Sensors and Materials, Vol. 38, No. 7 (2026) 3837–3856
MYU Tokyo

S & M 4537

Unsupervised Diffusion-trajectory Segmentation and Multiregion Fusion for Understanding RGB–thermal Scene

Muhammad Waqas Ahmed,[1] Tingting Xue,[2,3] Nouf Abdullah Almujally,[4] Ahmad Jalal,[1,5*] and Hui Liu[3,6,7**]

[1]Department of Computer Science, Air University, E-9, Islamabad 44000, Pakistan
[2]School of Environmental Science & Engineering, Nanjing University of Information Science and Technology, Nanjing 210044, China
[3]Cognitive Systems Lab, University of Bremen, Bremen 28359, Germany
[4]Department of Information System, College of Computer and Information Sciences, Princess Nourah bint Abdulrahman University, Riyadh 11671, Saudi Arabia
[5]Department of Computer Science and Engineering, College of Informatics, Korea University, Seoul 02841, South Korea
[6]Jiangsu Key Laboratory of Intelligent Medical Image Computing, School of Artificial Intelligence (School of Future Technology), Nanjing University of Information Science and Technology, Nanjing, China
[7]Guodian Nanjing Automation Co., Ltd, Nanjing, China

(Received January 15, 2026; accepted June 8, 2026)

Keywords: cross-modal sensor fusion, scene classification, multimodal, region encoding, segmentation

Multispectral scene understanding remains a critical sensing challenge: visible-spectrum (RGB) image sensors provide high-resolution color and texture cues but fail under poor illumination, while thermal infrared sensors operating on the principle of detecting heat radiation emitted by objects are inherently robust to low-light conditions yet deliver lower spatial resolution and less distinct contrast characteristics. The effective fusion of these two sensor modalities is therefore essential for reliable autonomous perception in real-world environments. To address the sensing limitations of individual modalities and the scarcity of large, densely annotated RGB–thermal (RGB–T) datasets, we propose a hierarchical, multistage RGB–T framework that (1) performs adaptive layer-wise fusion of multiscale features from a dual-stream backbone (ResNet-18 for RGB sensor data, and a custom convolutional neural network with squeeze-and-excitation blocks for thermal sensor data), (2) produces sharp, unsupervised region segmentations by mining latent diffusion trajectories of a pretrained diffusion model, and (3) aggregates region features via a graph-structured hierarchical fusion network for robust scene reasoning. The proposed sensing pipeline preserves complementary local and global cues from both modalities, yields semantically coherent region proposals without manual annotations, and models region-to-region context for stronger scene descriptors. Evaluated on RGB–T benchmarks, the integrated pipeline improves segmentation (Mean Intersection over Union ≈ 0.758) and scene classification (mean accuracy ≈ 88%) over representative baselines, demonstrating superior boundary precision, object separation, and robustness to clutter and occlusion, highlighting its value for sensor-driven autonomous scene understanding systems.

*Corresponding author: e-mail: ahmadjalal@au.edu.pk
**Corresponding author: e-mail: hui.liu@uni-bremen.de
https://doi.org/10.18494/SAM6151

ISSN 0914-4935
https://myukk.org/

1. Introduction

The extraction and analysis of features from RGB images have become a widely used processing technique in computer vision that has found its way into a diverse array of industrial, commercial, and everyday applications. However, this technique exhibits limitations, primarily owing to its confinement to the visible spectrum. The most significant constraint is that it only operates effectively under good lighting and clear visibility conditions. This has prompted researchers to explore the use of RGB–Depth (RGB–D) and thermal cameras for multispectral perception in recent years.[1,2]

RGB cameras, being ubiquitous and economical, offer high-resolution color images that can be readily interpreted by both human observers and computer vision algorithms. They excel in tasks such as object recognition, scene understanding, and texture analysis. However, their efficacy is heavily dependent on good lighting conditions. Thermal infrared sensors, in contrast to visible sensors, can sense slight temperature differences between objects and their surroundings. This capability is not hindered by low-light conditions or complete darkness, as the operation of these sensors is based on thermal radiation independent of any light source. This unique capability makes thermal sensing a valuable modality for object detection under challenging conditions.[3] However, thermal sensors typically offer lower resolution than RGB cameras and are more expensive. The overarching aim of sensor fusion is to harmonize the strengths of each sensor modality to mitigate their individual limitations. However, achieving effective fusion requires careful calibration and alignment of the sensors, along with sophisticated algorithms to integrate the different types of data.[4]

The complementary nature of RGB and thermal infrared sensing makes their fusion inherently attractive for robust scene understanding. RGB sensors capture high-resolution spatial and spectral details that are invaluable for texture discrimination and object recognition. Thermal sensors, by contrast, respond to emitted infrared radiation and are therefore insensitive to visible-light illumination, enabling robust detection under nighttime or shadowed conditions where RGB sensors fail. However, the two sensor modalities differ substantially in spatial resolution, contrast characteristics, dynamic range, and noise properties, creating significant challenges for effective data fusion. Achieving effective RGB–thermal (RGB–T) sensor fusion requires careful calibration and co-registration of the two sensing streams to establish pixel-level correspondence. Differences in sensor field of view, focal length, and lens distortion must be resolved using intrinsic and extrinsic calibration procedures based on the pinhole camera model. Beyond geometric alignment, the modality gap reflecting systematic differences in feature statistics between RGB and thermal representations poses a fundamental challenge for learned fusion models, particularly when training data are scarce. These challenges motivate the development of adaptive, multiscale fusion architectures that can learn to balance modality contributions depending on scene context and sensing conditions.[5,6]

To achieve an effective fusion of the different modalities, it is essential to calibrate each sensor and align them in the same coordinate system, which involves determining the intrinsic and extrinsic parameters using the pinhole camera model. Aligning the modalities correctly is crucial for achieving precise data fusion. Although descriptor-based methods utilizing feature

point matching algorithms can accomplish registration, they are often unsuitable for real-time applications involving moving cameras.[7] This is because of their high computational complexity and the challenges in implementing them with thermal data, which displays distinct characteristics compared with visual data.[8]

In this paper, we address the fundamental sensing challenge of combining complementary-sensor-modality RGB cameras and thermal infrared sensors for robust urban scene understanding. RGB cameras, as ubiquitous electro-optical sensors, provide dense texture and color information but are severely limited by illumination conditions. Thermal infrared sensors, which detect emitted heat radiation rather than reflected light, operate independently of ambient lighting and can resolve temperature contrasts invisible to RGB sensors. Fusing the outputs of these two sensor types is a core challenge in multimodal sensing systems, since the modalities differ in spatial resolution, contrast characteristics, and noise profiles. Sensor misalignment and the lack of cross-modal correspondence further complicate the integration. To address these sensing-system challenges, we propose a novel dual-stream feature extraction framework that combines the complementary strengths of RGB and thermal sensors through adaptive, layer-wise multiscale fusion. Our approach processes RGB sensor data through a truncated ResNet-18 backbone while simultaneously analyzing aligned thermal sensor data through a custom convolutional neural network (CNN) enriched with squeeze-and-excitation (SE) blocks. The resulting multimodal feature representations are fused through a learned adaptive mechanism that preserves the most discriminative characteristics from both sensing streams. The fused sensor features are then passed to a latent diffusion trajectory-based segmentation (LDTM) module that performs unsupervised scene decomposition, followed by a graph-structured region reasoning network for final scene classification. This sensing-aware architecture directly addresses several key limitations of existing multimodal perception approaches, particularly the inability to exploit fine-grained complementary sensor information across multiple feature scales.

