BEYOND UNIDIRECTIONAL PRIOR INJECTION: BIDIRECTIONAL SEMANTIC ALIGNMENT FOR RGB-T SEMANTIC SEGMENTATION

Anonymous authors
Paper under double-blind review

ABSTRACT

RGB-T semantic segmentation benefits from the complementary information of visible and thermal images, while pre-trained large-scale vision models (LVMs) provide rich semantic priors for scene understanding. However, transferring RGB-pretrained semantics to RGB-T representations is nontrivial because the pre-trained semantic space is optimized for visible-image distributions and may be mismatched with heterogeneous multimodal features. To address this issue, we propose a Bidirectional Semantic Alignment Network (BiSANet), which enables target-aware interaction between RGB-T representations and pre-trained semantic priors. A Hierarchical Dual-Modal Encoder first constructs stable and complementary RGB-T features. Bidirectional Semantic Adapters then inject target-domain information into the frozen semantic pathway and transfer the adapted priors back to the multimodal branch for semantic correction. Finally, a Semantic Enhancement Decoder integrates multi-level features for fine-grained prediction. BiSANet achieves state-of-the-art performance on MFNet, PST900, and FMB, demonstrating effective semantic-prior adaptation for RGB-T segmentation.

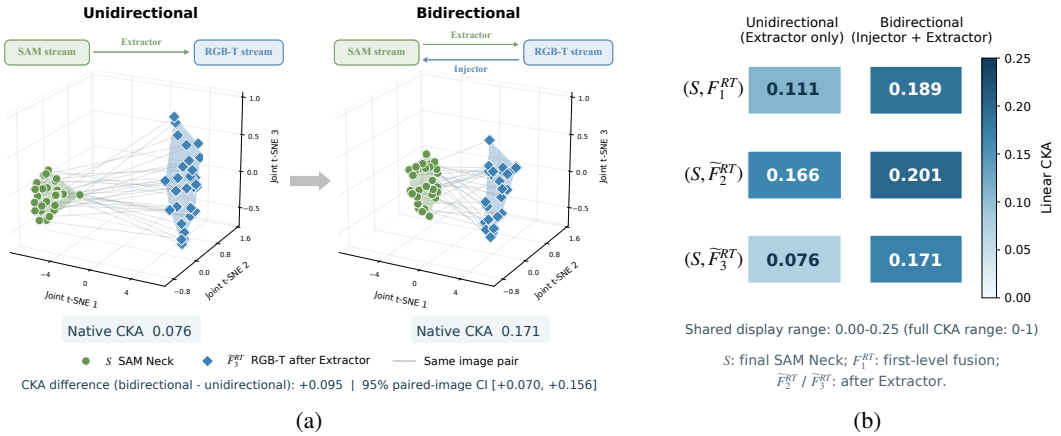
1 INTRODUCTION

Semantic segmentation assigns a semantic label to each pixel and is widely used in autonomous driving, robotic perception, intelligent surveillance, and remote sensing (Xie et al., 2021; Zheng et al., 2021b). Recent studies have increasingly exploited semantic priors learned from large-scale visual data to improve dense prediction (Hamilton et al., 2022; Cho et al., 2024; Xie et al., 2024b). Large-scale supervised models and self-supervised models such as DINO (Caron et al., 2021) and MoCo (Chen et al., 2021) learn rich and transferable knowledge for object localization, boundary perception, and semantic discrimination. RGB-T semantic segmentation further benefits from the complementary properties of visible and thermal images: RGB images provide rich appearance and structural cues, whereas thermal images offer relatively stable responses under low illumination, occlusion, and appearance degradation (Dong et al., 2023; Zhou et al., 2023). Therefore, transferring large-scale semantic priors to RGB-T representations provides a promising way to enhance multimodal scene understanding.

However, effectively transferring such priors to RGB-T segmentation remains challenging. Existing semantic priors are predominantly learned from RGB images and are therefore biased toward the visible-image distribution, while RGB-T representations combine heterogeneous information from different imaging mechanisms. Existing RGB-T methods typically introduce semantic knowledge through weight initialization, teacher supervision, or auxiliary features, resulting in largely unidirectional knowledge transfer (Liu et al., 2025; Zhang et al., 2026). Such strategies allow pre-trained semantics to guide RGB-T representations, but do not explicitly adapt the semantic-prior space to the RGB-T target domain. Consequently, unidirectional transfer may leave a semantic mismatch between RGB-pretrained priors and RGB-T features. This motivates a more interactive mechanism in which RGB-T features can adapt pre-trained semantics, while the adapted priors can in turn refine multimodal representations. As shown in Fig. 1, bidirectional adaptation improves their alignment.

Motivated by this observation, we propose a Bidirectional Semantic Alignment Network (BiSANet) for RGB-T semantic segmentation. Unlike conventional unidirectional prior transfer, BiSANet es-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) (b)

Figure 1: Empirical evidence for bidirectional semantic alignment based on experimentally extracted features. (a) Joint t-SNE distributions of S and $\tilde{F}_3^{RT}$ under unidirectional and bidirectional adaptation. (b) Linear CKA (Kornblith et al., 2019) between S and multi-level RGB-T features; higher values indicate greater representation similarity.

tablishes a learnable bidirectional bridge between RGB-T representations and RGB-pretrained semantic priors. A Hierarchical Dual-Modal Encoder first constructs stable multimodal representations through level-specific fusion: difference-aware low-level fusion preserves spatial details, while adaptive high-level fusion selects reliable modality information to alleviate cross-modal conflicts. A frozen Segment Anything Model (SAM) provides large-scale semantic priors, and the proposed Bidirectional Semantic Adapters enable mutual interaction between the two representation spaces. Specifically, RGB-T features are injected into the SAM pathway to adapt semantic priors toward the target domain, while the adapted semantics are transferred back to correct multimodal features. Finally, a Semantic Enhancement Decoder integrates low-level details, corrected RGB-T representations, and high-level semantic priors for fine-grained prediction. BiSANet achieves state-of-the-art performance on MFNet, PST900, and FMB, demonstrating the effectiveness of bidirectional semantic alignment for RGB-T segmentation.

