ThermalAssist: Towards Efficient Annotation of Thermal Imagery

Jingwei Zhu[1,2], Manoj Biswanath[1,3], Benjamin Busam[1,3]

[1] Chair of Photogrammetry and Remote Sensing, Technical University of Munich, Germany - (jingwei.zhu, manoj.biswanath, b.busam)@tum.de
[2] School of Geospatial Artificial Intelligence, East China Normal University, China - jwzhu@geoai.ecnu.edu.cn
[3] Munich Center for Machine Learning (MCML), Munich, Germany

Keywords: Thermal Infrared images, annotations, Thermal Bridge

Abstract

Thermal infrared (TIR) imaging provides a unique capability to reveal surface heat-transfer patterns of buildings and supports applications such as energy leakage detection, insulation inspection, and building energy monitoring. However, large-scale TIR image analysis by deep learning is still constrained by the lack of reliable annotations, as TIR images often exhibit blurred textures and weak boundaries, which makes manual labeling inconsistent and time-consuming. To address this challenge, we propose **ThermalAssist**, a geometry-aware and gradient-enhanced framework designed to assist thermal anomaly labeling in UAV-based TIR images. By combining sparse labeling, dense correspondence, and label propagation, the framework efficiently transfers labels across overlapping views while maintaining geometric and semantic consistency. Experiments on the TBBR (Thermal Bridges on Building Rooftops) dataset demonstrate that ThermalAssist achieves an F1-score of up to **0.87** and a mean IoU of **0.69**, effectively helping reduce missing annotations and inconsistent boundaries. Compared with the tracking-based SAMURAI method, our approach shows greater robustness under low-texture and low-overlap conditions. This work establishes a foundation for a broader thermal annotation system and validates an important step toward scalable, reliable, and more intelligent labeling of thermal anomaly imagery.

1. Introduction

Buildings account for nearly 40% of global energy consumption and approximately one-third of greenhouse gas emissions (Delmastro et al., 2022). As nations pursue carbon neutrality, improving the energy efficiency of buildings has become an urgent goal. Thermal Infrared (TIR) images capture the surface temperature of objects by sensing long-wavelength radiation, thereby revealing heat-transfer patterns that are invisible to the human eye. This unique capability makes TIR imaging indispensable for a range of applications, including energy leakage detection, insulation defect identification, environmental condition assessment, and cultural heritage preservation.

Despite its great potential, analyzing TIR images at scale remains highly challenging. Unlike visible-spectrum images that resemble human visual perception, TIR images are typically single-channel with blurred and low-texture features, making them inherently difficult to interpret. Effective analysis often requires professional knowledge and a familiarity of the target environment. Moreover, due to the limited field of view of thermal cameras, large-area analysis demands processing a substantial number of images, further increasing the complexity and cost.

With the rapid advancement of deep learning, automated interpretation of image data has achieved remarkable success in tasks such as object detection, semantic segmentation, and object tracking within the visual domain. However, the performance of such models critically depends on the availability of large, well-annotated datasets. Creating such datasets, particularly for TIR images, is labor-intensive and time-consuming. Existing datasets primarily focus on autonomous driving under low-light conditions, where TIR data are used jointly with RGB imagery. Consequently, publicly available TIR datasets remain limited in both task and quality, especially for thermal- or energy-related applications.

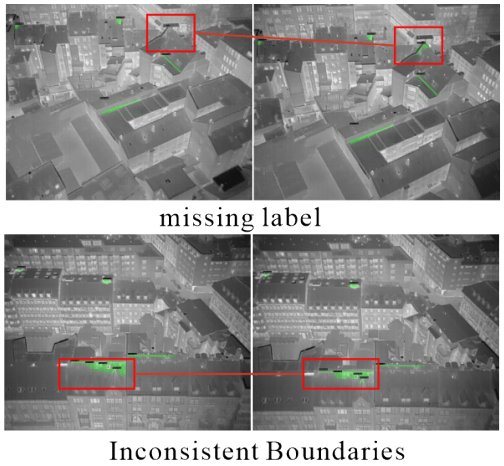

Figure 1. Problem of manual label thermal anomalies in TIR images with missing label and inconsistent boundaries.

Annotating TIR images poses challenges compared to RGB images. Thermal anomalies like thermal bridges usually lack clear geometric boundaries or distinct textures, and accurate labeling requires a thorough understanding of both thermal context and material properties. As shown in Figure 1, the difficulty is amplified in TIR pairs, where annotations must remain temporally consistent across consecutive frames, ensuring that (1) identical objects are labeled consistently to avoid missing labels and (2) boundaries remain spatially coherent. Due to subjective judgments and human errors, maintaining such consistency manually is both tedious and unreliable.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-53-2026

To address these challenges, this study proposes a novel annotation-assistance approach, ThermalAssist, for TIR images based on image-processing label propagation. We develop a prototype system that enables (1) automatic detection and tracking of identical objects across frames, and (2) rapid label transfer between temporally adjacent TIR images. The proposed approach can accelerate manual labeling, enhance inter-frame consistency, and support post-annotation quality control. In this work, we validate the feasibility of using flow-based label propagation to transfer sparse labels across overlapping TIR images for thermal annotation generation, facilitating future research in energy diagnostics and thermal anomaly detection.