2. Literature Review

This research builds on prior work in RGB–T fusion for segmentation and scene understanding. Existing studies explore multimodal alignment, attention-based fusion, edge-guided refinement, and multilevel feature integration; however, they remain limited by rigid fusion strategies, high computational complexity, and domain-specific performance, especially in nonurban scenes. To clearly position our contribution, we summarize key RGB–T fusion models in Table 1, highlighting their strengths and the persistent gaps they leave. These limitations motivate the development of our proposed framework, which aims to deliver more adaptive, efficient, and generalizable scene understanding across diverse environments.

Table 1 shows that existing RGB–T fusion models often use rigid or single-stage fusion, depend on heavy attention modules or pseudo-labels, and are limited mainly to urban scenes with high computational cost. To address these issues, our model introduces a lightweight multiscale adaptive fusion backbone, an unsupervised diffusion-based segmentation module, and a region-graph reasoning network. Together, these components reduce supervision

Table 1
Literature review of RGB-T fusion for segmentation and scene understanding.

Model Name	Main Contribution	Limitations
FuseSeg[9]	Introduces end-to-end RGB-T fusion with deformation fields for registration and semantic alignment loss for urban scene segmentation	The registration overhead affects real-time performance and might struggle with nonrigid transformations. It is also limited to urban environments
FEANet[10]	Introduces feature-enhanced attention network for real-time RGB-T segmentation, which is lightweight and designed with attention mechanisms	Trade-off between speed and accuracy. It has limited feature representation capacity so it might struggle with complex scenes.
GMNet[11]	Graded-feature multilabel learning network; divides features into two levels with different fusion strategies for urban scenes	Multilabel supervision increases annotation complexity; limited scalability to other domains; requires extensive labels
EGFNet[12]	Edge-aware guidance fusion with multiscale edge detection; focuses on boundary preservation for scene parsing	Edge detection preprocessing overhead; sensitive to noisy edges; may over-emphasize boundaries at expense of regions
SGFNet[13]	Semantic-guided fusion with asymmetric encoder; multimodal coordination and distillation units (MCDUs) for feature fusion	High computational complexity; dependence on thermal semantic quality; requires careful hyperparameter tuning
MEFNet[14]	Decouples 3D attention into 2D modal weights and channel attention; multiple expert modules for modal differences	Complex attention mechanism increases inference time; high memory requirements; risk of overfitting to specific patterns
MTANet[15]	Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding; joint learning framework	Multitask optimization complexity; requires multiple loss balancing; increased training difficulty
PAIF[16]	Perception-aware infrared-visible fusion for attack-tolerant semantic segmentation; robustness to adversarial attacks	Focused on adversarial robustness; complex adversarial training pipeline; may sacrifice clean accuracy
Multilevel Fusion[17]	Multilevel RGB-T fusion for object tracking; enhanced thermal representation training paradigm	Not designed for classification; requires temporal consistency; not suitable for static scene understanding
CACFNet[18]	Cross-modal attention cascaded fusion network; multilevel feature integration for urban scene parsing	Multiple cascade stages increase latency; complex training procedure; limited generalization beyond urban scenes
DCFNet[19]	Dense complementary fusion with complementary vote attention (CVA) module; exploits complementary information	Heavyweight architecture with high memory consumption; slow inference speed; struggles with real-time applications
RSFNet[20]	Residual spatial fusion with asymmetric dual-encoder; saliency detection for pseudo-label generation; structural re-parameterization	Pseudo-label quality affects performance; additional saliency detection overhead; limited to spatial fusion only

requirements, improve generalization to diverse scenes, and deliver stronger segmentation and scene understanding performance than prior RGB–T methods.

3. Data, Materials, and Methods

The proposed framework integrates a hierarchical RGB–T fusion backbone with an unsupervised diffusion-trajectory segmentation module to generate rich, semantically coherent region representations for scene understanding. The overall architecture is illustrated in Fig. 1. The methodology consists of three core components.

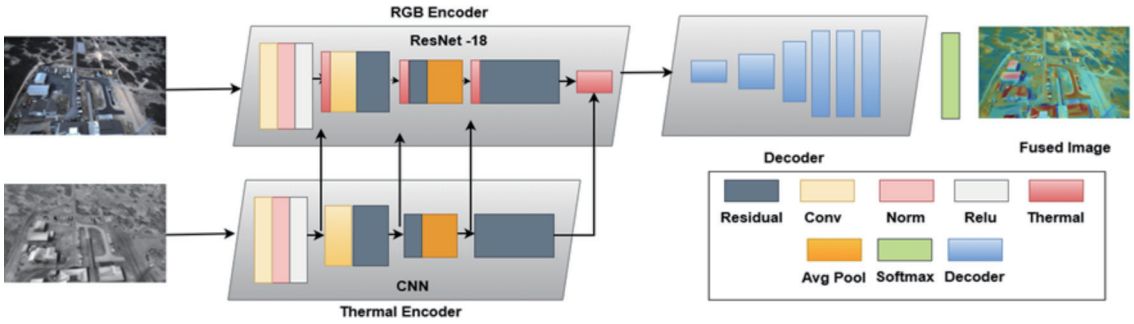


Fig. 1. (Color online) H-MSSF Network for RGB and thermal images.

- Hierarchical multiscale feature fusion (H-MSFF)
- LDTM
- Region-level encoding and structural representation

Each module is designed to contribute complementary information on local textures, global semantics, and diffusion-driven structural cues while maintaining computationally efficient backbone design, although the full pipeline operates in an offline regime owing to diffusion-based segmentation. The proposed system's architecture is visually represented in Fig. 1.

3.1 Preprocessing

All images were resized to 512 × 512 pixels. RGB images were normalized using ImageNet statistics (*mean* = [0.485, 0.456, 0.406], *std* = [0.229, 0.224, 0.225]), and thermal images were normalized independently to zero mean and unit variance over the training split. Bilinear upsampling was applied to the thermal channel where resolution differences remained after rectification. Training augmentation includes random horizontal flipping (p = 0.5), rotation within ±15°, RGB-only color jitter (*brightness* = 0.2, *contrast* = 0.2), and random cropping to 448 × 448 pixels.

3.2 H-MSFF

The first stage of our methodology focuses on extracting strong and complementary features from RGB and thermal images. RGB images capture textures, edges, and colors, while thermal images provide heat signatures and structural information that are often robust to illumination changes. To leverage these complementary modalities, we propose the H-MSFF network, which consists of two parallel pathways. The first pathway is a truncated ResNet-18, consisting of four layers, to capture deep semantic features. Each layer gradually extracts increasingly abstract representations, from low-level textures to high-level semantic patterns. The second pathway is a custom CNN enhanced with SE blocks, which processes thermal images. SE blocks dynamically adjust the importance of feature channels, allowing the network to focus on relevant thermal patterns.