2 Related Work

2.1 RGB Semantic Segmentation

RGB semantic segmentation assigns a semantic category to each image pixel. Early methods (Boykov et al., 2002; Starck et al., 2005) relied on handcrafted features and low-level cues, with limited representation capability. FCN (Long et al., 2015) enabled end-to-end prediction, while DeconvNet (Noh et al., 2015) and U-Net (Ronneberger et al., 2015) used encoder-decoder architectures to recover spatial details. Later methods focused on contextual and semantic reasoning. DeepLab (Chen et al., 2017; 2018), PSPNet (Zhao et al., 2017), and OCRNet (Yuan et al., 2020) exploited multi-scale, global, and object-level context. More recently, Transformer-based methods such as SETR (Zheng et al., 2021a), SegFormer (Xie et al., 2021), and Mask2Former (Cheng et al., 2022) have enhanced global semantics through self-attention.

Large-scale Vision Models (LVMs), including MoCo (Chen et al., 2021), DINO (Oquab et al., 2023), SAM (Kirillov et al., 2023), and CLIP (Radford et al., 2021), provide transferable priors for localization and regional consistency. DenseCL (Wang et al., 2021) and STEGO (Hamilton et al., 2022) leverage self-supervised representations for dense prediction and unsupervised segmentation, while CAT-Seg (Cho et al., 2024), SED (Xie et al., 2024a), and PA-SAM (Xie et al., 2024b) adapt CLIP or SAM priors to segmentation tasks. Nevertheless, RGB-only models remain vulnerable to illumination changes and appearance degradation, motivating RGB-T segmentation.

2.2 RGB-T Semantic Segmentation

RGB-T semantic segmentation combines RGB appearance cues with stable thermal responses to exploit cross-modal complementarity. MFNet (Ha et al., 2017) first demonstrated end-to-end RGB-T

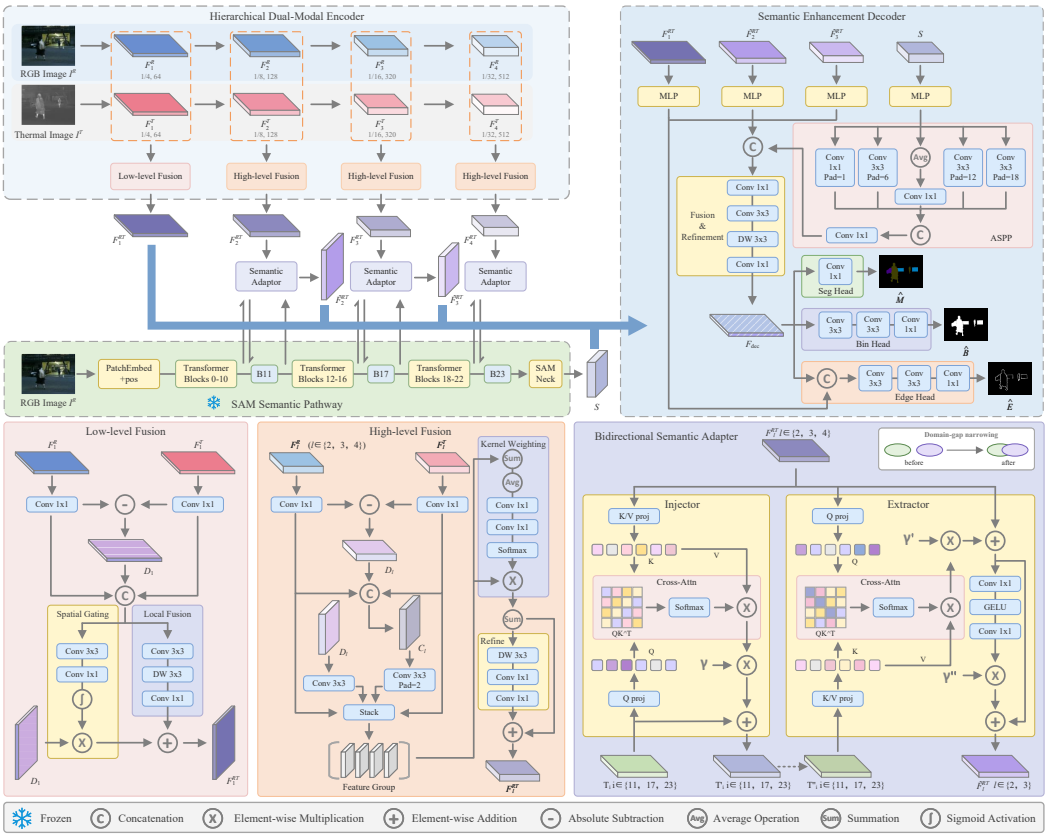


Figure 2: Overall architecture of the proposed BiSANet.

segmentation, followed by RTFNet (Sun et al., 2019) and FuseSeg (Sun et al., 2021), which hierarchically integrated thermal features into RGB encoders. However, these methods mainly relied on same-level fusion with limited modeling of cross-modal conflicts. Subsequent methods introduced discrepancy modeling and semantic constraints: ABMDRNet (Zhang et al., 2021) reduces modality shifts, EGFNet (Dong et al., 2023) exploits edge priors, LASNet (Li et al., 2023) adopts level-specific fusion, and CAINet (Lv et al., 2024) combines global context with auxiliary tasks.

Recent studies further introduce semantic priors through teacher networks or vision foundation models. CLNet (Guo et al., 2024) and KDSNet (Zhou et al., 2024) transfer semantic or structural knowledge from task-specific teachers through knowledge distillation, but their priors remain constrained by specific training data. MiLNet (Liu et al., 2025) employs the SAM encoder as a foundation-model teacher, while IG-SAM (Zhang et al., 2026) transfers SAM priors through multi-stage distillation and implicit prior fusion. Despite these advances, such priors are predominantly learned from RGB images and are not inherently adapted to thermal representations.