The main contributions of this work are:

- We propose ThermalAssist, a geometry-aware and gradient-enhanced framework for UAV-based thermal anomaly annotation, which utilizes gradient-guided inputs to overcome the inherent challenges of weak boundaries in TIR imagery.
- We introduce an adaptive, overlap-aware label propagation strategy that evaluates inter-frame similarity via gradient-based matching, robustly handling varying overlap conditions and mitigating label drift.
- We validate our approach on the TBBR dataset, demonstrating that ThermalAssist achieves superior spatial consistency (F1-score up to 0.87) compared to state-of-the-art tracking methods (e.g., SAMURAI) under low-texture and low-overlap conditions.

While individual components such as RAFT (Teed and Deng, 2020) and LightGlue (Lindenberger et al., 2023) are existing tools, their integration into a unified thermal annotation framework with thermal-specific adaptations constitutes the novelty of this work.

2. Related Work

The availability of annotated TIR images has historically been far more limited than that of RGB images. Since TIR images record the long-wavelength radiation emitted from objects, it enables the detection of phenomena invisible to the human eye, such as heat leakage, wild animal activity, and wildfire outbreaks (Bondi et al., 2020, Barrios et al., 2024, Zhang et al., 2025). Early datasets and models therefore focused primarily on natural scenes, where the single heat source and distinct temperature gradients make TIR data particularly useful for environmental monitoring.

In urban environments, however, the situation becomes more complex. Semantic understanding of diverse urban objects such as buildings, pedestrians, and vehicles requires fine-grained annotation and temporal consistency. Visible imagery is widely used for semantic segmentation, yet it struggles under low illumination or night-time conditions. TIR imagery compensates for these limitations and has been increasingly adopted for autonomous driving and UAV-based perception tasks. For example, the *HIT-UAV* dataset (Suo et al., 2023) provides 2,898 thermal images extracted from 43,470 UAV frames across multiple scenarios, and (Riz et al., 2023) introduced the *MONET* dataset containing annotated UAV-based thermal images of pedestrians and vehicles in rural areas. Despite their usefulness, these datasets rely heavily on fully manual annotation, which is time-consuming.

To reduce the effort of thermal labeling, recent studies have explored cross-domain and semi-supervised learning strategies. A common idea is to exploit spatially aligned external data to transfer labels to thermal images. For instance, (Kibe et al., 2025) aligned satellite-based thermal images with land-use maps using georeferenced positioning to automatically generate type annotations for the land. Similarly, (Ji et al., 2024) introduced the *MVUAV* benchmark, a multispectral UAV-view video dataset containing synchronized RGB-Thermal sequences with over thirty semantic categories. In their semi-supervised framework, manual labels are provided only for RGB key frames, and the corresponding thermal frames are co-registered to share the same labels. Through cross-modal consistency regularization and pseudo-label propagation over time, each manually labeled frame can expand to approximately twenty-five additional labeled RGB-Thermal pairs, achieving an impressive 1:25 manual-to-pseudo label ratio. The RGB images play a vital role in labeling to achieve semantic segmentation. Consistency regularization and pseudo-labeling techniques (Chen et al., 2020, Zhuang et al., 2022) demonstrate the potential of leveraging unlabeled data for large-scale thermal video annotation. However, these methods mainly target general semantic segmentation or object detection tasks, rather than thermal anomaly diagnosis or temperature-related interpretation with irregular shape under unpredictable circumstances, where frame-to-frame coherence and radiometric accuracy are critical. Moreover, low frame rates and scene-dependent temperature variation further complicate annotation propagation in TIR images. As a result, most domain-specific datasets for building energy inspection or thermal anomaly detection (Mayer et al., 2023, Vollmer et al., 2025) still rely on manual labeling and expert interpretation, making the process costly and inconsistent.

3. Method

The core idea behind **ThermalAssist** is to exploit correspondences and shared thermal anomalies between overlapping TIR images so that sparse manual labels can be transferred to unlabeled frames. As illustrated in Figure 1, the system takes a collection of overlapping or partially aligned TIR images, where only a small subset is manually labeled. The method estimates pixel-wise correspondences between labeled and unlabeled images using dense matching, and then propagates label information accordingly to produce dense or multi-label masks.

Given a small set of selected images and a larger collection of target ones, ThermalAssist performs three major steps (Figure 2): (1) defining the sparse-labeled image setting and labeling; (2) estimating dense correspondences using RAFT; and (3) propagating and filtering the labels according to the estimated flow fields. Each component is detailed in the following subsections.

3.1 Sparse Label image

Let $\mathcal{D} = (I_i, L_i)_{i=1}^{N}$ denote a set of TIR images. Only a subset $\mathcal{D}_L \subset \mathcal{D}$ is provided with manual annotations L_i, while the majority remain unlabeled ($L_i = \emptyset$). These images are not necessarily sequential video frames but are captured with partial overlap from neighboring viewpoints or adjacent time steps. The goal is to estimate pseudo-labels $\hat{L}_j$ for each unlabeled I_j.