What makes H-MSFF distinctive is its layer-wise adaptive fusion. Instead of fusing features only at the last layer, H-MSFF merges features after each corresponding stage of the two backbones. Before fusion, feature maps are aligned using 1 × 1 convolution layers:

$$`F_i^R = W_i^R\left(F_i^R\right),\ `F_i^C = W_i^C\left(F_i^C\right), \tag{1}$$

where W_i^R, W_i^C are 1 × 1 convolution layers. The fusion is performed using learnable weights α_i and β_i to control the contribution from each network.

$$\alpha_i, \beta_i = Softmax\left(\left[\alpha_i, \beta_i\right]\right) \tag{2}$$

$$F_i^{Fused} = \alpha_i,\ `F_i^R + \beta_i,\ `F_i^C \tag{3}$$

This fusion mechanism ensures balanced information exchange between ResNet and CNN at multiple scales, preventing feature dominance from a single network and enabling the model to adaptively prioritize relevant features. Once features from all levels have been fused, the hierarchical feature aggregation (HFA) module integrates the fused representations into a final feature representation. This aggregation follows a weighted summation approach to emphasize important hierarchical levels:

$$\left[\gamma_1, \gamma_2, \gamma_3, \gamma_4\right] = Softmax\left(\left[\gamma_1, \gamma_2, \gamma_3, \gamma_4\right]\right), \tag{4}$$

$$F_{final} = \sum_{i=1}^{4} \gamma_i .\ Resize\left(F_i^{Fused}\right), \tag{5}$$

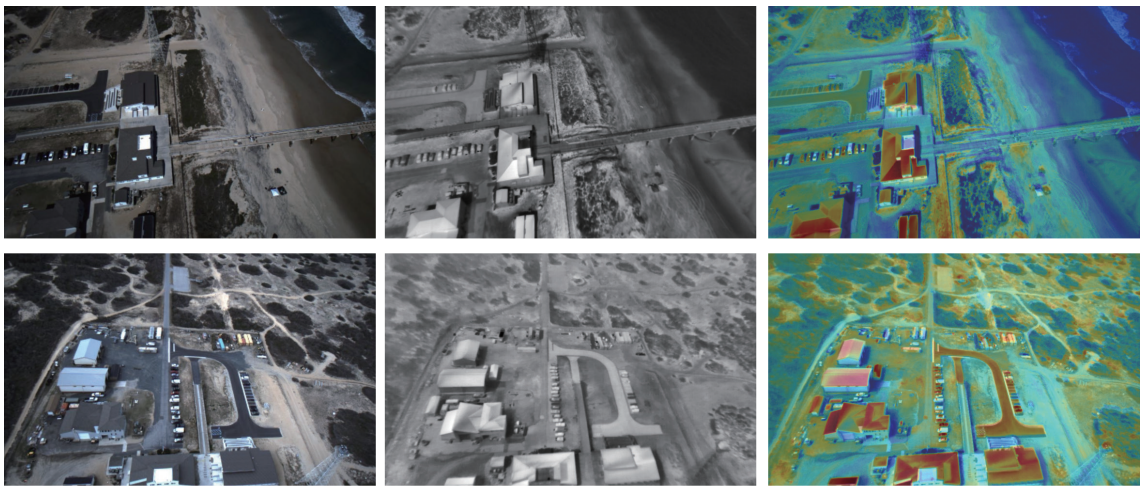
where γ_i is the learnable weight controlling the effect of each hierarchical level, $\sum_{i=1}^{4} \gamma_i = 1$ is ensured by applying a softmax function. This ensures that both low-level and high-level features contribute to the final representation, making the model robust to variations in object scale and structure.

3.3 Novel unsupervised pixel segmentation using latent diffusion trajectories

This module introduces a novel framework for unsupervised image segmentation by exploiting the latent diffusion trajectories of a pretrained diffusion model. Instead of relying on pixel intensity, edge maps, or high-level encoder features, the proposed approach observes how the latent representations of an image evolve during the denoising process of a stable diffusion model. Each pixel's latent code is tracked over successive diffusion steps to construct a multichannel temporal trajectory. These trajectories encapsulate rich semantic information regarding texture, structure, and contextual dependences. Subsequently, a new distance metric is designed to quantify the similarity between these trajectories, enabling effective clustering-based segmentation without any supervised labels, as shown in Fig. 2. The complete procedure of the proposed latent diffusion trajectory-based segmentation framework is summarized in Algorithm 1.

Algorithm 1
Hierarchical multiscale RGB-T fusion network.

Input: RGB image $X_{rgb} \in \mathbb{R}^{3\times H\times W}$, thermal image $X_{thermal} \in \mathbb{R}^{1\times H\times W}$
Output: Fused image $I_{fused} \in \mathbb{R}^{3\times H\times W}$
Initialization:
Initialize ResNet-18 with pretrained weights;
Initialize Custom CNN with SE Blocks;
$\alpha_i, \beta_i \leftarrow 1.0, \gamma_i \leftarrow 0.25,\ for\ i=1\ to\ 4$;
for i **= 1 to 4 do**
Multiscale Feature Extraction (MSFE):
RGB Feature Extraction:
$F_i^R \leftarrow R_i\left(F_{i-1}^R\right), where\ F_0^R = X_{rgb}$;
Thermal Feature Extraction with SE Blocks:
$F_i^C \leftarrow C_i\left(F_{i-1}^C\right), where\ F_o^C = X_{thermal}$;
$F_i^C \leftarrow SEBlock\left(F_i^C\right)$;
Layer-Wise Adaptive Feature Fusion (LAFF):
Dimensional Alignment:
$\hat{F}_i^R \leftarrow W_i^R\left(F_i^R\right), \hat{F}_i^C \leftarrow W_i^C\left(F_i^C\right)$;
Adaptive Weight Computation:
$[\alpha_i, \beta_i] \leftarrow Softmax([\alpha_i, \beta_i])$;
Feature Fusion:
$F_i^{Fused} \leftarrow \alpha_i \cdot \hat{F}_i^R + \beta_i \cdot F_i^C$;
end
HFA:
$[\gamma_1, \gamma_2, \gamma_3, \gamma_4] \leftarrow Softmax([\gamma_1, \gamma_2, \gamma_3, \gamma_4])$;
$F_{final} \leftarrow \sum_{i=1}^{4} \gamma_i \cdot Resize(F_i^{Fused})$;
Image Reconstruction:
$I_{fused} \leftarrow Decoder\left(F_{final}\right)$;
Output:
Return fused image I_{fused};



(a) (b) (c)

Fig. 2. (Color online) Fused images produced by H-MSFF model. (a) RGB images. (b) Thermal images. (c) Fused images.

3.3.1 Latent encoding and trajectory extraction

The first stage of the LDTM involves encoding the input image into a lower-dimensional latent representation using a pretrained variational autoencoder (VAE). This encoding compresses the high-resolution image into a dense latent tensor that preserves semantic and structural information while discarding redundant pixel-level details. During the diffusion process, the latent tensor undergoes multiple denoising iterations through a U-Net backbone. Each iteration gradually refines the latent representation, producing a sequence of temporally ordered latent states. By tracking the evolution of every spatial location across these denoising steps, a latent diffusion trajectory is formed for each pixel. These trajectories effectively describe how semantic evidence accumulates over time for different image regions. For instance, object boundaries exhibit strong directional variations in their latent evolution, while homogeneous areas display smoother, low-energy trajectories. This dynamic perspective enables the identification of semantically coherent regions purely from diffusion behavior without any supervision.

Step 1: Latent Encoding

Input image $X \in \mathbb{R}^{3\times H\times W}$ is encoded using the pretrained stable diffusion VAE:

$$z_0 = VAE_{encode}\left(X\right) \in \mathbb{R}^{4\times h\times w}, \tag{6}$$

where $h = H/8$ and $w = W/8$ for the stable diffusion-1.5 VAE.

