1* in a two-stage pipeline: extracting pixel-accurate thermal building footprints from mid-wave IR mosaics. Step 2—linking those masks to building-level EUI—is covered in a companion manuscript now under review. Downstream energy models can directly adopt the isolated segmentation results.

Main contributions.

- **Thermal-SAM.** First unsupervised thermal-image model that segments buildings without human labels.
- **Vision–language synergy.** Hierarchical captioning supplies robust semantic cues for pseudo-labels.
- **Adversarial prompt generation.** A novel prompt scheme expands SAM's capabilities to distorted IR imagery and boosts IoU by +10 pp over strong baselines.

2. Related Works

The field of unsupervised image segmentation has advanced significantly, addressing the challenges posed by reliance on human labeling. Notable methods include CutLER (Wang et al., 2023), which introduces MaskCut, leveraging self-supervised learning with Vision Transformers (Dosovitskiy, 2020) and DINO (Caron et al., 2021) for class-agnostic object segmentation. Similarly, STEGO (Hamilton et al.) employs clustering on DINO-extracted features for semantic segmentation, while U2Seg (Niu et al., 2024) bridges semantic and instance segmentation using pseudo-labels from MaskCut, DINO, and clustering. Despite these advancements, extending such unsupervised methods to building segmentation in thermal images remains largely unexplored and challenging. This is primarily due to the scarcity of high-quality, large-scale datasets of aerial-view thermal building images, as existing ones predominantly focus on pedestrians and vehicles (Liu et al., 2018; Li et al., 2020).

3. Methodology

3.1. Overview

As shown in Figure 2, Thermal-SAM adopt SAM, OneFormer, CLIP-Seg, and Sentence Transformers to achieve robust unsupervised thermal building segmentation (descriptions are illustrated in Appendix A).

3.2. Zero-Shot Panoptic Segmentation

We first present our Thermal-SAM for panoptic segmentation of color imagery. Given an UAV image in visible view (color) $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$ and a paired infrared view (thermal) $\boldsymbol{T} \in \mathbb{R}^{H\times W}$, we map these two image by rotation $\mathcal{T}$ and shift $\boldsymbol{\Delta}$:

$$\boldsymbol{I}' = \mathcal{T}(\boldsymbol{I}) + \boldsymbol{\Delta}_{\boldsymbol{I}}, \tag{1}$$

where $\boldsymbol{I}'$ is the transformed color image. Subsequently, we crop and partition the color image and thermal image into smaller patches $\boldsymbol{P}_N^{(I)}$ and $\boldsymbol{P}_N^{(T)}$, respectively, where N is number of patches, each of which retaining only valid regions unobscured by black pixels in the thermal image.

Next, we pass each visible view patch $\boldsymbol{P}_N^{(I)}$ into SAM, which outputs an instance mask and a set of segmented instance images:

$$\boldsymbol{M}_N^{(I)}; \boldsymbol{F}^{(I)} = \text{SAM}\Big(\boldsymbol{P}_N^{(I)}\Big), \tag{2}$$

where $\boldsymbol{M}_N^{(I)}$ denotes the instance mask, and $\boldsymbol{F}^{(I)} = \{\boldsymbol{f}_1^I, .., \boldsymbol{f}_n^I\}$ represents the set of segmented instances extracted from $\boldsymbol{P}_{i,j}^{(I)}$ (with n being the total number of instances).

Then, for each segmented instance $\boldsymbol{f}_n^I$, we generate a coarse-grained image label $\boldsymbol{C}_n^{(I)(O)}$ using OneFormer and fine-grained image labels $\boldsymbol{C}_n^{(I)(B)}$ using BLIP2:

$$\boldsymbol{C}_n^{(I)(B)}, \boldsymbol{C}_n^{(I)(O)} = \text{BLIP2}\big(\boldsymbol{f}_n^I\big), \text{OneFormer}\big(\boldsymbol{f}_n^I\big). \tag{3}$$

The fine-grained labels include more details of the instance description. Finally, we apply a Clip-Seg model to generate the final accurate labels:

$$\boldsymbol{C}_n^{(I)} = \text{argmax}\big(\text{Clip-Seg}\big(\boldsymbol{f}_n^I, [\boldsymbol{C}_n^{(I)(O)}, \boldsymbol{C}_n^{(I)(B)}]\big)\big). \tag{4}$$