3 Proposed Method

As illustrated in Fig. 2, BiSANet contains a Hierarchical Dual-Modal Encoder (HDE), three Bidirectional Semantic Adapters (BSAs), and a Semantic Enhancement Decoder (SED). Given aligned RGB and thermal images I^R and I^T, HDE produces multi-scale RGB-T representations $\{F_l^{RT}\}_{l=1}^4$. Meanwhile, a frozen Segment Anything Model (SAM) (Kirillov et al., 2023), taking only I^R as input, provides RGB-pretrained semantic priors. The features F_2^{RT}, F_3^{RT}, and F_4^{RT} interact with the 11-th, 17-th, and 23-rd SAM blocks through BSAs, respectively, enabling target-domain information to adapt the semantic pathway and the adapted priors to correct multimodal representations. Finally, SED integrates F_1^{RT}, corrected features $\tilde{F}_2^{RT}$ and $\tilde{F}_3^{RT}$, and the adapted high-level semantic feature S for pixel-level prediction. It produces three predictions $\hat{M}$, $\hat{B}$, and $\hat{E}$. All predictions are supervised during training, while only $\hat{M}$ is retained during inference.

3.1 HIERARCHICAL DUAL-MODAL ENCODER

A two-stream MiT-B3 backbone (Xie et al., 2021) extracts RGB and thermal features F_l^R and F_l^T, where $l \in \{1, 2, 3, 4\}$. Since cross-modal complementarity varies across feature levels, we adopt level-specific fusion to construct stable RGB-T representations for subsequent semantic alignment.

At the low level, projected RGB and thermal features are combined with their absolute difference:

$$\begin{aligned} \bar{F}_1^* &= \mathrm{Conv}_{1\times1}(F_1^*), \\ D_1 &= \left|\bar{F}_1^R - \bar{F}_1^T\right|. \end{aligned} \tag{1}$$

Subsequently, the three features are concatenated along the channel dimension:

$$C_1 = \left[\bar{F}_1^R, \bar{F}_1^T, D_1\right]. \tag{2}$$

Then, a local branch extracts complementary information, while a spatial gate selects useful cross-modal differences:

$$F_1^{RT} = \phi_{\mathrm{local}}(C_1) + \sigma\left(\omega(C_1)\right) \odot D_1, \tag{3}$$

where $\phi_{\mathrm{local}}(\cdot)$ contains a 3×3 convolution, a depth-wise convolution, and a 1×1 convolution. $\sigma(\cdot)$ denotes the Sigmoid function, and the gating function $\omega(\cdot)$ is implemented by 3×3 and 1×1 convolutions.

For $l \in \{2, 3, 4\}$, we construct four candidate representations, covering modality-specific, discrepancy, and mixed information:

$$B_l^k \in \{\bar{F}_l^R, \psi_{\mathrm{diff}}(D_l), \psi_{\mathrm{mix}}(C_l), \bar{F}_l^T\} \tag{4}$$

where $\psi_{\mathrm{diff}}(\cdot)$ and $\psi_{\mathrm{mix}}(\cdot)$ are implemented by standard and dilated 3×3 convolutions, respectively. Their channel-wise reliability weights are estimated by:

$$\begin{aligned} B_l &= \mathrm{GAP}\left(\sum_{k=1}^{4} B_l^k\right), \\ W_l &= \mathrm{Softmax}\left(\mathrm{Conv}_{1\times1}\left(\mathrm{Conv}_{1\times1}(B_l)\right)\right), \end{aligned} \tag{5}$$

where $\mathrm{GAP}(\cdot)$ means global average pooling. The fused representation is:

$$F_l^{RT} = F_l' + \psi_{\mathrm{refine}}(F_l'), \qquad F_l' = \sum_{k=1}^{4} W_l^k \odot B_l^k. \tag{6}$$

where $\psi_{\mathrm{refine}}(\cdot)$ contains a 3×3 depth-wise convolution and two 1×1 convolutions.

Thus, HDE preserves low-level structures while adaptively selecting reliable higher-level multi-modal information, providing stable target representations for semantic alignment.

3.2 BIDIRECTIONAL SEMANTIC ADAPTER

Although HDE constructs complementary RGB-T representations, they remain mismatched with the RGB-pretrained semantic space of SAM. We therefore introduce a Bidirectional Semantic Adapter consisting of an injector and an extractor. The injector enables SAM to perceive RGB-T target-domain cues, while the extractor uses the adapted SAM priors to correct multimodal features.

The adapters are inserted into the 11-th, 17-th, and 23-rd Transformer blocks. Let T_i, $i \in \{11, 17, 23\}$, denote the corresponding SAM feature. For clarity, we define the cross-attention as:

$$\mathcal{A}(Q, K, V) = \mathrm{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right)V, \tag{7}$$

where d is the projected channel dimension. Features are projected into a shared embedding space before interaction, with pooling optionally applied to reduce computation.

For the injector, T_i serves as the query and the corresponding RGB-T feature provides the key and value. The three SAM features are then updated respectively as follows:

$$\begin{cases} T_{11}' = T_{11} + \gamma_1 \mathcal{A}(T_{11}, F_2^{RT}, F_2^{RT}), \\ T_{17}' = T_{17} + \gamma_2 \mathcal{A}(T_{17}, F_3^{RT}, F_3^{RT}), \\ T_{23}' = T_{23} + \gamma_3 \mathcal{A}(T_{23}, F_4^{RT}, F_4^{RT}). \end{cases} \tag{8}$$

where γ_i is a learnable scaling parameter initialized to zero to preserve the original SAM representation at the beginning of training. The updated features are further processed by their corresponding Transformer blocks, producing $\bar{T}_i$.

The extractor uses RGB-T features as queries and adapted SAM features as keys and values:

$$\begin{cases} \bar{F}_2^{RT} = F_2^{RT} + \gamma_1' \mathcal{A}(F_2^{RT}, \bar{T}_{11}, \bar{T}_{11}), \\ \bar{F}_3^{RT} = F_3^{RT} + \gamma_2' \mathcal{A}(F_3^{RT}, \bar{T}_{17}, \bar{T}_{17}). \end{cases} \tag{9}$$

A lightweight refinement branch further improves local semantic consistency:

$$\tilde{F}_l^{RT} = \bar{F}_l^{RT} + \gamma'' \phi_{\text{refine}}\left(\bar{F}_l^{RT}\right), \quad l \in \{2, 3\}, \tag{10}$$

where $\phi_{\text{refine}}(\cdot)$ consists of two 1×1 convolutions and a GELU activation. Both γ_i' and γ'' are initialized to zero.