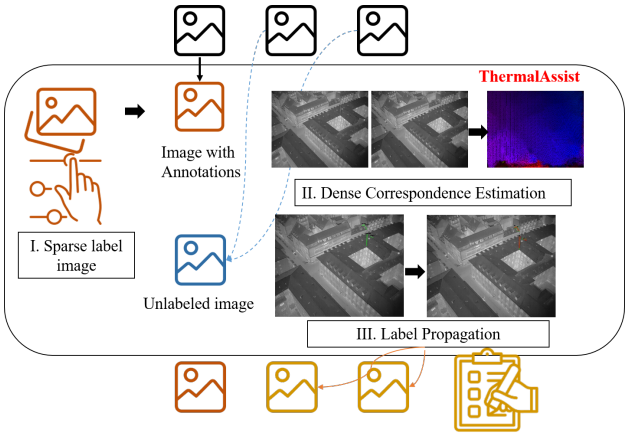

Figure 2. Core workflow for ThermalAssist

Labeling thermal bridges differs from conventional semantic segmentation. A thermal bridge refers to a localized area where heat transfer is intensified due to discontinuities in insulation or variations in material conductivity (ISO 10211:2017 – Thermal bridges in building construction — Heat flows and surface temperatures — Detailed calculations, 2017, Balaras and Argiriou, 2002). Unlike semantic objects in RGB images whose boundaries are visually explicit and defined by color or texture, thermal bridges manifest as subtle temperature gradients or anomalous heat patterns in TIR images. Their identification relies on relative temperature contrasts and structural continuity within the building envelope, requiring annotators to possess a basic understanding of thermography and heat diffusion principles to correctly delineate these regions.

To ensure annotation reliability, each image in $\mathcal{D}_L$ should be labeled by at least two trained annotators and subsequently cross-reviewed by a domain expert to resolve ambiguities and maintain consistency, following a quality-control process similar to that adopted in the Monet dataset (Riz et al., 2023). This protocol minimizes subjectivity and enforces consistent interpretation of thermal patterns across the dataset.

3.2 Dense Correspondence Estimation

We employ the RAFT (Recurrent All-Pairs Field Transforms) model (Teed and Deng, 2020) to estimate pixel-wise correspondences between overlapping TIR images. The RAFT model is employed with weights pretrained on RGB data due to the lack of large-scale TIR datasets with ground-truth optical flow. Our gradient-enhanced input effectively narrows the domain gap by providing structural priors familiar to the model. Dense correspondence aims to compute a dense flow field that aligns two images at the pixel level, denoted as $F_{t\to t+k}(x,y) = (\Delta x, \Delta y)$, where each pixel in I_t is assigned a displacement vector pointing to its corresponding location in I_{t+k}. Compared with sparse feature matching, dense correspondence explicitly models geometric deformation for every pixel, which is crucial for low-texture and viewpoint-varied TIR images.

Unlike previous optical flow networks which predict flow in a single forward pass based on local correlations, RAFT introduces a global all-pairs correlation volume and a recurrent refinement mechanism that iteratively updates the flow with context-aware feedback. Given two feature maps Φ_t and Φ_{t+k} extracted by a shared CNN encoder, RAFT first constructs a 4D correlation volume:

$$C(\mathbf{x}_1, \mathbf{x}_2) = \Phi_t(\mathbf{x}_1)^\top \Phi_{t+k}(\mathbf{x}_2), \tag{1}$$

where $\mathbf{x}_1 \in I_t$ and $\mathbf{x}_2 \in I_{t+k}$. This volume encodes the matching similarity between all pixel pairs, enabling long-range motion reasoning beyond local neighborhoods.

Then, a recurrent update operator $\mathcal{R}$ iteratively refines the flow estimate through a gated recurrent unit (GRU)-based architecture:

$$F^{(n+1)}_{t\to t+k} = F^{(n)}_{t\to t+k} + \Delta F^{(n)}, \quad \Delta F^{(n)} = \mathcal{R}\Big(G(C, F^{(n)}), h^{(n)}\Big), \tag{2}$$

where $G(C, F^{(n)})$ samples local correlation features guided by the current flow estimate, and $h^{(n)}$ is the hidden state of the recurrent unit. This iterative refinement allows consistent updates and subpixel accuracy while maintaining robustness to illumination changes or low texture-conditions typical in TIR images.

RAFT is originally trained under a supervised regime using the endpoint error (EPE) between predicted and ground-truth flow fields. Given ground-truth flow $F^*_{t\to t+k}$, the EPE loss is defined as:

$$\mathcal{L}_{EPE} = \frac{1}{N} \sum_{\mathbf{x}} \|F_{t\to t+k}(\mathbf{x}) - F^*_{t\to t+k}(\mathbf{x})\|_2, \tag{3}$$

where N is the number of valid pixels. During training, RAFT produces multiple intermediate flow estimates, and the total flow loss aggregates the endpoint errors with exponential decay:

$$\mathcal{L}_{flow} = \sum_{n=1}^{N_{iter}} \gamma^{N_{iter}-n} \mathcal{L}^{(n)}_{EPE}, \tag{4}$$

where $\gamma \in (0,1)$ emphasizes later, more accurate iterations. To encourage spatial coherence, a first-order smoothness regularization is applied:

$$\mathcal{L}_{smooth} = \sum_{\mathbf{x}} \|\nabla_x F_{t\to t+k}(\mathbf{x})\|_1 + \|\nabla_y F_{t\to t+k}(\mathbf{x})\|_1, \tag{5}$$

penalizing abrupt flow variations while preserving motion boundaries. The overall objective combines both terms:

$$\mathcal{L}_{total} = \mathcal{L}_{flow} + \lambda_s \mathcal{L}_{smooth}, \tag{6}$$

where λ_s controls the trade-off between flow accuracy and smoothness.