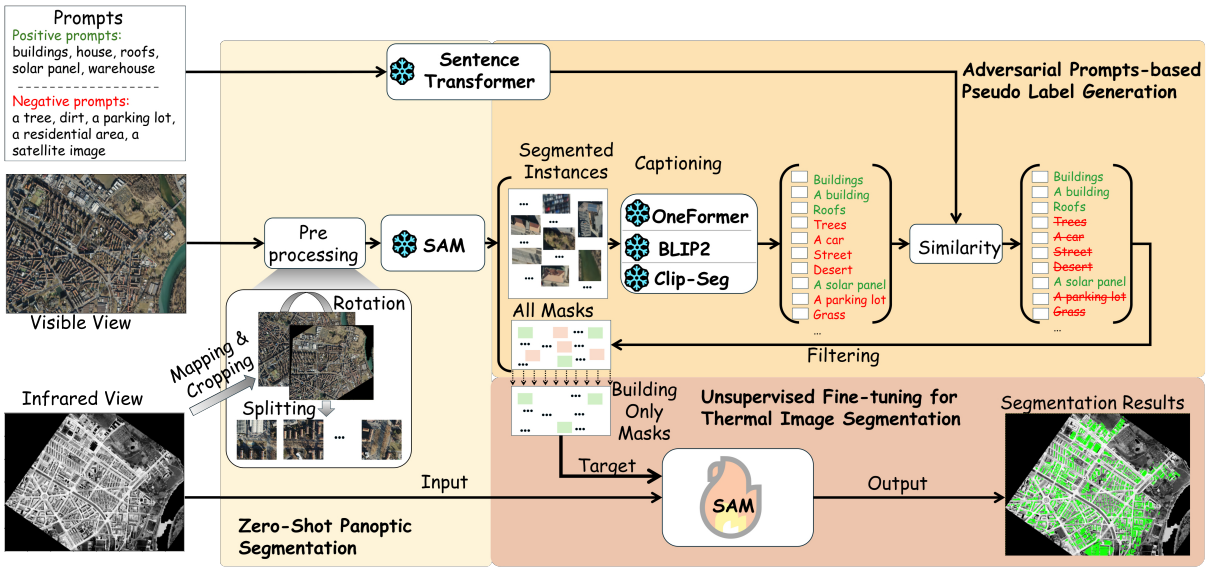

Figure 2. Our unsupervised thermal image segmentation framework, Thermal-SAM, includes three modules: Zero-Shot Panoptic Segmentation, Adversarial Prompts-based Pseudo Label Generation, and Unsupervised Fine-tuning for Thermal Image Segmentation.

The final step aggregates the labels of all instances in a patch:

$$\boldsymbol{C}_N^{(I)} = \{\boldsymbol{C}_1^{(I)}, .., \boldsymbol{C}_n^{(I)}\}. \tag{5}$$

3.3. Adversarial Prompts-based Pseudo Label Generation

Given that Thermal-SAM is designed to segment all buildings from thermal images, we employ an adversarial prompts-based pseudo-label generation approach to select only building-related instances in the color aerial image.

We first compile a list of building-related words to serve as positive prompts $\boldsymbol{S}^+$ and a set of words that are semantically similar but do not actually refer to buildings as negative prompts $\boldsymbol{S}^-$. Then, we apply a Sentence Transformer to filter out building-irrelevant labels from the image labels $\boldsymbol{C}^{(I)}$ using cosine similarity cos:

$$\begin{aligned}\boldsymbol{Y}_N^{(I)} =& \{\boldsymbol{y}_k^{(I)} \mid \cos(\boldsymbol{C}_n^{(I)}, \boldsymbol{S}^+) \geq \tau^+\} \\ & \cap \{\boldsymbol{y}_k^{(I)} \mid \cos(\boldsymbol{C}_n^{(I)}, \boldsymbol{S}^+) < \tau^-\},\end{aligned} \tag{6}$$

where $\boldsymbol{Y}_N^{(I)} = \{\boldsymbol{y}_1^{(I)}, .., \boldsymbol{y}_k^{(I)}\}$, with $k < n$, τ^+ is the threshold for selecting building-relevant labels, and τ^- is the threshold for filtering out building-irrelevant labels.