The low-level feature F_1^{RT} bypasses semantic adaptation to preserve spatial details. At the highest level, F_4^{RT} is injected into SAM without an additional extraction step; the adapted semantic pathway is propagated through the SAM neck to produce S. In this way, the model preserves low-level details, corrects middle-level multimodal features using the SAM semantic space, and introduces the final adapted SAM semantic prior into the decoding stage.

3.3 Semantic Enhancement Decoder

SED integrates F_1^{RT}, $\tilde{F}_2^{RT}$, $\tilde{F}_3^{RT}$, and S to generate pixel-level predictions. Firstly, SegFormer-style MLP layers project them into a unified embedding space, producing $\{P_l\}_{l=1}^4$. To enhance the high-level semantic feature, an ASPP module is applied to P_4:

$$\bar{P}_4 = \text{Conv}_{1\times1}([\psi_1(P_4), \psi_6(P_4), \psi_{12}(P_4), \psi_{18}(P_4), \text{Conv}_{1\times1}(\text{GAP}(P_4))]), \tag{11}$$

where $\psi_1(\cdot)$ is a 1×1 convolution, and $\psi_6(\cdot)$, $\psi_{12}(\cdot)$, and $\psi_{18}(\cdot)$ are 3×3 atrous convolutions with dilation rates of 6, 12, and 18. The decoder feature is obtained as:

$$F_{\text{dec}} = \phi_{\text{FR}}([P_1, P_2, P_3, \bar{P}_4]), \tag{12}$$

where $\phi_{\text{FR}}(\cdot)$ consists of 1×1, 3×3, depth-wise 3×3, and 1×1 convolutions. Finally, three prediction heads are constructed:

$$\hat{M} = \phi_{\text{seg}}(F_{\text{dec}}), \qquad \hat{B} = \phi_{\text{bin}}(F_{\text{dec}}), \qquad \hat{E} = \phi_{\text{edge}}([P_1, F_{\text{dec}}]). \tag{13}$$

Here, $\hat{M} \in \mathbb{R}^{C\times H\times W}$ denotes the semantic prediction, whereas $\hat{B}, \hat{E} \in \mathbb{R}^{2\times H\times W}$ represent foreground/background and boundary predictions.

During training, the three prediction branches are jointly optimized. The semantic segmentation prediction $\hat{M}$ is supervised by the ground-truth mask M using weighted cross-entropy $\mathcal{L}_{\text{CE}}$ and Lovász-Softmax loss $\mathcal{L}_{\text{Lov}}$ Berman et al. (2018). The auxiliary foreground/background prediction $\hat{B}$ and boundary prediction $\hat{E}$ are respectively supervised by their ground-truth masks B and E through cross-entropy loss, $\mathcal{L}_{\text{bin}}$ and $\mathcal{L}_{\text{edge}}$. Finally, the overall training objective is defined as:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{CE}} + \lambda_2 \mathcal{L}_{\text{Lov}} + \lambda_3 \mathcal{L}_{\text{bin}} + \lambda_4 \mathcal{L}_{\text{edge}}, \tag{14}$$

where λ_i are balancing coefficients and are set to 1 in our experiments.

By jointly exploiting high-level priors, intermediate RGB-T semantics, and low-level details, the decoder improves category discrimination, region completeness, and boundary localization.

4 Experiments

4.1 Datasets and Implementation Details

Experiments are conducted on three public benchmarks: MFNet, PST900, and FMB. MFNet contains 1,569 aligned RGB-T pairs at 480×640 with nine semantic categories. Following previous works (Li et al., 2023; Zhou et al., 2025; Liu et al., 2025), 784/392/393 pairs are used for training/validation/testing. PST900 contains 894 pairs at 1280×720 with five categories, split into

Table 1: Quantitative comparison on the MFNet dataset.