Since TIR images are single-channel images with weak texture and blurry edges, we construct a gradient-enhanced pseudo-RGB input for RAFT instead of directly replicating the intensity channel. Given a thermal image I, we first normalize it to $[0,1]$ to obtain $\tilde{I}$, and compute its spatial gradient magnitude G using Sobel filters in the x and y directions. We then form a three-channel input as:

$$X = [G_x, G_y, \tilde{I}], \tag{7}$$

where the first and second channel emphasizes structural and boundary cues, and the remaining channels preserve radiometric information.

This design injects edge and contrast information into the correlation lookup of RAFT, improving correspondence estimation in low-texture thermal regions while remaining compatible with pretrained weights originally trained on RGB datasets.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-53-2026 | © Author(s) 2026. CC BY 4.0 License.

3.3 Label Propagation and filtering

a b

Figure 3. Problem of directly using RAFT for label transfer in (a) high overlap and (b) Low overlap. Green: GT, Red: transferred label.

Due to the varying frame rate and partial overlap among thermal image pairs, not all frames can be directly aligned via a flow-based method. Figure 3 shows an example of using RAFT only in both high overlap and low overlap situations. The misalignment becomes severe when images share low overlap. We therefore first evaluate the **frame similarity** between a labeled image I_t and an unlabeled image I_{t+k}. Since direct registration on TIR images is often unreliable due to weak texture and blurry boundaries, we compute the gradient maps $G_t = \nabla I_t$ and $G_{t+k} = \nabla I_{t+k}$, and use the LightGlue (Lindenberger et al., 2023) matcher to establish keypoint correspondences. The similarity between two frames is quantified by the pixel-wise overlap ratio:

$$S(I_t, I_{t+k}) = \frac{|M_t \cap M_{t+k}|}{|M_t \cup M_{t+k}|}, \tag{8}$$

where M_t and M_{t+k} denote the valid matched pixel sets after outlier filtering.

If $S(I_t, I_{t+k}) < \delta_s$, a **coarse-to-fine alignment** is first performed to roughly register the two images based on matched keypoints before estimating the dense correspondence field; otherwise, we directly apply the RAFT model. In our experiments, the overlap threshold is set to $\delta_s = 0.7$ based on empirical analysis of the TBBR dataset, where image pairs below this value exhibit insufficient geometric correspondence for reliable direct flow estimation.

Once the dense flow $F_{t\to t+k}$ is obtained, label propagation becomes a deterministic process. Given a labeled image I_t with its annotation mask L_t, and an unlabeled image I_{t+k}, each label is transferred as:

$$(x', y') = (x, y) + F_{t\to t+k}(x, y), \qquad L_{t+k}(x', y') = L_t(x, y). \tag{9}$$

Confidence Weighting Each propagated pixel is assigned a confidence weight $w(x, y)$ according to the correlation magnitude or the internal convergence stability of RAFT:

$$w(x, y) = \sigma\big(\|C(I_t, I_{t+k})(x, y)\|\big), \tag{10}$$

where $\sigma(\cdot)$ denotes a min-max normalization that maps the correlation magnitude to $[0, 1]$, defined as $\sigma(x) = (x - x_{\min})/(x_{\max} - x_{\min})$ computed over all pixels in the target image. These confidence maps are later used to weight multiple propagated masks or suppress uncertain regions.

Algorithm 1 ThermalAssist Label Propagation Workflow

Require: Labeled TIR images $\mathcal{D}_L = \{I_i, L_i\}$, Unlabeled images I_j, Overlap threshold δ_s
Ensure: Pseudo-labels $\hat{L}_j$ for unlabeled frames
Step 1: Gradient Enhancement
For each I, compute Sobel gradients G_x, G_y and construct pseudo-RGB input $X = [G_x, G_y, \tilde{I}]$.
Step 2: Overlap-aware Alignment
Calculate similarity $S(I_t, I_{t+k})$ using LightGlue keypoint matching.
if $S(I_t, I_{t+k}) < \delta_s$ **then**
  Apply coarse-to-fine registration based on matched keypoints.
else
  Estimate dense flow $F_{t\to t+k}$ using RAFT with input X.
end if
Step 3: Label Propagation & Fusion
Warp L_t to I_{t+k} via flow field F and assign confidence weights $w(x, y)$.
Step 4: Post-processing
Refine $\hat{L}_{t+k}$ using morphological smoothing and bounding-box alignment.
return Final pseudo-labels $\hat{L}_j$

Post-Processing To enhance spatial coherence and geometric consistency, we refine the propagated labels through *bounding-box based adjustment* and *morphological smoothing*. Small isolated components are removed, and object boundaries are aligned with existing bounding boxes or semantic segments. The final pseudo-label $\hat{L}_{t+k}$ thus preserves both geometric integrity and semantic continuity with the source annotation. In cases where a target image receives labels from multiple source frames, conflicts are resolved by selecting the label with the highest confidence weight $w(x, y)$, ensuring radiometric and geometric consistency.