Finally, we extract the building-related masks as segmentation labels for the next-step fine-tuning:

$$\boldsymbol{M}_N^{(I)*} = \{\boldsymbol{m}_k^{(I)} \mid \boldsymbol{y}_k^{(I)} \in \boldsymbol{Y}_N^{(I)}\} \text{ and } \boldsymbol{m}_k^{(I)} \in \boldsymbol{M}_N^{(I)}. \tag{7}$$

3.4. Unsupervised Fine-tuning for Thermal Image Segmentation

The building-only mask by using Eq (7) does not perfectly align with thermal images due to distortions between visible and infrared views. To address this, we use the masks $\boldsymbol{M}_N^{(I)*}$ as pseudo labels to fine-tune SAM over a few epochs, enhancing SAM's ability to segment thermal images without being compromised by the inaccuracies of the labels:

$$\boldsymbol{M}_N^{(T)*} = \text{SAM}(\boldsymbol{M}_N^{(I)*}, \boldsymbol{P}_N^{(T)}; \theta), \tag{8}$$

where θ is the tunable parameter. Finally, we manually select the best output between $\boldsymbol{M}_N^{(I)*}$ and $\boldsymbol{M}_N^{(T)*}$. Refer to Appendix A.1 for the detailed Thermal-SAM algorithm.

4. Experiments

4.1. Dataset Description

In this study, we captured aerial imagery of Turin, Italy, on March 23, 2023. Thermal images were captured by FLIR A8581 MWIR HD camera, concurrently, color aerial images were captured from the same vantage (More description and data preprocessing please refer to Appendix A.2 and A.3).

4.2. Baseline Models

We benchmark our approach against several methods: SAM, a semantic segmentation model that does not assign labels to individual segmented objects; MaskFormer (Cheng et al., 2021) and OneFormer, both panoptic segmentation models from which we retain only building-labeled masks. Addi-

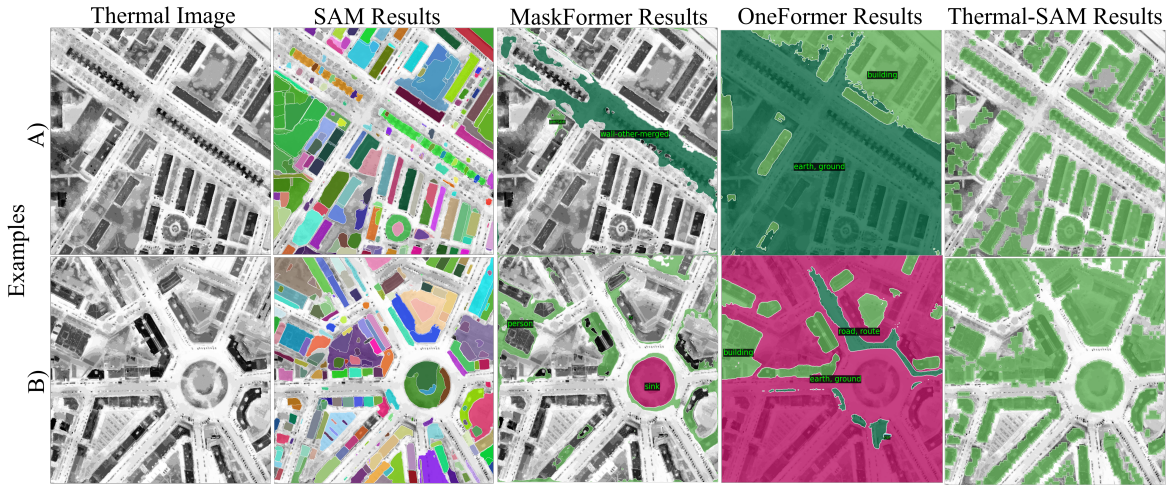

Figure 3. Qualitative Comparison of Building Segmentation: Thermal-SAM vs. Baseline Models

tionally, we evaluate Thermal-SAM, without fine-tuning on pseudo labels (Thermal-SAM w/o FT). Reported results are averaged over five runs with different random seeds. Implementation details please refer to Appendix A.5.