Methods	Car		Person		Bike		Curve		Car-Stop		Guardrail		Color-Cone		Bump		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$MFNet_{17}$	77.2	65.9	67.0	58.9	53.9	42.9	36.2	29.9	19.1	9.9	0.1	8.5	30.3	25.2	30.0	27.7	45.1	39.7
$RTFNet_{19}$	93.0	87.4	79.3	70.3	76.8	62.7	60.7	45.3	38.5	29.8	0.0	0.0	45.5	29.1	74.7	55.7	63.1	53.2
$MMNet_{22}$	–	83.9	–	69.3	–	59.0	–	43.2	–	24.7	–	4.6	–	42.2	–	50.7	62.7	52.8
$CCFFNet_{22}$	94.5	89.6	83.6	74.2	73.2	63.1	67.2	50.5	38.7	31.9	30.6	4.8	55.2	49.7	72.9	56.3	68.3	57.6
$GCNet_{22}$	94.2	86.0	89.6	72.0	77.5	60.0	68.9	42.8	38.3	30.7	45.8	6.2	59.6	49.5	82.1	52.6	72.7	55.3
$EGFNet_{23}$	95.8	87.6	89.0	69.8	80.6	58.8	71.5	42.8	48.7	33.8	33.6	7.0	65.3	48.3	71.1	47.1	72.7	54.8
$DSGBINet_{23}$	95.2	87.4	89.2	69.5	**85.2**	64.7	66.0	46.3	56.7	43.4	7.8	3.3	**82.0**	**61.7**	72.8	48.9	72.6	58.1
$LASNet_{23}$	94.9	84.2	81.7	67.1	82.1	56.9	70.7	41.1	56.8	39.6	59.5	18.9	58.1	48.8	77.2	40.1	**75.4**	54.9
$ECGFNet_{23}$	89.4	83.5	85.2	72.1	72.9	61.6	62.8	40.5	44.8	30.8	45.2	11.1	57.2	49.7	65.1	50.9	69.1	55.3
CMX_{23}	–	90.1	–	75.2	–	64.5	–	50.2	–	35.3	–	8.5	–	54.2	–	60.6	–	59.7
$FDCNet_{23}$	94.1	87.5	91.4	72.4	78.1	61.7	70.1	43.8	34.4	27.2	61.5	7.3	64.0	52.0	74.5	56.6	74.1	56.3
$MDRNet_{23}$	95.2	87.1	**92.5**	69.8	76.2	60.9	72.0	47.8	42.3	34.2	66.8	8.2	64.8	50.2	63.5	55.0	74.7	56.8
$SGFNet_{23}$	94.3	88.4	90.3	**77.6**	77.9	64.3	67.7	45.8	37.7	31.0	57.3	6.0	64.2	57.1	73.4	55.0	73.6	57.6
$RSFNet_{24}$	93.4	83.5	91.8	72.7	83.7	61.9	75.1	40.7	39.7	31.7	59.1	11.2	59.1	50.9	60.5	45.5	73.2	55.1
$CAINet_{24}$	93.0	88.5	74.6	66.3	**85.2**	**68.7**	65.9	55.4	34.7	31.5	65.6	9.0	55.6	48.9	**85.0**	60.7	73.2	58.6
$KDSNet_{24}$	93.7	88.4	74.2	45.7	79.5	61.6	**80.2**	**69.7**	48.0	38.4	**79.6**	**49.9**	64.0	52.2	59.0	8.7	75.2	56.9
C^4Net_{25}	93.0	87.3	89.0	74.1	75.3	61.8	66.3	44.9	51.1	43.4	49.2	11.3	59.7	50.9	70.9	**63.8**	72.6	59.5
$MDBFNet_{25}$	94.5	89.1	89.5	73.5	78.4	62.6	71.3	48.8	42.5	34.6	65.0	8.0	62.4	54.5	75.4	51.3	75.3	57.8
$AGFNet_{25}$	95.4	88.0	87.6	73.7	79.4	64.6	67.1	45.3	40.9	34.3	64.9	13.8	64.4	57.0	72.1	52.5	74.5	58.6
$MiLNet_{25}$	93.2	87.9	85.9	72.3	76.7	64.5	70.8	46.2	57.7	**45.8**	39.4	10.1	63.8	55.8	72.1	60.1	73.2	60.1
$USformer_{26}$	93.1	88.6	84.4	73.1	81.1	64.4	65.6	48.6	**57.8**	44.0	26.6	21.1	65.1	54.3	79.3	50.4	72.5	60.3
SAR_{26}	94.6	88.6	90.1	74.0	76.5	63.8	61.0	43.8	48.7	41.7	46.8	10.0	67.1	56.7	70.3	58.4	72.5	59.4
$TUNI_{26}$	94.8	**90.6**	86.3	75.7	79.2	66.4	62.1	49.4	45.8	40.5	7.8	2.3	64.7	58.6	78.7	61.5	68.7	60.4
Ours	**95.9**	89.9	89.6	75.3	82.9	65.0	64.4	47.3	53.4	45.3	41.4	15.7	65.1	56.6	67.4	57.2	73.2	**61.2**

Table 2: Performance comparison under day and night conditions on MFNet.

Methods		EGFNet	LASNet	ECGFNet	FDCNet	SGFNet	RSFNet	CAINet	KDSNet	MDBFNet	AGFNet	MiLNet	TUNI	**Ours**
Day	mAcc	74.4	73.3	67.5	62.0	70.7	72.9	70.4	**75.6**	74.6	73.0	71.5	66.7	73.4
	mIoU	47.3	45.2	47.1	47.8	49.2	49.1	48.5	50.5	50.1	49.9	51.8	**54.9**	53.7
Night	mAcc	68.0	72.8	67.3	72.2	72.2	**73.8**	69.8	71.5	72.3	71.4	71.3	65.9	72.0
	mIoU	55.0	58.7	55.9	56.8	58.4	57.2	59.1	56.3	58.4	58.7	60.0	59.1	**61.3**

597/288 training/testing pairs. FMB contains 1,500 pairs at 800×600 with 14 categories, split into 1,220/280 training/testing pairs.

Experiments are implemented in PyTorch on an Intel Core i7-13700KF CPU, 16 GB RAM, and an NVIDIA RTX 4080 GPU. RGB and TIR inputs retain their original resolutions. A frozen SAM ViT-L encoder provides semantic priors. Training uses the Ranger optimizer with an initial learning rate of 1×10^{-4}, weight decay of 5×10^{-4}, and batch size of 2 for 200 epochs, with polynomial learning-rate decay (power=0.9) and mixed precision.

4.2 Comparison with State of the Art

We compare BiSANet with 23 representative RGB-T semantic segmentation methods, including MFNet (Ha et al., 2017), RTFNet (ResNet152) (Sun et al., 2019), MMNet (Lan et al., 2022), CCFFNet (Wu et al., 2022b), GCNet (Liu et al., 2022), EGFNet (Dong et al., 2023), DSGBINet (Xu et al., 2023), LASNet (Li et al., 2023), ECGFNet (Zhou et al., 2023), CMX (Zhang et al., 2023b), FDCNet (Zhao & Zhang, 2023), MDRNet (Zhao et al., 2023), SGFNet (Wang et al., 2023), RSFNet (Li et al., 2024), CAINet (Lv et al., 2024), KDSNet (Zhou et al., 2024), C^4Net (Zhao et al., 2025), MDBFNet (Liang et al., 2025), AGFNet (Zhou et al., 2025), MiLNet (Liu et al., 2025), USformer (Yang et al., 2026), SAR Zhao et al. (2026), and TUNI (Guo et al., 2026).

4.2.1 Comparison on the MFNet Dataset

Table 1 reports the quantitative results on MFNet. BiSANet achieves the best mIoU of 61.2% and a competitive mAcc of 73.2%. It also maintains competitive performance across categories with different scales and structures, demonstrating strong overall segmentation quality and balanced recognition capability. As shown in Table 2, BiSANet achieves a competitive daytime mIoU of 53.7% and the best nighttime mIoU of 61.3%. The stronger nighttime performance suggests that complementary thermal cues and adapted semantic priors can effectively compensate for degraded RGB appearance, improving robustness under challenging illumination conditions.