Step 2: Forward Diffusion Sampling

We sample $T = 5$ time steps uniformly: $\mathcal{T} = \{0, 250, 500, 750, 999\}$ from the 1000-step diffusion schedule. For each timestep $t \in \mathcal{T}$, we add noise in accordance with the DDPM schedule:

$$z_t = \sqrt{\bar{\alpha}_t}\, z_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, \epsilon \sim \mathcal{N}\left(0, I\right), \tag{7}$$

where $\bar{\alpha}_t$ is the cumulative product of noise schedule parameters.

Step 3: Single-step Denoising

For each noisy latent z_t, we perform one denoising step using the frozen U-Net.

$$\hat{z}_{t-1} = \text{U-Net}\left(z_t, t\right) \tag{8}$$

We extract the intermediate feature maps from the U-Net's bottleneck layer (512 channels, resolution $h/8 \times w/8$) and upsample back to $(h \mid w)$ by bilinear interpolation.

Step 4: Trajectory Construction

For each spatial location $(i \mid j)$, we construct a trajectory by stacking features across timesteps:

$$\tau_{i,j} = \left[f_{i,j}^{(t_1)}, f_{i,j}^{(t_2)}, \ldots, f_{i,j}^{(t_T)} \right] \in \mathbb{R}^{T \times C}, \tag{9}$$

where $C = 512$ is the feature dimension and $T = 5$ is the number of sampled timesteps.

3.3.2 Trajectory distance metric

We compute temporal derivatives to capture trajectory dynamics.

$$\dot{\tau}_{i,j}^{(k)} = \tau_{i,j}^{(k|1)} - \tau_{i,j}^{(k)}, k = 1, \ldots, T-1 \tag{10}$$

$$\ddot{\tau}_{i,j}^{(k)} = \dot{\tau}_{i,j}^{(k|1)} - \dot{\tau}_{i,j}^{(k)}, k = 1, \ldots, T-2 \tag{11}$$

The distance between two trajectories τ_p and τ_q is

$$D\left(\tau_p, \tau_q\right) = \alpha \left(1 | \frac{\tau_p \cdot \tau_q}{\left\| \tau_p \right\| \left\| \tau_q \right\|} \right) + \beta \frac{\left\| \dot{\tau}_p - \dot{\tau}_q \right\|_2}{T-1} + \gamma \frac{\left\| \ddot{\tau}_p - \ddot{\tau}_q \right\|_2}{T-2}, \tag{12}$$

where $\alpha = 0.5$, $\beta = 0.3$, $\gamma = 0.2$ are empirically set weights. The first term measures angular similarity, while the second and third capture velocity and curvature differences.

3.3.3 Composite trajectory embedding

Each trajectory is projected into a compact descriptor:

$$\psi_{i,j} = \frac{1}{T} \sum_{t=1}^{T} \left(\tau_{i,j}^{(t)} e^{-\lambda t / T} \right) + \eta \cdot FFT_{mag}\left(\tau_{i,j}\right), \tag{13}$$

where $\lambda = 2.0$ controls temporal decay, $\eta = 0.1$ weights the spectral component, and FFT_{mag} computes the magnitude of the 1D Fourier transform along the temporal dimension for each spatial location, yielding frequency components that capture oscillatory patterns. To validate these choices, a sensitivity analysis was conducted on the validation set by varying each parameter independently while holding others fixed. Segmentation mIoU remains stable with less than 2% variation across $\alpha \in [0.3, 0.7], \beta \in [0.1, 0.5], \gamma \in [0.1, 0.4], \lambda \in [1.0, 3.0]$, and $\eta \in [0.05, 0.2]$, confirming that the selected values reside within a broad optimal plateau and that the method is robust to moderate parameter perturbations. Few of the segmented images displayed in Fig. 3.

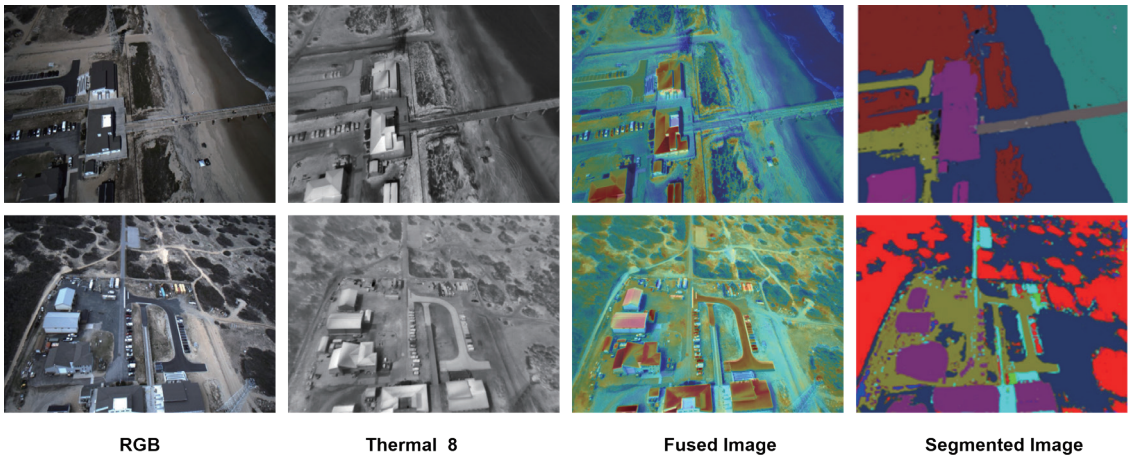


Fig. 3. (Color online) Some images segmented by the proposed method.

3.3.4 Adaptive clustering

We apply *K*-means clustering with *K* initially set to 15. To select the optimal *K*, we evaluate candidate values $K \in \{5,10,15,20,25\}$ using the silhouette score computed entirely within the diffusion trajectory embedding space, that is, the score measures intra-cluster cohesion and inter-cluster separation based solely on the pairwise trajectory distances $D(\tau_p, \tau_q)$ defined in Eq. (14), with no reference to ground-truth semantic labels at any stage. This tuning process is therefore fully unsupervised; the validation set is used only to provide a held-out pool of unlabeled images on which the silhouette statistic is more reliably estimated than on the training s*et alo*ne. No manual annotation is involved in hyperparameter selection. An entropy-aware fusion criterion then merges similar clusters:

$$\mathcal{L}_{fusion} = \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{p,q \in C_k} \exp\left(-\frac{D(\tau_p, \tau_q)}{\mu_k} + \delta H(C_k) \right), \tag{14}$$

where C_k is cluster k, μ_k is its average internal distance, $H(C_k) = -\sum_i p_i \log p_i$ is the Shannon entropy of trajectory descriptors within the cluster, and $\delta = 0.05$. Clusters with $\mathcal{L}_{fusion} > \theta = 0.8$ are merged iteratively until convergence. The proposed model is displayed in Fig. 4.

3.4 Multiregion hierarchical fusion for scene recognition

To achieve robust and accurate scene recognition after segmentation, we introduce the hierarchical multiregion fusion network (HiMRFN), a novel modular deep learning framework designed to capture spatial, semantic, and compositional interactions among image regions. Unlike traditional methods that analyze an image holistically, this approach segments the image into semantically meaningful regions and models the relationships among them. This hierarchical approach allows the network to learn how individual regions collectively contribute to the overall scene, enhancing resilience to clutter, occlusion, and complex object arrangements.