Table 1. Evaluation Metrics for Comparative Models

Model	IoU	Dice Coefficient	Precision	Recall	F1 Score
SAM	0.0003	0.0006	0.0003	0.1107	0.0006
MaskFormer	0.0981	0.1787	0.1821	0.1754	0.1787
OneFormer	0.1862	0.3140	0.6733	0.2047	0.3140
Thermal-SAM	0.2873	0.4463	0.4421	0.4506	0.4463
w/o FT	0.2659	0.4201	0.3544	0.5157	0.4201

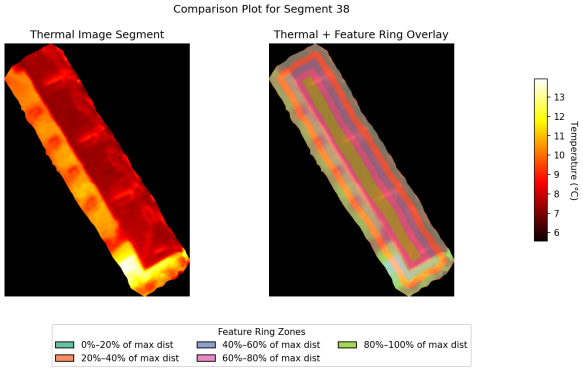

Figure 4. Illustration of a segmented patch (left) and thermal information used for EUI prediction (right).

5. Results

We conduct both quantitative and qualitative comparisons with the state-of-the-art baselines and also comparison with Microsoft-footprint in Appendix A.4 and A.6.

5.1. Quantitative Comparison of Segmentation Results

We evaluate Thermal-SAM against SAM, MaskFormer, and OneFormer for building segmentation in Turinese thermal images using standard metrics (Table 1). SAM's poor performance highlighted the need for supervision. MaskFormer and OneFormer improved by over 17% with building-related label selection; OneFormer nearly doubled MaskFormer's scores, benefiting from its task-conditioning and multi-task architecture. In contrast, our Thermal-SAM, after pseudo-label generation and fine-tuning with SAM-base, accurately detects and segments individual building-related objects in thermal images.

5.2. Qualitative Comparison of Segmentation Results

Thermal segmentation. Figure 3 qualitatively compares segmentation by Thermal-SAM against baselines. SAM fails to identify individual instances, only grouping broad classes. While MaskFormer and OneFormer preserve building-related masks, MaskFormer lacks precision while OneFormer lacks reval, which is consistent with quantitative results. In contrast, Thermal-SAM, after pseudo-labeling and fine-tuning SAM-base, accurately segments individual building-related objects in thermal images.

Energy proxy. Figure 4 presents a segmented thermal building image patch (left), used to examine how rooftop surface temperature variations (core vs. periphery, via ring circles) correlate with EUI prediction (right). Using these masks, the mean roof–ambient temperature difference (ΔT_{roof}) for 857 buildings explains $R^2 = 0.46$ of the vari-

ance in measured EPC-based EUIs. Our future work, by adding five additional mask-derived features, increases this explanatory power to $R^2 = 0.78$, underscoring the proposed segmentation's downstream value.

6. Conclusion and Future Works

We introduce Thermal-SAM, a novel unsupervised thermal building segmentation method, specifically tailored for Turin. Our approach combines zero-shot panoptic segmentation of color images with adversarial prompt-based pseudo-label generation to extract building-related objects. These labels then fine-tune SAM for thermal imagery, correcting distortions between visible and infrared data. Evaluations show Thermal-SAM outperforms all baselines and provides more robust segmentation than Microsoft Footprints. Furthermore, our method provides a generalizable framework for segmenting thermal UAV imagery to support building-level energy approximation. In future work, we aim to extend our method to additional object categories, such as vehicles and Vegetation (Fire prevention), and apply it to broader geographic areas beyond Turin.

Impact Statement

This paper presents work whose goal is to advance the field of energy consumption and machine learning. The dataset will not be opened to the public due to the usage policy.

References

Altay, F. and Velipasalar, S. The use of thermal cameras for pedestrian detection. *IEEE Sensors Journal*, 22(12): 11489–11498, 2022.

Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., and Joulin, A. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9650–9660, 2021.

Chen, H., Sun, K., Tian, Z., Shen, C., Huang, Y., and Yan, Y. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8573–8581, 2020.

Cheng, B., Schwing, A., and Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *Advances in neural information processing systems*, 34: 17864–17875, 2021.

Devlin, J. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Dosovitskiy, A. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Gade, R. and Moeslund, T. B. Thermal cameras and applications: a survey. *Machine vision and applications*, 25: 245–262, 2014.

Ham, Y. and Golparvar-Fard, M. An automated vision-based method for rapid 3d energy performance modeling of existing buildings using thermal and digital imagery. *Advanced Engineering Informatics*, 27(3):395–409, 2013.