This hierarchical pipeline from gradient-based similarity estimation to flow-guided label propagation, enables robust pseudo-label generation under heterogeneous thermal imaging conditions, forming a core element of the proposed **ThermalAssist** framework.

4. Experiments and Result

We evaluated our proposed method on the **Thermal Bridges on Building Rooftops (TBBR)** dataset (Mayer et al., 2023). It is a multi-channel UAV dataset acquired over the city center of Karlsruhe, Germany, during multiple flight campaigns. Each image is co-registered across RGB, TIR, and height map channels, forming a comprehensive multi-modal benchmark for building-level heat-loss analysis. [1]

In this study, we exclusively used the **thermal infrared channel** to evaluate the robustness of our label propagation framework under realistic single-modality conditions. We conducted experiments on two subsets of TBBR images that were roughly divided according to spatial coverage and image continuity. **Site 1** contains eight overlapping TIR images representing a compact area with high inter-frame overlap, while **Site 2** includes twenty-nine images covering a larger and more heterogeneous region. Neither site forms a fully continuous video sequence; instead, the images were selected based on spatial adjacency to ensure meaningful correspondence for label propagation.

[1] The implementation of the proposed **ThermalAssist** framework, along with the experimental scripts and pretrained models, will be made publicly available at: `https://github.com/Ano/ThermalAssist`.

Following the labeling protocol of the TBBR dataset, each thermal bridge or heat-loss region was manually delineated by expert annotators using polygonal masks provided by the dataset. To ensure annotation reliability, all masks were cross-validated by multiple experts and refined to correct boundary ambiguities inherent to low-texture thermal data. For object detection evaluation, bounding boxes were derived from the polygonal annotations to maintain consistency with standard metrics.

To comprehensively assess propagation behavior, we constructed *pairwise experiments* between adjacent images to estimate the similarities between pairs. For each pair (I_s, I_t), the earlier image served as the *source* and the subsequent image as the *target*. This pairwise design enables direct evaluation of label transfer accuracy between overlapping views without relying on temporal continuity. Three representative configurations were considered: (i) **C-P1**, corresponding to two consecutive and highly overlapping images; (ii) **D-P2**, representing a direct propagation between nonadjacent images with step of 2; and (iii) **CC-P2**, a cross-continuous case that shares the same source and target as D-P2 but includes an intermediate overlapping image as a transitional step. For each target image, the propagated label was compared with its manual annotation, which served as the ground truth (GT). This design allows us to analyze how spatial continuity and frame overlap influence label transfer quality across different site configurations.

To quantify performance, we adopted standard object detection and segmentation metrics, including **Precision**, **Recall**, **F1-score**, and the mean **Intersection over Union (mIoU)** between predicted and ground-truth bounding boxes or masks. These metrics jointly evaluate the completeness and geometric accuracy of the transferred annotations, providing a comprehensive assessment of label propagation quality under varying overlap conditions.

Figure 4 illustrates representative examples of the proposed label propagation results under different frame-overlap conditions on TIR images. Manual annotations (GT) are shown in green, and automatically transferred labels are indicated in red for visual comparison. Across both sites, most thermal anomaly labels are successfully transferred to adjacent views, exhibiting strong spatial and semantic correspondence with the ground truth. The visual results confirm that the majority of thermal anomaly labels can be reliably transferred across different viewpoints, maintaining strong spatial correlation with the manually annotated ground truth.

When two TIR images exhibit high overlap (e.g., *DJI_0179_R* and *DJI_0181_R*), most building structures and thermal patterns can be precisely aligned through the proposed gradient-enhanced flow estimation. As a result, the transferred labels exhibit smooth, well-defined boundaries that nearly coincide perfectly with the manual annotations. In contrast, when the overlap between frames is small (e.g., *DJI_0181_R* and *DJI_0183_R*), some anomalies appear partially missing or slightly distorted due to viewpoint shifts and parallax. In these challenging cases, the propagated masks may lose fine boundary details or drift along facade edges. Nevertheless, the confidence weighting and bounding-box–guided post-processing effectively suppress outliers and maintain overall geometric coherence.

Quantitative results for both **Site 1** and **Site 2** are summarized in Table 1. Across all experiments, the proposed method achieves stable and accurate performance, with mean F1-scores above **0.80** for bounding-box detection and above **0.70** for segmentation. On Site 1, which features dense overlaps and similar viewpoints, the propagated labels achieve the highest IoU values, indicating strong pixel-level consistency. Site 2, containing a larger number of images with more diverse viewpoints, shows a moderate decrease in IoU but maintains comparable F1-scores, confirming the framework’s robustness under less-overlapped and more heterogeneous conditions.

Site	Setting	BBox				Segmentation			
		Prec.	Rec.	F1	mIoU	Prec.	Rec.	F1	mIoU
1	C-P1	0.85	0.76	0.80	0.69	0.77	0.69	0.73	0.61
	D-P2	0.67	0.57	0.61	0.66	0.60	0.51	0.55	0.58
	CC-P2	0.88	0.57	0.69	0.66	0.79	0.51	0.62	0.57
2	C-P1	0.92	0.86	0.89	0.75	0.85	0.80	0.82	0.58
	D-P2	0.84	0.73	0.78	0.70	0.76	0.66	0.70	0.54
	CC-P2	0.83	0.73	0.77	0.71	0.74	0.65	0.70	0.54

Table 1. Quantitative results of label propagation on **Site 1** and **Site 2**. Detection (BBox) and segmentation (mask) metrics are reported.