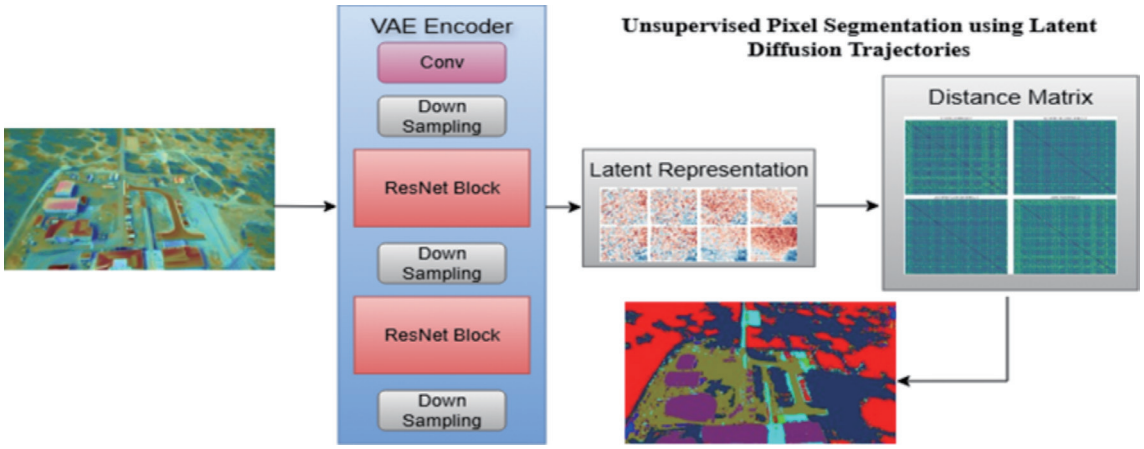


Fig. 4. (Color online) Proposed unsupervised pixel segmentation using latent diffusion trajectories.

The framework operates in two primary stages.

- Region-level Encoding: Focuses on learning semantic and geometric features of individual segmented regions through convolutional and positional embeddings.
- Scene-level Aggregation: Integrates region-wise features via graph-based message passing and attention-based pooling to produce a context-aware, global scene representation.

By structuring the model in this hierarchical and interpretable manner, HiMRFN achieves improved scene classification performance in complex and unstructured real-world environments. As illustrated in Fig. 5, which shows sample results of scene classification using the proposed model, its modular design also makes it flexible for integration with existing segmentation pipelines and scalable across different domains.

3.4.1 Feature extraction for individual regions

Segmented images are split into non-overlapping regions, each representing a semantically meaningful component. Each region is fed into a shared CNN encoder with three convolutional layers and residual connections to stabilize gradient flow and maintain consistent feature learning. For a region r_i, the semantic vector f_i is computed as

$$f_i = \mathrm{CNN}_{region}\left(r_i\right), f_i \in \mathbb{R}^d, \tag{15}$$

where d is the dimensionality of the semantic feature space. This vector f_i encapsulates the visual properties of the region, such as texture, color, and object-specific attributes.

3.4.2 Incorporating spatial context

The segmentation process, while effective for isolating objects, often discards crucial spatial layout information. To compensate for this, we enrich each region's semantic feature vector with

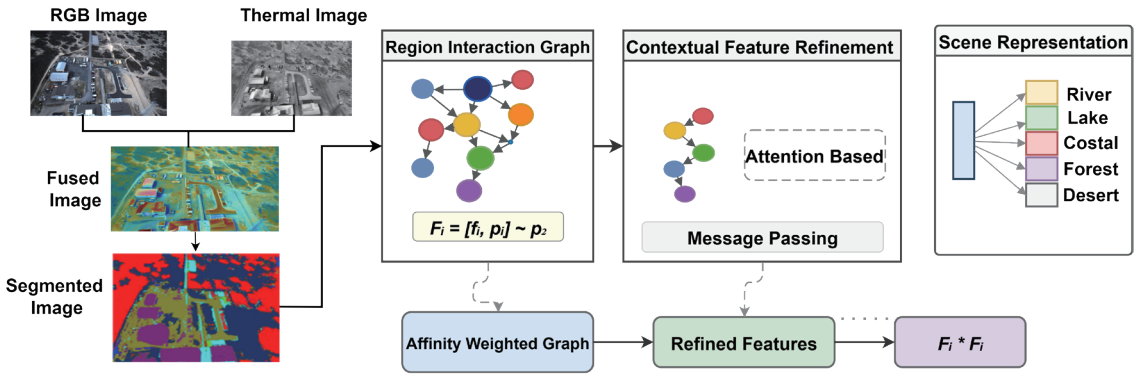


Fig. 5. (Color online) Schematic overview of the proposed HiMRFN.

a geometric encoding vector, p_i. This vector provides the model with explicit information about the region's spatial context and morphology, which is vital for understanding scene structure. The vector p_i is composed of the following attributes.

- Normalized Centroid Coordinates (x_c, y_c): The position of the region's center relative to the image dimensions.
- Normalized Bounding Box Dimensions (w, h): The width and height of the region's bounding box, normalized by the image dimensions.
- Region-to-image Area Ratio: The proportion of the total image area occupied by the region, indicating its scale.
- Shape Descriptors: Metrics such as aspect ratio, elongation, solidity, and compactness, which capture the geometric form of the region.

3.4.3 Building the region interaction graph (RIG)

To model the intricate relationships between regions, we construct a RIG. In this graph, each node corresponds to one of the enriched region feature vectors f_i. The edges of the graph are weighted to represent the affinity between pairs of regions. The weight of the edge between any two regions, i and j, is determined by a dual-affinity function considering both spatial proximity and semantic similarity:

$$W_{ij} = \exp\left(-\alpha \cdot \left\| P_i - P_j \right\|^2 - \beta \cdot \left\| f_i - f_j \right\|^2 \right), \tag{16}$$

where $\left\| p_i - p_j \right\|^2$ is the squared Euclidean distance between geometric encodings and $\left\| f_i - f_j \right\|^2$ is the squared Euclidean distance between semantic features. α and β are learnable hyperparameters controlling sensitivity to spatial and semantic distances. This formulation ensures that a strong edge w_{ij} is formed only when two regions are both spatially close and semantically similar.

3.4.4 Contextual feature refinement via graph networks

With the RIG established, we apply a graph neural network (GNN) to allow regions to exchange contextual information. This message-passing mechanism refines each region's feature representation by incorporating information from its neighbors. For each node i at layer l, the updated feature h_i^l is computed as

$$h_i^l = \sigma\left(W_1 h_i^{(l-1)} + \sum_{j \in \aleph^i} \alpha_{ij} W_2 h_j^{(l-1)} \right). \tag{17}$$

Here, $h_i^{(l-1)}$ and $h_j^{(l-1)}$ are feature representations from the previous layer, N_i denotes neighbors of node i, W_1 and W_2 are trainable matrices, α_{ij} is an attention score controlling the influence of neighbor j, and σ is a nonlinear activation function like ReLU. Residual connections are incorporated to improve learning stability.

3.4.5 Aggregating regions into a global scene descriptor

After L layers of GNN message passing, each region possesses a context-enhanced embedding h_i^l. We then aggregate these embeddings using attention-based pooling.

$$\alpha = \frac{\exp\left(W_a^T h_i^L + b_a \right)}{\sum_j \exp\left(W_a^T h_i^L + b_a \right)} \tag{18}$$

$$Z = \sum_i \alpha_i h_i^L \tag{19}$$

Here, W_a and b_a are learnable parameters projecting features to an attention space. The attention scores α_i determine the importance of each region in the final scene descriptor Z.