Hamilton, M., Zhang, Z., Hariharan, B., Snavely, N., and Freeman, W. T. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*.

He, K., Gkioxari, G., Dollár, P., and Girshick, R. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pp. 2961–2969, 2017.

Jain, J., Li, J., Chiu, M. T., Hassani, A., Orlov, N., and Shi, H. Oneformer: One transformer to rule universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2989–2998, 2023.

Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A. C., Lo, W.-Y., et al. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4015–4026, 2023.

Li, C., Xia, W., Yan, Y., Luo, B., and Tang, J. Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 32(7): 3069–3082, 2020.

Li, J., Li, D., Savarese, S., and Hoi, S. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pp. 19730–19742. PMLR, 2023.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, X., Yang, T., and Li, J. Real-time ground vehicle detection in aerial infrared imagery based on convolutional neural network. *Electronics*, 7(6):78, 2018.

Lüddecke, T. and Ecker, A. Image segmentation using text and image prompts. In *Proceedings of the IEEE/CVF*

conference on computer vision and pattern recognition, pp. 7086–7096, 2022.

Masouleh, M. K. and Shah-Hosseini, R. Development and evaluation of a deep learning model for real-time ground vehicle semantic segmentation from uav-based thermal infrared imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 155:172–186, 2019.

Nejat, P., Jomehzadeh, F., Taheri, M. M., Gohari, M., and Majid, M. Z. A. A global review of energy consumption, co2 emissions and policy in the residential sector (with an overview of the top ten co2 emitting countries). *Renewable and sustainable energy reviews*, 43:843–862, 2015.

Niu, D., Wang, X., Han, X., Lian, L., Herzig, R., and Darrell, T. Unsupervised universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22744–22754, 2024.

Reimers, N. and Gurevych, I. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019. URL http://arxiv.org/abs/1908.10084.

Rogelj, J., McCollum, D. L., O'Neill, B. C., and Riahi, K. 2020 emissions levels required to limit warming to below 2° c. *Nature Climate Change*, 3(4):405–412, 2013.

Strudel, R., Garcia, R., Laptev, I., and Schmid, C. Segmenter: Transformer for semantic segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 7262–7272, 2021.

Wang, P. and Bai, X. Thermal infrared pedestrian segmentation based on conditional gan. *IEEE transactions on image processing*, 28(12):6007–6021, 2019.

Wang, X., Girdhar, R., Yu, S. X., and Misra, I. Cut and learn for unsupervised object detection and instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 3124–3134, 2023.

Yang, D. and Park, H. A new shape feature for vehicle classification in thermal video sequences. *IEEE Transactions on Circuits and Systems for Video Technology*, 26 (7):1363–1375, 2015.

Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., and Torralba, A. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 633–641, 2017.

A. Appendix

A.1. Preliminaries for Large-Scale Models

SAM (Kirillov et al., 2023) is a foundation model for image segmentation, which includes an image encoder, a flexible prompt encoder, and a fast mask decoder. SAM accepts both sparse (points, boxes, text) and dense (masks) inputs and is capable of segmenting any objects without labels in an image with the given prompts.

OneFormer (Jain et al., 2023) is a unified image segmentation model for instance segmentation, semantic segmentation, and panoptic segmentation. By using different task prompts and contrastive learning, OneFormer achieved state-of-the-art evaluation performance on diverse datasets.

BLIP2 (Li et al., 2023) is a versatile and efficient pretraining strategy for vision–language understanding that leverages a lightweight query transformer. This transformer encodes visual features into image prompts that are aligned with the language encoding space, enabling Large Language Models (LLMs) to more effectively process and understand multimodal inputs.

CLIP-Seg (Lüddecke & Ecker, 2022) builds a segmentation decoder based on CLIP, which accepts both image and text prompts for guiding query image segmentation. It can also be used to filter out the most appropriate classes from candidate prompt lists for the input query image.

Sentence Transformers (Reimers & Gurevych, 2019) is built based on BERT (Devlin, 2018), while modifying the structure of BERT. It uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity.

The implementation of SAM, OneFormer, BLIP2, and CLIP-Seg is based on Semantic-SAM[1].