A clear relationship is observed between the degree of overlap and label quality. Frames with larger spatial overlap generally yield higher precision and smoother segmentation boundaries, while pairs with smaller overlap exhibit reduced recall because many ground-truth objects fall outside the shared field of view between the source and target images. These omissions are therefore not algorithmic failures but rather physical non-overlaps, highlighting the importance of overlap estimation in guiding reliable label propagation.

Among the tested configurations, the **C-P1** case representing continuous and high-overlap pairs achieves the best results on both sites, with F1-scores of 0.80 and 0.89 for bounding-box detection and 0.73 and 0.82 for segmentation. These results confirm that when sufficient geometric correspondence exists between consecutive frames, the propagation process can accurately preserve object completeness and boundary integrity. In contrast, performance declines under the **D-P2** and **CC-P2** settings, where overlap is smaller and viewpoint variation is more pronounced. Nevertheless, the mean IoU remains relatively stable (around 0.66–0.71), suggesting that once correspondences are correctly established, the propagated labels remain geometrically well aligned.

A closer comparison between **CC-P2** and **D-P2** further reveals the role of intermediate-frame transitions. Both settings share the same source and target images, but CC-P2 introduces one additional overlapping frame as a transitional step, whereas D-P2 performs a direct propagation. CC-P2 achieves slightly higher precision while maintaining similar IoU, indicating that intermediate frames can help stabilize propagation by reducing spatial discontinuities between distant viewpoints. However, this benefit diminishes when the intermediate image contains weak thermal texture or poor registration quality. These observations suggest that the choice of propagation path—whether to include intermediate overlaps—should be adaptively determined based on inter-frame similarity rather than fixed by sequence order. Such adaptive path selection could further improve label coverage and reduce omission in large-scale, multi-view thermal datasets.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-53-2026

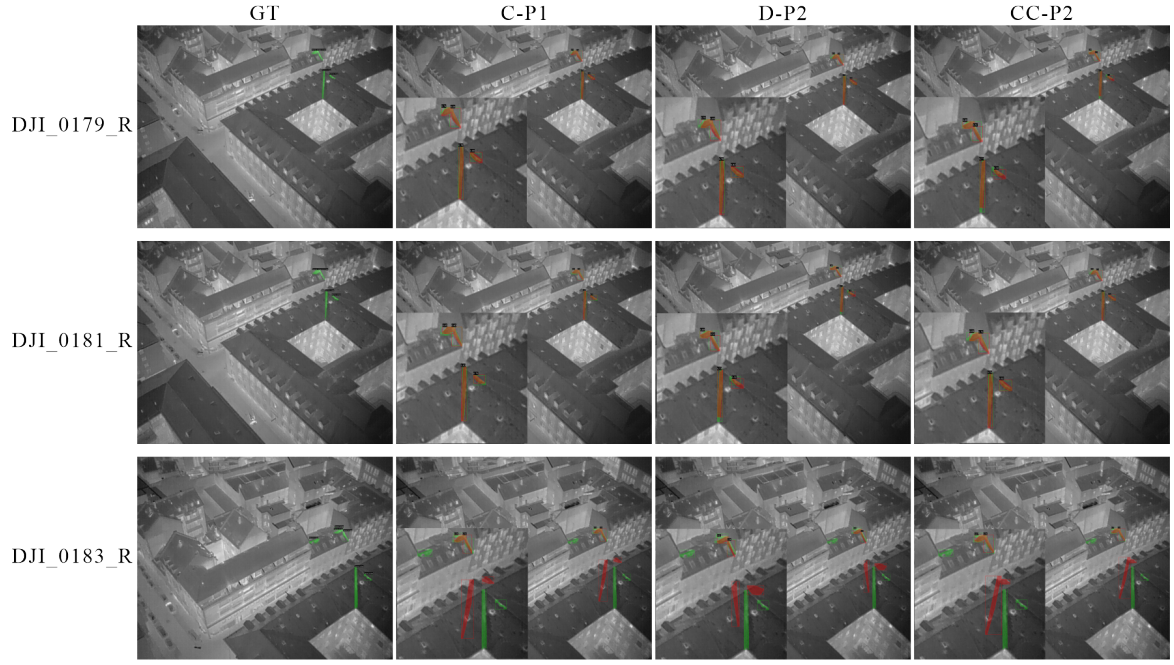

Figure 4. Example results for label transfer.

5. Discussion

To evaluate the efficiency of our method, our results are compared with the **SAMURAI** tracking-based label propagation framework (Yang et al., 2024), a state-of-the-art 2D tracking method that combines flow estimation and object tracking to establish temporal associations across consecutive frames. Utilizing SAMURAI, we perform dense motion estimation between adjacent images and transfer labeled masks through iterative warping and linking, achieving label propagation in image sequences.

Table 2 reports the quantitative comparison on **Site 1** of the TBBR dataset. Our method consistently outperforms SAMURAI on the TIR data across both continuous and cross-continuous configurations. In the continuous case (C-P1), the proposed framework achieves an F1-score of **0.87** and a mean IoU of **0.69**, notably higher than SAMURAI (F1 = 0.56, mIoU = 0.68). Similarly, under the cross-continuous setting (CC-P2), where inter-frame overlap decreases, our model maintains stable performance (F1 = 0.70), while SAMURAI drops sharply to 0.31. These results demonstrate that introducing gradient-based similarity and geometric alignment substantially improves label stability under low-texture and low-overlap conditions.