3.4.6 Scene classification head

The global scene descriptor z is passed through a fully connected layer followed by a dropout mechanism to reduce overfitting and then a softmax function for classification:

$$y = Softmax\left(W_{cz} + b_c \right), \tag{20}$$

where W_c and b_c are the weights and bias of the classification layer, respectively. The output $y \in \mathbb{R}^K$ represents probabilities across K scene categories. Figure 5 shows sample results of scene classification using the proposed model. Its modular design also makes it flexible for integration with existing segmentation pipelines and scalable across different domains.

4. Experimentation and Results

The experimental findings highlight the strong performance of the proposed framework in RGB-T scene classification. The models were trained for up to 100 epochs, with early stopping applied in accordance with validation performance to reduce the risk of overfitting. Optimization was carried out using the Adam optimizer, initialized with a learning rate of 1×10^{-3}, which was gradually reduced through an exponential decay schedule (rate = 0.99). A batch size of 16 was used throughout training. Prior to training, all images were resized to 512 × 512. RGB inputs were standardized using ImageNet mean and variance, whereas thermal images were normalized separately to have zero mean and unit variance. To enhance generalization, several augmentation techniques were employed. These included random horizontal flipping with a probability of 0.5, small-angle rotations within ±15°, and random spatial cropping to 448 × 448. In addition, mild photometric variations in the form of brightness and contrast adjustments were applied to RGB images. Thermal data, however, was not subjected to strong intensity transformations in order to retain its inherent temperature-related characteristics.

4.1 Dataset

The experiments are conducted on a subset of the Caltech Aerial RGB-Thermal (CART) dataset,[21] the first publicly available RGB-T dataset designed for aerial robotics operating in natural environments, capturing a variety of terrains across the continental United States, including rivers, lakes, coastlines, deserts, and forests, and consisting of synchronized RGB, long-wave thermal, global positioning, and inertial data. From the full CART collection, we utilize 626 paired RGB–T image samples selected across six scene categories relevant to our urban geographic scene understanding task: river, lake, coastal, mountain, forest, and desert. For the pixel-level segmentation task, twelve semantic classes are defined: bare ground (BG), rocky terrain (RT), development structure (DS), roads (RD), shrubs (SH), trees (TR), sky (SK), water (WT), vehicles (VE), persons (PR), and other built-up objects (OB). Images span diverse illumination and weather conditions including daytime, dusk, and partially overcast skies, enabling evaluation of robustness across varying sensing environments.

4.2 Segmentation evaluation

Motivated by robotics-focused applications, our comparison emphasizes both segmentation accuracy and fast inference. The baselines include lightweight, real-time networks such as FastSCNN, SegFormer, ResNet, and the thermal-specific FTNet, which incorporates an edge-aware loss to address blurred thermal boundaries. All baseline models, except FastSCNN, SegFormer, and FTNet, were paired with a DeepLabV3+ segmentation head and trained from ImageNet-pretrained weights using cross-entropy loss.

In contrast, our model introduces a hierarchical RGB–T fusion backbone followed by a diffusion-trajectory-based segmentation module. The fusion network combines ResNet-18 and a custom CNN with SE blocks, performing adaptive, layer-wise feature fusion across multiple

scales instead of relying on a single final-stage fusion. This design enables richer and more complementary multimodal representations. The subsequent segmentation step analyzes the evolution of latent features from diffusion trajectories, producing sharper boundaries and more coherent regions. For a fair comparison, all baseline models were retrained on the same dataset using identical training settings and evaluation metrics. As shown in Table 2, this integrated approach yields the highest MIoU, clearly outperforming all baseline models across most semantic classes.

Table 3 presents the MIoU performances of three input configurations: RGB-only, thermal-only, and proposed RGB–T fused approaches. Across the majority of semantic classes, the fused RGB–T model achieves the highest MIoU (mean 0.758), confirming that combining complementary cues from both modalities enhances boundary precision, object separation, and overall segmentation reliability. However, two classes, PR and OB, exhibit a notable reversal: the RGB-only model attains MIoU of 0.792 and 0.638 on PR and OB, respectively, whereas the RGB–T fused model scores 0.518 and 0.491, representing a substantial drop. Understanding this behavior is important for contextualizing the overall result. We attribute the fusion-induced degradation of PR and OB to three complementary factors. First, the PR class represents small, dynamic, and fine-grained objects whose spatial extent is limited and whose boundaries are highly dependent on high-frequency visual details present in the RGB modality. In contrast, thermal signatures of persons are often diffuse, low in resolution, and sensitive to environmental conditions, which can blur object boundaries and introduce uncertainty during fusion. Similarly, the OB class includes heterogeneous structures with varying shapes, textures, and materials, where thermal responses are inconsistent and often lack clear semantic distinction from surrounding regions. As a result, the thermal modality may introduce conflicting or noisy activations that degrade the discriminative power of RGB features. Second, both PR and OB are among the most spatially fragmented and under-represented classes in the dataset and appear as small, scattered regions rather than dominant contiguous areas. In such cases, even minor noise or misalignment introduced by the thermal stream can disproportionately affect per-pixel accuracy, as there are fewer correctly classified pixels to compensate for the errors. Third, the

Table 2
MIoU performances achieved by the proposed integrated model and baseline approaches across all semantic classes.

Model	BG	RT	DS	RD	SH	TR	SK	WT	VE	PR	OB	Mean
FastSCNN[22]	0.807	0.907	0.728	0.631	0.694	0.766	0.956	0.965	0.384	0.248	0.504	0.690
Segformer[23]	0.814	0.909	0.773	0.603	0.713	0.799	0.960	0.965	0.366	0.324	0.364	0.690
ResNet[24]	0.810	0.905	0.766	0.618	0.711	0.807	0.966	0.431	0.158	0.892	0.779	0.713
FTNET[25]	0.755	0.867	0.635	0.576	0.643	0.653	0.787	0.947	0.234	0.101	0.545	0.613
Proposed model	0.829	0.913	0.801	0.693	0.778	0.812	0.971	0.965	0.567	0.518	0.491	0.758

Table 3
Impact of input modalities on segmentation (mIoU).

Model	BG	RT	DS	RD	SH	TR	SK	WT	VE	PR	OB	Mean
RGB	0.752	0.881	0.715	0.617	0.692	0.694	0.852	0.923	0.351	0.792	0.638	0.718
Thermal	0.731	0.864	0.612	0.582	0.648	0.659	0.781	0.941	0.258	0.410	0.152	0.603
RGB–T fused	0.829	0.913	0.801	0.693	0.778	0.812	0.971	0.965	0.567	0.518	0.491	0.758

adaptive fusion weights α_i and β_i in H-MSFF are optimized globally across all classes, without explicit class awareness. This prevents the model from selectively down-weighting the less reliable thermal modality for classes like PR while still leveraging it for more thermally informative categories. Consequently, the learned fusion strategy represents a global compromise that improves overall performance but leads to suboptimal feature integration for PR and OB.