Algorithm 1 Thermal-SAM Framework for Thermal Image Building Segmentation

Require: UAV color image $\boldsymbol{I} \in \mathbb{R}^{H\times W\times 3}$, thermal image $\boldsymbol{T} \in \mathbb{R}^{H\times W}$, patch size $h \times w$, thresholds τ^+, τ^-, trainable parameters θ

Ensure: Final building segmentation mask $\boldsymbol{M}^{(T)}$

Alignment: Obtain the aligned color image $\boldsymbol{I}'$ using Eq. (1).
Partitioning: Split $\boldsymbol{I}'$ and $\boldsymbol{T}'$ into patches $\boldsymbol{P}_N^{(I)}$ and $\boldsymbol{P}_N^{(T)}$.
for each image patch $\boldsymbol{P}_N^{(I)}$ **do**
  Segmentation: Run SAM on $\boldsymbol{P}_N^{(I)}$ to obtain the instance mask $\boldsymbol{M}_N^{(I)}$ and instances $\boldsymbol{F}^{(I)}$ (Eq. (2)).
  for each instance $\boldsymbol{f}_n^I \in \boldsymbol{F}^{(I)}$ **do**
    Generate pseudo label $\boldsymbol{C}_n^{(I)}$ for the instance (Eqs. (3) and (4)).
  end for
  Aggregate instance labels to form $\boldsymbol{C}_N^{(I)}$ (Eq. (5)).
end for
Pseudo-label Generation: Filter $\boldsymbol{C}^{(I)}$ using adversarial prompts $\boldsymbol{S}^+$, $\boldsymbol{S}^-$,thresholds τ^+ and τ^-, to obtain $\boldsymbol{Y}_N^{(I)}$ (Eq. (6)).
Extract building-related masks $\boldsymbol{M}_N^{(I)*}$ based on $\boldsymbol{Y}_N^{(I)}$ (Eq. (7)).
SAM Fine Tuning:
while not converge **do**
  for mini-batch B **do**
    Fine-tune SAM to generate the building mask
    $\boldsymbol{M}_N^{(T)*}$ using Eq. (8).
  end for
end while
for each thermal image patch $\boldsymbol{P}_N^{(T)}$ **do**
  Manually select the best mask between $\boldsymbol{M}_N^{(T)}$ and $\boldsymbol{M}_N^{(T)*}$.
end for
return $\boldsymbol{M}^{(T)}$