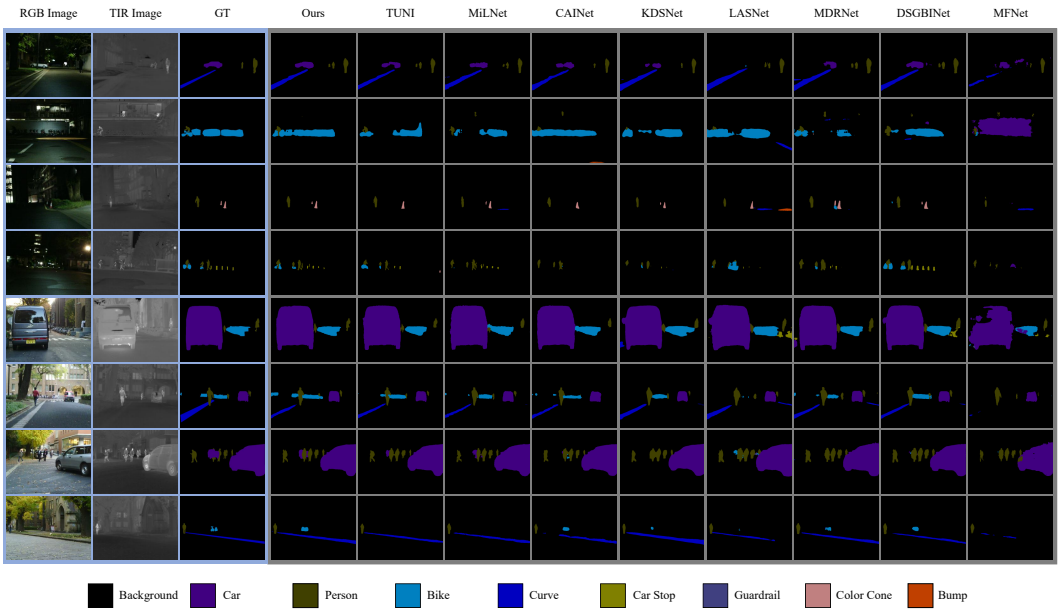


Figure 3: Qualitative comparison on the MFNet dataset.

Table 3: Quantitative comparison on the PST900 dataset.

Methods	Background		Hand-Drill		Backpack		Extinguisher		Survivor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$EGFNet_{23}$	99.5	99.3	**98.0**	64.7	94.2	83.1	**95.2**	71.3	83.3	74.3	94.0	78.5
$DSGBINet_{23}$	99.7	99.4	94.5	75.0	88.7	85.1	94.8	79.3	81.4	75.6	91.8	82.9
$LASNet_{23}$	99.8	99.5	91.8	82.8	90.8	86.5	92.4	77.8	83.4	75.5	91.6	84.4
$MDRNet_{23}$	99.3	99.1	86.6	63.5	93.5	76.3	90.2	64.0	85.6	71.3	91.0	74.6
$FDCNet_{23}$	99.7	99.2	82.5	70.4	77.5	72.2	91.8	71.5	78.4	72.4	86.0	77.1
$SGFNet_{23}$	99.8	99.4	94.0	76.7	90.4	85.4	89.4	75.6	82.7	76.7	91.2	82.8
$RSFNet_{24}$	–	99.4	–	72.9	–	84.9	–	75.4	–	70.1	–	80.5
$CAINet_{24}$	99.7	99.5	95.9	80.3	**96.1**	88.0	88.4	77.2	**91.4**	78.7	94.3	84.7
C^4Net_{25}	99.7	99.5	83.5	69.4	94.8	88.4	94.7	79.6	83.4	76.0	91.2	82.6
$AGFNet_{25}$	99.8	99.5	94.0	80.3	90.7	85.3	91.0	80.2	83.5	78.7	91.8	84.8
$MiLNet_{25}$	99.8	99.5	95.4	76.3	93.5	89.6	94.9	**82.2**	81.3	78.0	93.0	85.1
$USformer_{26}$	–	**99.6**	–	82.2	–	89.3	–	77.8	–	79.5	92.6	85.7
SAR_{26}	–	–	–	–	–	–	–	–	–	–	92.3	78.2
$TUNI_{26}$	**99.9**	**99.6**	94.4	87.2	91.3	88.5	91.8	80.2	85.7	82.4	92.6	87.6
Ours	99.8	**99.6**	96.5	**87.5**	94.5	**90.8**	93.3	80.8	88.6	**82.5**	**94.5**	**88.2**

Fig. 3 presents qualitative comparisons on MFNet. BiSANet produces more complete regions for partially occluded cars in rows 1 and 7, and more accurate boundaries for distant small objects and thin-structured bicycles. It also preserves clearer object structures in cluttered scenes. In the dark scenes of the first four rows, BiSANet maintains more complete target regions under degraded RGB appearance, further demonstrating its robustness to low-light conditions.

4.2.2 Comparison on the PST900 Dataset and the FMB Dataset

As shown in Table 3, BiSANet achieves the highest mAcc of 94.5% and mIoU of 88.2%, improving the previous best mIoU by 0.6 percentage points. It obtains the best or tied-best IoU in four of the five categories and ranks second on Extinguisher. These results demonstrate effective semantic-prior utilization across categories. Qualitative comparisons are provided in the supplementary material.

As shown in Table 4, BiSANet achieves the best mAcc of 79.8% and mIoU of 67.9%, improving the previous best mIoU by 1.6 percentage points. It ranks first or second in IoU across all eight reported categories, further demonstrating consistent performance across RGB-T benchmarks. Notably, the larger gains on FMB may stem from its dense pixel-wise setting, where properly adapted SAM priors provide richer global semantics and structural cues to complement local RGB-T features.

Table 4: Quantitative comparison on the FMB dataset, showing 8 of the 14 classes for brevity.