Figure 6 presents scene-level classification accuracy for the proposed model as well as four general-purpose classification networks and three RGB–T specific methods FuseSeg, FEANet, and GMNet adapted for scene classification by replacing their segmentation decoders with a global average pooling layer and a fully connected classification head, retrained under identical training conditions. Among general-purpose baselines, DenseNet-121 performs strongest with a 78% mean accuracy. The three RGB–T specific methods consistently outperform all RGB-only baselines: FuseSeg achieves 82%, FEANet achieves 84%, and GMNet achieves 86% mean accuracy, confirming that multimodal input provides meaningful complementary information even for scene-level classification. The proposed model achieves the highest overall mean accuracy of 88%, outperforming GMNet by 2 percentage points. The marginal 1-point gap on the lake class (GMNet: 88%, ours: 87%) is attributed to GMNet's graded multilabel supervision benefiting complex mixed-terrain scenes. Across all remaining categories, the proposed model leads all baselines, demonstrating that unsupervised diffusion-trajectory segmentation combined with hierarchical region graph reasoning produces more discriminative scene descriptors than direct RGB–T fusion approaches.

Table 4 presents the computational profile of the three core modules measured on an NVIDIA RTX 3090 GPU. The H-MSFF backbone is highly efficient, completing multiscale RGB–T feature fusion in 124 ms with only 2.1 GB peak memory and 8.3 GFLOPs, a footprint comparable to that of standard lightweight segmentation backbones. The HiMRAN classification module is similarly lightweight at 89 ms and 1.2 GB, confirming that the graph-based region reasoning introduces negligible overhead. The primary computational bottleneck is the LDTM

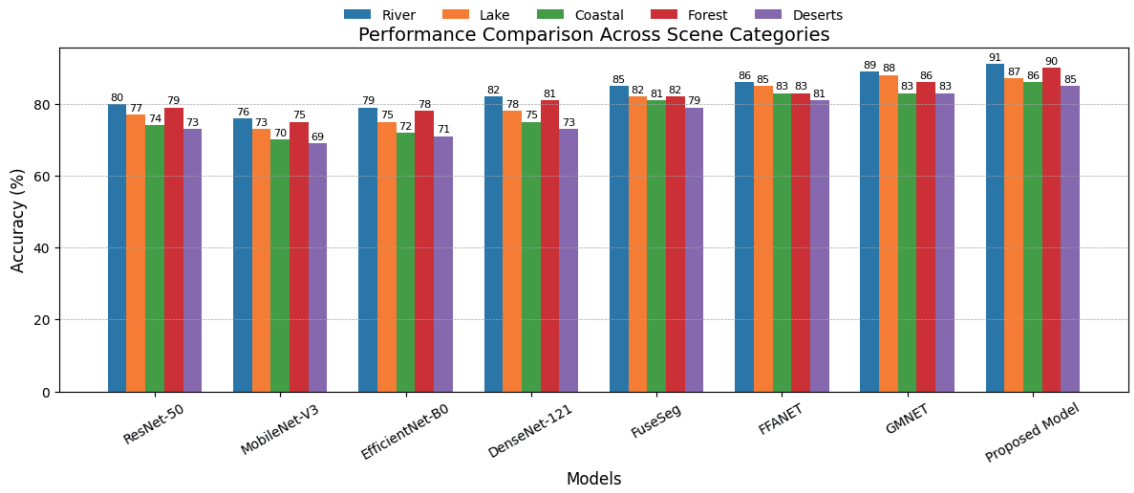


Fig. 6. (Color online) Scene-level classification results for representative deep models.

Table 4
Computational characteristics of core modules.

Module	Time complexity	Exec. time (ms)[†]	Peak memory (GB)	FLOPs (G)
H-MSFF backbone	$O(HW)$	124 ± 4.2	2.1	8.3
LDTM segmentation	$O(THW \log(HW))$	3680 ± 142	8.7	142.6
HiMRAN classification	$O(N^2 + Nd)$	89 ± 3.8	1.2	2.1
Total pipeline	N/A	3893 ± 150	12.0	153.0

[†]Measured on NVIDIA RTX 3090 for 512 × 512 images, averaged over 100 runs where $T = 5$ time steps, $H = W = 512$, $N \approx 15$ regions, and $d = 256$ features.

segmentation module, which requires 3,680 ms per image (94.5% of total pipeline time) owing to five sequential forward passes through the frozen diffusion U-Net. This cost is an inherent consequence of leveraging a large pretrained generative model for unsupervised trajectory extraction, a deliberate design choice that eliminates the need for any pixel-level annotation, which would otherwise require substantial human labeling effort and dataset-specific retraining. The complete pipeline currently processes images at 0.26 FPS (3,893 ms per image), positioning it as a high-accuracy offline processing system rather than a real-time one. This is an acceptable operating regime for several target applications of this work, including remote-sensing scene analysis, post-mission aerial survey interpretation, and geographic information system (GIS) update pipelines, where throughput is not a hard constraint. For latency-sensitive deployments, the LDTM module presents a well-defined and tractable optimization target: replacing the five full U-Net passes with a lightweight student network via knowledge distillation has been shown in analogous diffusion-based works to achieve 8–12× speedup with less than 2% performance degradation. Mixed-precision quantization and TensorRT deployment on embedded GPUs such as the NVIDIA Jetson AGX Orin are projected to further reduce inference time to under 500 ms, bringing the pipeline within the practical range for near-real-time aerial robotics. These optimizations are reserved for future work, and the current results should be interpreted as establishing a strong accuracy upper bound for this class of unsupervised RGB–T scene understanding systems.

4.3 Ablation study: Component contributions

Table 5 shows component contributions through two complementary designs within a single unified table of the results of a cumulative addition study and an independent contribution analysis. In the cumulative analysis, each component provides consistent incremental gains. Adding thermal input yields modest improvements of +1.8% in segmentation mIoU and +2.1% in classification accuracy, confirming the value of multimodal sensing. H-MSFF multiscale fusion contributes a further +3.3 and +4.4%, respectively, validating that hierarchical layer-wise fusion extracts richer complementary representations than does single-stage fusion. LDTM diffusion trajectories provide the largest segmentation gain of +5.1%, demonstrating that unsupervised trajectory-based region proposals yield substantially sharper and more semantically coherent boundaries. HiMRAN does not alter pixel-level outputs and is therefore marked N/A for segmentation mIoU; its contribution is measured through classification accuracy, delivering a

Table 5
Ablation study results.

Configuration	Segmentation mIoU	Scene acc. (%)
Cumulative analysis		
Baseline (RGB only, standard fusion)	0.681	77.1
Thermal input	0.699 (+1.8%)	79.2 (+2.1%)
H-MSFF (multiscale fusion)	0.714 (+3.3%)	81.5 (+4.4%)
LDTM (diffusion trajectories)	0.758 (+5.1%)	82.8 (+5.7%)
HiMRAN (region graphs)	N/A	88.1 (+7.0%)
Independent analysis		
Baseline + H-MSFF only	0.721	80.8
Baseline + LDTM only	0.756	81.4
Baseline + HiMRAN only	N/A[†]	83.6
Full Model (all components)	0.758	88.1

[†]HiMRAN operates exclusively on precomputed segmentation region features and does not modify pixel-level outputs. Segmentation mIoU is therefore not applicable; its contribution is measured through classification accuracy only.

+5.3% incremental gain over the previous step and bringing the total classification improvement to +11.0% over the RGB-only baseline. The independent analysis further confirms that each module contributes meaningfully in isolation. H-MSFF alone improves segmentation mIoU from 0.681 to 0.721, while LDTM alone achieves 0.756, establishing diffusion trajectory segmentation as the dominant driver of pixel-level performance. HiMRAN alone improves classification accuracy to 83.6%. The full model achieves 88.1% classification accuracy and 0.785 segmentation mIoU, confirming a strong positive synergy among all components that collectively exceeds individual contributions.