[1]https://github.com/fudan-zvg/Semantic-Segment-Anything/tree/main

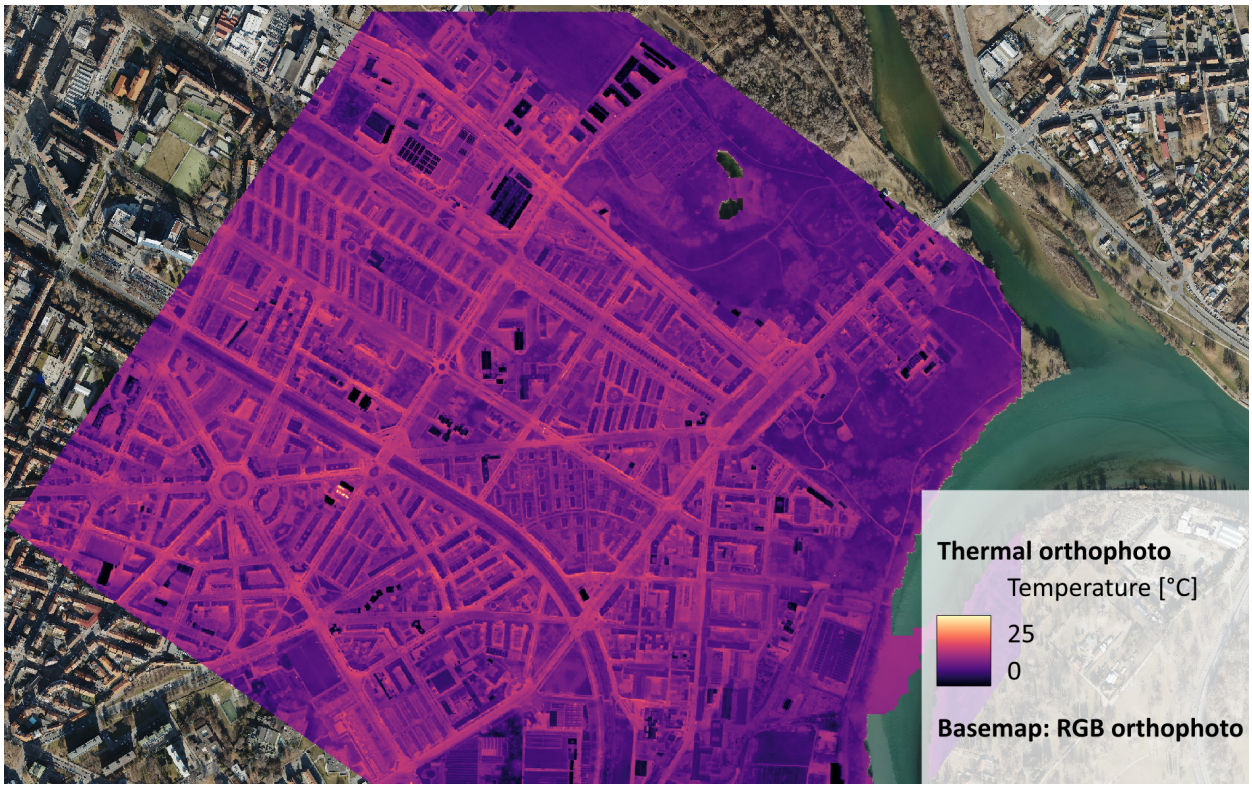

Figure 5. Turin Aerial Imagery

A.2. Infrared View and Visible View

We use a UAV with a thermal camera to capture the dataset. During the UAV flight, skies were mostly clear with an ambient temperature around 20 C, ensuring stable conditions for midwave infrared capture despite possible variability from clouds or wind. Thermal images were captured by a FLIR A8581 MWIR HD camera (equipped in a UAV), shortly after sunset, to minimize biases from solar heating, flying at or above 300m in a nadir orientation as legally mandated. This altitude was chosen to balance coverage of our $2km^2$ study region with achieving a fine ground sampling distance (20cm) for distinguishing building rooftops. We also verified the camera's $\pm$ 1 C accuracy by referencing a known temperature target on the ground before and after the flight. Midwave IR generally provides a higher spatial resolution than longwave IR, resulting in a ground sampling distance of approximately 20cm—relatively fine for thermal imaging. However, due to the lower photon energy compared to the visible spectrum (0.4–0.7 μm), a larger instantaneous field of view is required to capture sufficient radiation, which can make thermal images coarser or noisier than their RGB counterparts. Nevertheless, we produced a final orthomosaic of roughly 7,814 $\times$ 6,000 pixels, as shown in Figure 5. Concurrently, color aerial images were captured from the same vantage, offering a higher resolution of 39,070 × 30,000 pixels, as shown by the purple overlay in Figure 5. Although these RGB images exhibit sharper detail due to their shorter wavelengths, combining them with midwave IR data allows us to capture both temperature patterns and fine-grained scene structure—a key advantage for unsupervised segmentation tasks.

A.3. Data Preprocessing

Both the infrared and visible views underwent an orthorectification process to correct for perspective distortions and terrain effects, ensuring that each pixel is accurately aligned with real-world coordinates in WGS84. We used standard aerial triangulation and ground control points to refine the mosaics, then performed a global shift-rotation alignment of the IR mosaic to match the RGB mosaic, ensuring that patches cover the same region. Although this alignment simplifies subsequent patching and segmentation, it cannot entirely eliminate the minor distortions between corresponding buildings in the two views.