Although SAMURAI occasionally achieves comparable mean IoU values (around 0.67-0.68), its *recall* and overall detection rate are significantly lower, primarily because the method often misses small or partially occluded objects. This limitation stems from its reliance on *pixel-level motion consistency* without incorporating camera geometry or 3D structural constraints. Such a design performs well for high-frame-rate RGB videos but becomes unstable for UAV-based TIR images, where parallax, occlusion, and illumination variation lead to unreliable optical flow. Consequently, SAMURAI exhibits label drift, duplicated detections, and shape distortions when applied to low-overlap thermal data.

In contrast, our proposed framework explicitly models spatial geometry and gradient-based similarity between frames. By estimating overlap ratios, it adaptively applies coarse-to-fine alignment or RAFT-based dense correspondence and refines propagated labels via confidence weighting and bounding-box–guided filtering. This integration of geometric priors and structural cues preserves object location, shape, and overlap consistency throughout the propagation process.

Furthermore, when the same model is applied to the RGB modality, the performance drops substantially (F1 $<$ 0.3), despite similar geometric configurations. This observation implies that thermal signals encode more stable radiometric features for anomaly localization, whereas RGB textures are sensitive to illumination and viewpoint variation. By incorporating structural correspondence, gradient information, and overlap-aware matching, our framework effectively mitigates label drift, improves spatial consistency, and enhances pseudo-label reliability for UAV-based thermal anomaly mapping.

Data	Method	Setting	Prec.	Rec.	F1	mIoU
TIR	Our	C-P1	**0.88**	**0.86**	**0.87**	0.69
TIR	Our	CC-P2	0.76	0.66	0.70	0.63
TIR	SAMURAI	C-P1	0.53	0.59	0.56	0.68
TIR	SAMURAI	CC-P2	0.31	0.32	0.31	0.67
RGB	Our	C-P1	0.24	0.27	0.26	0.64
RGB	Our	CC-P2	0.13	0.13	0.13	0.67

Table 2. Comparison between the proposed method and SAMURAI under different modalities and overlap conditions. C-P1 and CC-P2 denote continuous step1 and cross-continuous pairs step2, respectively.

Our experiments confirm that the proposed assisted propagation enables successful label transfer across multiple thermal views, yielding quantitatively stable and visually coherent results. Nevertheless, this work represents only an initial step toward a fully integrated annotation framework for automatically identifying corresponding thermal annotations. How to synchronize object boundaries while maintaining reasonable shape consistency and optimizing for low-overlap precision remains an open question for investigation.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
https://doi.org/10.5194/isprs-annals-XI-2-2026-53-2026 | © Author(s) 2026. CC BY 4.0 License.

While the framework involves dense matching, it drastically reduces human effort. ThermalAssist achieves a manual-to-pseudo label ratio of up to 1:10 in the test area, enabling large-scale survey annotation in a fraction of the time required for manual labeling. The inference time per image pair is approximately 1.2s on a modern GPU, making it highly scalable for extensive UAV datasets. Although validated on the TBBR dataset, the geometry-based nature of ThermalAssist allows for potential adaptation to other TIR applications, such as forest fire detection or wildlife monitoring, where inter-frame overlap is maintained.

Future efforts will focus on extending this concept toward a more comprehensive system, tentatively designed to support large-scale, multi-view, and multi-condition annotation of TIR images. It is envisioned as a four-stage pipeline consisting of: (1) *image-graph reconstruction* to model inter-image relationships, (2) *keyframe selection* to select the reasonable TIR images for labeling while minimizing manual annotation cost, (3) *label propagation* (the focus of this study), and (4) *converged label fusion* to integrate multi-view annotations into a spatially consistent global map. Within this roadmap, **ThermalAssist** corresponds to Stage (3) and (4) and serves as a proof-of-concept, validating the feasibility of automatic label transfer under real-world thermal imaging conditions. This work therefore establishes a solid foundation for developing a complete, scalable framework for reliable TIR annotation in future studies.

6. Conclusion

In this study, we proposed **ThermalAssist**, a geometry-aware and gradient-enhanced framework designed to assist the labeling of thermal anomalies in UAV-based TIR images. Unlike conventional manual annotation workflows that are often labor-intensive and inconsistent due to low contrast and blurred boundaries in TIR data, ThermalAssist enables the automatic propagation of labels across overlapping views. Through pairwise experiments on the TBBR dataset, we demonstrated that the proposed approach can achieve accurate and spatially consistent label transfer, with F1-scores of up to **0.87** and mean IoU of **0.69** in high-overlap settings. The framework effectively reduces *missing annotations* and *inconsistent boundaries*, while also serving as a practical tool for quality checking and accelerating manual labeling.

Compared with the state-of-the-art **SAMURAI** tracking-based propagation method, ThermalAssist shows superior robustness under low-texture and low-overlap conditions. By explicitly incorporating geometric alignment and gradient-based similarity, our method maintains stronger spatial correspondence with ground-truth annotations and produces fewer drifted or duplicated detections. These results confirm the advantage of integrating structural priors and inter-frame similarity estimation for reliable label transfer in TIR images.