Methods	Car		Person		Truck		T-Lamp		T-Sign		Building		Vegetation		Pole		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$EGFNet_{23}$	83.6	77.4	79.5	63.0	33.5	17.1	58.6	25.2	82.6	66.6	88.5	77.2	89.3	83.5	63.8	41.5	63.0	47.3
$LASNet_{23}$	81.3	72.6	76.4	48.6	29.6	14.8	20.7	2.9	79.3	59.0	86.9	75.4	87.6	81.6	56.6	36.7	56.9	42.5
$SGFNet_{23}$	91.2	73.9	66.9	54.9	30.8	18.7	–	–	79.9	66.8	88.3	79.7	**92.4**	84.5	53.4	42.7	58.1	50.2
$CAINet_{24}$	88.1	73.8	69.4	54.6	33.0	18.3	–	–	80.9	68.5	87.9	79.6	**92.4**	84.3	54.1	42.6	59.1	50.5
$AGFNet_{25}$	**95.3**	78.9	80.0	62.1	46.7	36.9	56.8	46.3	85.2	72.3	91.6	81.0	91.4	85.4	65.5	**63.0**	70.6	60.5
SAR_{26}	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	66.0	55.1
$TUNI\text{-}B_{26}$	–	**82.7**	–	73.1	–	–	–	–	–	–	–	–	–	–	–	50.7	–	66.3
Ours	94.7	82.6	**87.2**	**73.5**	**62.0**	**50.4**	**72.5**	**56.7**	**90.9**	**77.7**	**93.8**	**84.9**	**92.4**	**87.3**	**76.3**	53.0	**79.8**	**67.9**

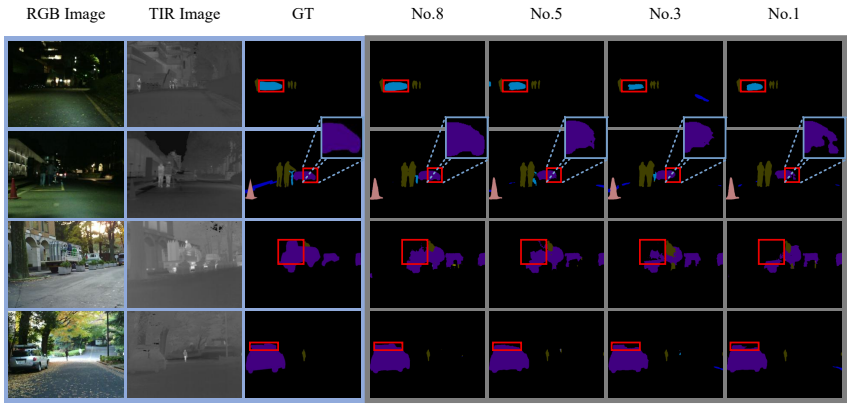


Figure 4: Qualitative comparison of the ablation variants.

4.3 Ablation Studies

To investigate the contributions of each component, Table 5 evaluates the Hierarchical Dual-Modal Encoder (HDE), Bidirectional Semantic Adapters (BSA), and Semantic Enhancement Decoder (SED) on the MFNet dataset. The baseline adopts MiT-B3 encoders with MFNet-style fusion (Ha et al., 2017) to generate four-level RGB-T features, which are combined with SAM semantic features and fed into a SegFormer-style decoder (Xie et al., 2021), representing a lightweight unidirectional prior injection scheme. Adding HDE, BSA, or SED individually improves mIoU, verifying their effectiveness. BSA alone brings limited gains due to the lack of stable RGB-T representations for semantic alignment. All two-component variants outperform single-component ones, while the full model achieves the best performance, demonstrating the complementary roles of multimodal fusion, semantic alignment, and refined decoding.

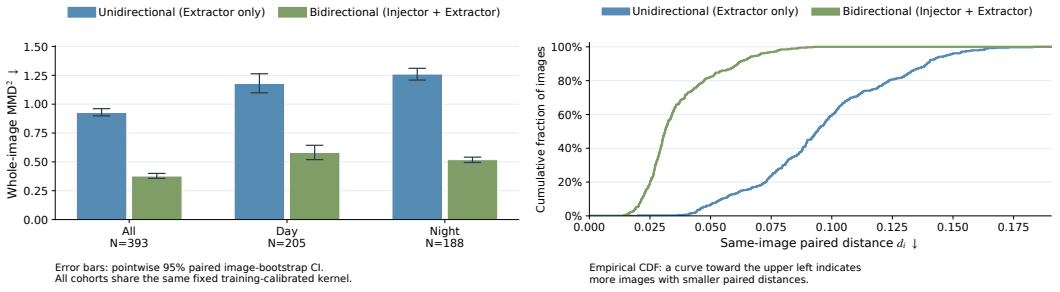
Table 5: Ablation study of the main components in BiSANet.

No.	Baseline	HDE	BSA	SED	mAcc	mIoU
1	✓				71.1	58.7
2	✓	✓			72.6	59.6
3	✓		✓		70.6	59.2
4	✓			✓	71.5	59.5
5	✓	✓	✓		73.2	60.2
6	✓	✓		✓	72.0	59.9
7	✓		✓	✓	72.6	60.3
8	✓	✓	✓	✓	**73.2**	**61.2**

Fig. 4 compares No. 1, No. 3, No. 5, and No. 8. The baseline produces incomplete regions and coarse boundaries. Adding BSA improves semantic consistency, while further introducing HDE provides stable RGB-T representations and brings substantial improvements. Finally, SED further refines object regions and boundaries, achieving the best qualitative results. Ablation studies on different loss combinations are provided in the supplementary material.

4.4 Effectiveness of the Adapter Design

To assess whether bidirectional interaction narrows the semantic gap between RGB-pretrained semantics and RGB-T representations, we compare the Extractor-only and full adapters. MMD^2 measures distribution discrepancy (Gretton et al., 2012), while the paired-distance CDF (Wasserman, 2004) reflects image-level correspondence. As shown in Fig. 5, bidirectional adaptation reduces MMD^2 from 0.929 to 0.377 and the mean paired distance from 0.0962 to 0.0365.



(a) (b)

Figure 5: Semantic alignment between S and $\tilde{F}_3^{RT}$. (a) MMD^2. (b) Empirical CDF of same-image paired distances. Lower MMD^2 and a leftward CDF shift indicate closer alignment.