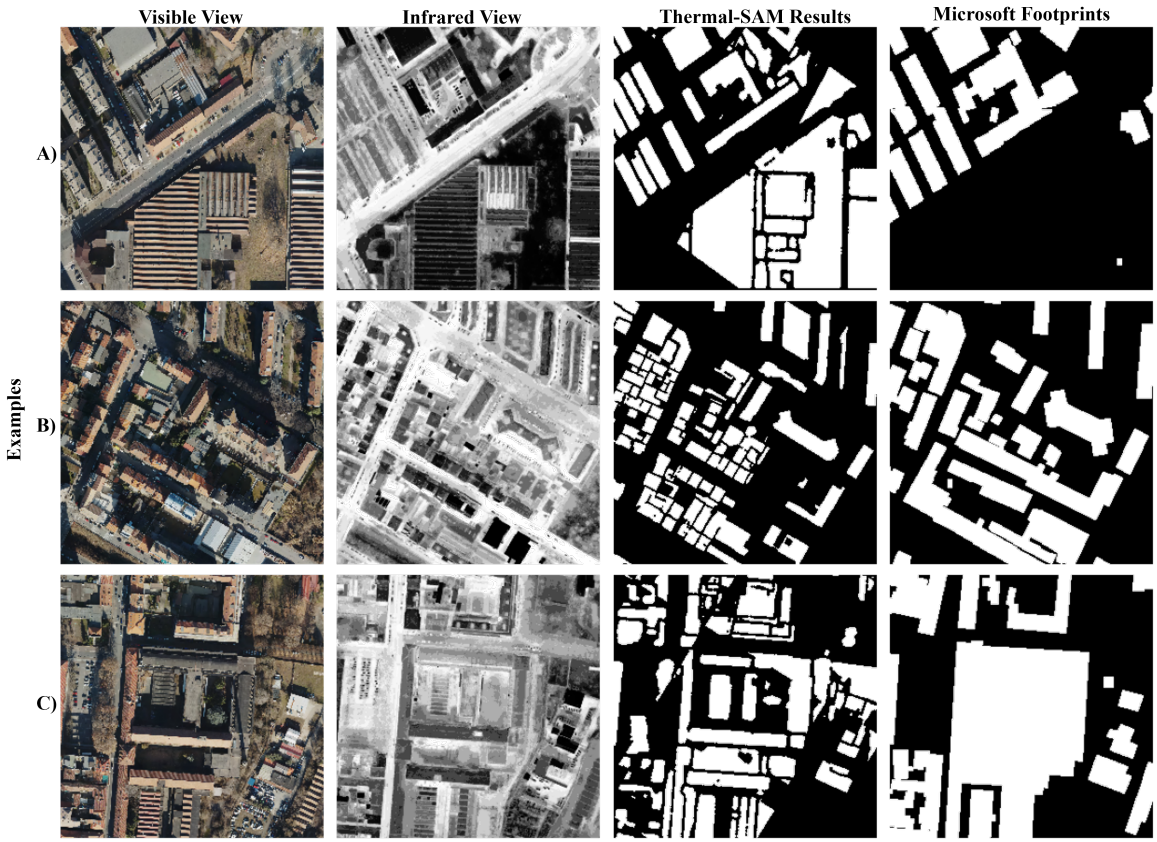

Figure 6. Qualitative Comparison of Building Segmentation: Thermal-SAM vs. Microsoft Footprints

A.4. Microsoft Footprints

Microsoft Footprint (MF) [2] is a global dataset released by Bing Maps that comprises 1.4 billion building footprints extracted from imagery captured between 2014 and 2024. The building labels are generated by deep neural networks (DNNs) trained for semantic segmentation, which detect building pixels in color aerial images and convert these detections into polygonal representations. Although these labels may not be perfectly accurate due to the absence of human annotation, we still use the subset of Footprint data covering Turin, Italy, as an evaluation indicator for our quantitative evaluation.

A.5. Implementation Details

In our experiments, we utilized the PyTorch framework (version 2.0.1) within a CUDA 11.7 environment. We employed the Adam optimizer with an initial learning rate of $1e^{-5}$, and a scheduled learning rate adjustment. The experiments were conducted on high-performance NVIDIA Tesla V100 GPUs. For pseudo-label generation, we used SAM-huge, OneFormer-large, and Clip-Seg-refined. For fine-tuning, we used SAM-base.

A.6. Qualitative Comparison of MF

In Section 4.3, although we employed Microsoft Footprints as an evaluation indicator to compute metrics, it is important to note that these footprints do not represent the true mask labels. Therefore, in this section, we present three examples (A, B, and C) to visually compare our results with Microsoft Footprints. The first two columns display the visible and infrared views of three aerial images.

[2] https://github.com/microsoft/GlobalMLBuildingFootprints

When comparing the segmentation results generated by our Thermal-SAM model with those of Microsoft Footprints, our approach demonstrates higher precision by detecting more complete building structures and capturing finer details, particularly evident in Example A. In this example, a large area in the lower portion of the visible view is easily misclassified as farmland from visible view; Microsoft Footprints do not include labels for this region, whereas our Thermal-SAM successfully segments the buildings using only the infrared view.

Furthermore, in Examples B and C, Microsoft Footprints assigns a single, large mask to high-density building areas, failing to distinguish individual small buildings. In contrast, our model effectively segments each small building, with Example B being particularly noteworthy. These results underscore our model's effectiveness in achieving accurate building segmentation in thermal imagery.