ThermalAssist represents a key intermediate step toward a more comprehensive framework for thermal annotation, which we tentatively refer to as **ThermalAnno**. This envisioned system will extend the current work from pairwise propagation to a fully integrated multi-view annotation pipeline, including (1) image-graph reconstruction, (2) keyframe selection, (3) optical-flow–based label propagation, and (4) converged multi-view label fusion. Future research will focus on implementing these additional components to enable scalable, cross-view consistent annotation of large-scale thermal datasets. Ultimately, we aim to develop a unified thermal annotation system that bridges manual expertise and automated geometry-aware assistance for efficient, reliable, and reproducible TIR data labeling.

References

Balaras, C. A., Argiriou, A. A., 2002. Infrared thermography for building diagnostics. *Energy and Buildings*, 34(2), 171–183.

Barrios, D. B., Valente, J., Langevelde, F. v., 2024. Monitoring mammalian herbivores via convolutional neural networks implemented on thermal UAV imagery. *Computers and Electronics in Agriculture*, 218, 108713.

Bondi, E., Jain, R., Aggrawal, P., Anand, S., Hannaford, R., Kapoor, A., Piavis, J., Shah, S., Joppa, L., Dilkina, B. et al., 2020. Birdsai: A dataset for detection and tracking in aerial thermal infrared videos. *Proceedings of the IEEE/CVF Winter conference on applications of computer vision*, 1747–1756.

Chen, L.-C., Lopes, R. G., Cheng, B., Collins, M. D., Cubuk, E. D., Zoph, B., Adam, H., Shlens, J., 2020. Naive-student: Leveraging semi-supervised learning in video sequences for urban scene segmentation. *European conference on computer vision*, Springer, 695–714.

Delmastro, C., Bienassis, T. D., Goodson, T., Lane, K., Marois, J.-B. L., Martinez-Gordon, R., Husek, M., 2022. Buildings. `https://www.iea.org/reports/buildings`. [Online; accessed 19-Oct-2022].

ISO 10211:2017 – Thermal bridges in building construction — Heat flows and surface temperatures — Detailed calculations, 2017.

Ji, W., Li, J., Li, W., Shen, Y., Cheng, L., Jin, H., 2024. Unleashing Multispectral Video’ s Potential in Semantic Segmentation: A Semi-supervised Viewpoint and New UAV-View Benchmark. A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, C. Zhang (eds), *Advances in Neural Information Processing Systems*, 37, Curran Associates, Inc., 65717–65737. https://nips.cc/virtual/2024/poster/93562?utm_source=chatgpt.com.

Kibe, M., Inoue, T., Morioka, J., Miyamoto, R., 2025. Automatic annotation method for day–night aerial infrared image dataset creation and its application to semantic segmentation. *Optical Engineering*, 64(9), 092203. Publisher: SPIE.

Lindenberger, P., Sarlin, P.-E., Pollefeys, M., 2023. Lightglue: Local feature matching at light speed. *Proceedings of the IEEE/CVF international conference on computer vision*, 17627–17638.

Mayer, Z., Kahn, J., Götz, M., Hou, Y., Beiersdörfer, T., Blumenröhr, N., Volk, R., Streit, A., Schultmann, F., 2023. Thermal Bridges on Building Rooftops. *Scientific Data*, 10(1), 268. https://doi.org/10.1038/s41597-023-02140-z.

Riz, L., Caraffa, A., Bortolon, M., Mekhalfi, M. L., Boscaini, D., Moura, A., Antunes, J., Dias, A., Silva, H., Leonidou, A., Constantinides, C., Keleshis, C., Abate, D., Poiesi, F., 2023. The MONET dataset: Multimodal drone thermal dataset recorded in rural scenarios. *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2546–2554.

Suo, J., Wang, T., Zhang, X., Chen, H., Zhou, W., Shi, W., 2023. HIT-UAV: A high-altitude infrared thermal dataset for Unmanned Aerial Vehicle-based object detection. *Scientific Data*, 10(1), 227. https://pegasus.ac.cn.

Teed, Z., Deng, J., 2020. Raft: Recurrent all-pairs field transforms for optical flow. *European conference on computer vision*, Springer, 402–419.

Vollmer, E., Ruck, J., Volk, R., Schultmann, F., 2025. Leak detection using thermal imagery: Deep learning versus traditional computer vision state-of-the-art. *ISPRS Journal of Photogrammetry and Remote Sensing*, 228, 505–518.

Yang, C.-Y., Huang, H.-W., Chai, W., Jiang, Z., Hwang, J.-N., 2024. SAMURAI: Adapting Segment Anything Model for Zero-Shot Visual Tracking with Motion-Aware Memory. _eprint: 2411.11922.

Zhang, Y., Rui, X., Song, W., 2025. A UAV-Based Multi-Scenario RGB-Thermal Dataset and Fusion Model for Enhanced Forest Fire Detection. *Remote Sensing*, 17(15). https://www.mdpi.com/2072-4292/17/15/2593.

Zhuang, J., Wang, Z., Gao, Y., 2022. Semi-supervised video semantic segmentation with inter-frame feature reconstruction. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3253–3261.