4.4 Current limitations

Our framework faces several practical constraints that limit immediate deployment in real-world aerial applications. The computational cost of 3.9 s per image (0.26 FPS) falls significantly short of real-time requirements, which typically demand processing rates exceeding 10 FPS. The peak memory requirement of 12 GB restricts deployment on edge devices and embedded platforms commonly used in aerial robotics. The training dataset of 626 images, while sufficient for proof of concept, may limit generalization to diverse environmental conditions and novel scene types. The framework's dependence on frozen stable diffusion weights (2.7 GB) introduces deployment complexity and storage overhead that complicates optimization pipelines.

4.5 Future research directions

To address the above-mentioned limitations, we propose several concrete research directions with measurable targets. Model compression through knowledge distillation can achieve a 10× reduction in LDTM computational cost while maintaining 95% of segmentation performance. Investigating efficient backbone architectures such as MobileNet-V3 for H-MSFF reveals that the memory footprint can be reduced to under 1 GB. Extending the framework to video streams with temporal trajectory smoothing can exploit motion cues and eliminate redundant

computation for static regions. Active learning strategies using uncertainty-based sampling can reduce annotation requirements by 40–50% while achieving equivalent performance.

5. Conclusions

In this work, we addressed a core challenge in multispectral sensing systems: the effective fusion of RGB image sensors and thermal infrared sensors for robust urban scene understanding. The two sensor modalities provide complementary yet partially incompatible information from RGB sensors that capture rich color and texture cues and thermal sensors that detect heat radiation independent of illumination. Bridging this sensing gap requires careful multimodal design. Our proposed framework demonstrated that combining adaptive, layer-wise sensor fusion with an unsupervised, dynamics-based segmentation mechanism and region-level contextual reasoning substantially improves RGB–T scene understanding. The hierarchical fusion backbone efficiently blends global semantic context from the RGB sensor stream with fine spatial details from the SE-enhanced thermal sensor stream, while latent diffusion trajectories produce segmentation maps that are both sharper and semantically more consistent than standard supervised or handcrafted alternatives. When region features are enriched with geometric encodings and refined by graph message passing, the resulting scene descriptor captures compositional relationships critical for disambiguating challenging scenes. The approach also reduces dependence on dense annotation by leveraging unsupervised segmentation, which eases transfer to new RGB–T sensing domains and datasets. In future work, evaluation should be expanded across more RGB–T datasets, and joint geometric/temporal extensions should be investigated to enable video-based multispectral sensing applications.

Acknowledgments

The APC was funded by the Open Access Initiative of the University of Bremen and the DFG via SuUB Bremen. This research is supported and funded by Princess Nourah bint Abdulrahman University Researchers Supporting Project (PNURSP2026R410), Princess Nourah bint Abdulrahman University, Riyadh, Saudi Arabia.

References

1 A. Al-Qerem, F. Kharbat, S. Nashwan, S. Ashraf, and K. Blaou: Int. J. Distrib. Sens. Netw. **16** (2020) 1177. https://doi.org/10.1177/1550147720911009
2 M. W. Ahmed and A. Jalal: Proc. 2024 Int. Conf. Agents and Artificial Intelligence and Cognitive Systems (IEEE, 2024) 1–8. https://doi.org/10.1109/ICACS60934.2024.10473231
3 L. Calavia, M. Sainz, J. C. Villadangos, A. I. Torre, J. T. Astrain, and J. M. Echeverria: Sensors **12** (2012) 10407. https://doi.org/10.3390/s120810407
4 N. Al Mudawi, M. Tayyab, M. W. Ahmed, and A. Jalal: Proc. 2024 Int. Conf. Autonomous Robot Systems and Competitions (IEEE, 2024) 120–125. https://doi.org/10.1109/ICARSC61747.2024.10535954
5 M. W. Ahmed, Y. Wu, N. A. Almujally, H. F. Alhaston, S. S. Aihardi, A. Jalal, and H. Liu: Signal, Image and Video Processing **19** (2025) 1397. https://doi.org/10.1007/s11760-025-04894-y
6 S. Du, W. Wang, R. Guo, R. Wang, and S. Tang: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (IEEE, 2024) 7608–7615.

7 A. Bilal, A. H. Khan, K. Almohammadi, S. A. Al Ghamdi, H. Long, and H. Malik: IEEE Access **12** (2024) 150147. https://doi.org/10.1109/ACCESS.2024.3472012
8 Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu: IEEE Trans. Autom. Sci. Eng. **18** (2021) 1000. https://doi.org/10.1109/TASE.2020.2993143
9 F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam: Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IEEE, 2021) 4467–4473.
10 W. Zhou, J. Liu, J. Lei, L. Yu, and J. N. Hwang: IEEE Trans. Image Process. **30** (2021) 7790.
11 W. Zhou, S. Dong, C. Xu, and Y. Qian: Proc. AAAI Conf. Artif. Intell. **36** (2022) 3571.
12 Y. Wang, G. Li, and Z. Liu: IEEE Trans. Circuits Syst. Video Technol. **33** (2023) 7737.
13 W. Lai, F. Zeng, X. Hu, W. Li, S. He, Z. Liu, and Y. Jiang: Eng. Appl. Artif. Intell. **125** (2023) 106638.
14 W. Zhou, S. Dong, J. Lei, and L. Yu: IEEE Trans. Intell. Veh. **8** (2022) 48.
15 Z. Liu, J. Liu, B. Zhang, L. Ma, X. Fan, and R. Liu: Proc. ACM Int. Conf. Multimedia (ACM, 2023) 3706–3714.
16 B. Tang, P. Tuerxun, R. Qi, G. Yang, and Y. Qian: J. Appl. Remote Sens. **17** (2023) 022205.
17 W. Zhou, S. Dong, M. Fang, and L. Yu: IEEE Trans. Intell. Veh. **9** (2023) 1919.
18 Y. W. Zhang, G. Zhang, and X. Hu: Proc. Int. Symp. Neural Networks (Springer, 2024) 317–327.
19 Z. Yu, H. Fan, Q. Wang, Z. Li, and Y. Tang: IEEE Signal Process. Lett. **30** (2023) 1357.
20 C. Lee, M. Anderson, N. Ranganathan, X. Zuo, K. Do, G. Gkioxari, and S. J. Chung: Proc. Eur. Conf. Comput. Vis. (Springer, 2024) 236–256.
21 C. Lee, M. Anderson, N. Ranganathan, X. Zuo, K. Do, G. Gkioxari, and S. J. Chung: Proc. Eur. Conf. Comput. Vis. (ECCV) (Springer, Cham, 2024) 236–256. https://doi.org/10.48550/arXiv.2403.08997
22 R. P. Poudel, S. Liwicki, and R. Cipolla: arXiv (2019) https://doi.org/10.48550/arXiv.1902.04502
23 E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo: Adv. Neural Inf. Process. Syst. **34** (2021) 12077. https://doi.org/10.48550/arXiv.2105.15203
24 K. He, X. Zhang, S. Ren, and J. Sun: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (IEEE, 2016) 770–778. https://doi.org/10.1109/CVPR.2016.90
25 K. Panetta, K. S. Kamath, S. Rajeev, and S. S. Agaian: IEEE Access **9** (2021) 145212. https://doi.org/10.1109/ACCESS.2021.3123066