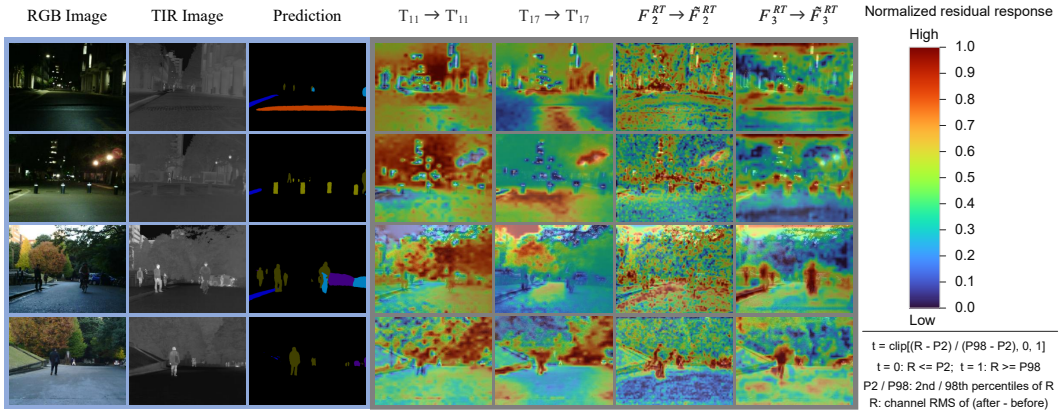


Figure 6: Residual response maps of bidirectional feature adaptation, obtained by aggregating channel-wise differences for $T_{11} \to T'_{11}$, $T_{17} \to T'_{17}$, $F_2^{RT} \to \tilde{F}_2^{RT}$, and $F_3^{RT} \to \tilde{F}_3^{RT}$.

With this gap reduced, the two streams can better exploit complementary information from each other. As shown in Fig. 6, injecting RGB-T features produces clear residual changes in the SAM stream, especially around dark or appearance-degraded targets, while extracting the adapted SAM priors increasingly concentrates changes on semantically coherent target regions in the RGB-T stream. This suggests that RGB-T cues adapt SAM semantics toward the target domain, which in turn better refine multimodal representations.

Table 6: Comparison of Adapter variants and insertion positions. "w/o" denotes "without."

Bidirectional Adapter			Insertion Positions		
Variant	mAcc	mIoU	Position	mAcc	mIoU
w/o Adapter	72.0	59.9	[5, 11, 17]	72.3	58.8
Injector only	72.9	60.4	[5, 11, 23]	**73.6**	59.6
Extractor only	70.5	59.8	[5, 17, 23]	72.8	59.4
Full Adapter	**73.2**	**61.2**	[11, 17, 23]	73.2	**61.2**

Finally, we quantitatively evaluate the effectiveness of the bidirectional adapters and their insertion positions. As shown in Table 6, the full Adapter outperforms using only the Injector or Extractor. Among the three-block combinations selected from global-attention layers $5, 11, 17, 23$, $11, 17, 23$ achieves the best mIoU, validating the proposed adapter design.

5 Conclusion

In this paper, we propose BiSANet for RGB-T semantic segmentation, enabling bidirectional alignment between RGB-T features and RGB-pretrained semantic priors. The Hierarchical Dual-Modal Encoder, Bidirectional Semantic Adapters, and Semantic Enhancement Decoder jointly improve multimodal fusion, semantic alignment, and fine-grained prediction. Experiments on MFNet, PST900, and FMB validate its effectiveness in category discrimination, region completeness, and boundary localization. Failure cases involving low-contrast and confusing targets are provided in the supplementary material. Future work will explore lightweight global semantic modeling.

AI USE STATEMENT

In this work, we used generative AI tools to assist with software implementation, code completion, debugging, and refactoring. We did not use generative AI tools to generate or modify datasets, formulate mathematical claims, propose hypotheses, design the research methodology or experiments, or interpret experimental results. Tasks related to theoretical proofs and qualitative data analysis are not applicable to this work.

Additionally, we used generative AI tools to improve the readability, grammar, and style of the manuscript and to assist with code documentation and formatting. The research ideas, model design, experimental protocol, analysis, and conclusions were developed and determined by the authors. All AI-assisted text was manually reviewed and revised by the authors. All AI-assisted code was manually inspected and tested against its intended functionality, and the reported experimental results were independently checked by the authors.

We take full responsibility for the final content of this work, including all text, claims, results, and software artifacts produced with the aid of generative AI.

REFERENCES

Maxim Berman, Amal Rannen Triki, and Matthew B Blaschko. The lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4413–4421, 2018.

Yuri Boykov, Olga Veksler, and Ramin Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on pattern analysis and machine intelligence*, 23(11):1222–1239, 2002.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 9650–9660, 2021.

Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4): 834–848, 2017.

Liang-Chieh Chen, Yukun Zhu, George Papandreou, Florian Schroff, and Hartwig Adam. Encoder-decoder with atrous separable convolution for semantic image segmentation. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 801–818, 2018.

Xinlei Chen, Saining Xie, and Kaiming He. An empirical study of training self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 9640–9649, 2021.

Bowen Cheng, Ishan Misra, Alexander G. Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1290–1299, 2022.

Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4113–4123, 2024.

Shaohua Dong, Wujie Zhou, Caie Xu, and Weiqing Yan. Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing. *IEEE Transactions on Intelligent Transportation Systems*, 25(1):657–669, 2023.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Xiaodong Guo, Wujie Zhou, and Tong Liu. Contrastive learning-based knowledge distillation for rgb-thermal urban scene semantic segmentation. *Knowledge-Based Systems*, 292:111588, 2024.

Xiaodong Guo, Xianda Guo, Tong Liu, Zhihong Deng, Yanlun Peng, Xiang Li, and Wujie Zhou. Tuni: Unifying pre-training and fine-tuning with modality-aware mutual learning and rectification for rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2026.

Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 5108–5115. IEEE, 2017.

Mark Hamilton, Zhoutong Zhang, Bharath Hariharan, Noah Snavely, and William T. Freeman. Unsupervised semantic segmentation by distilling feature correspondences. In *International Conference on Learning Representations*, 2022.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 4015–4026, 2023.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *International Conference on Machine Learning*, pp. 3519–3529, 2019.

Xin Lan, Xiaojing Gu, and Xingsheng Gu. Mmnet: Multi-modal multi-stage network for rgb-t image semantic segmentation. *Applied Intelligence*, 52(5):5817–5829, 2022.

Gongyang Li, Yike Wang, Zhi Liu, Xinpeng Zhang, and Dan Zeng. Rgb-t semantic segmentation with location, activation, and sharpening. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(3):1223–1235, 2023.

Ping Li, Junjie Chen, Binbin Lin, and Xianghua Xu. Residual spatial fusion network for rgb-thermal semantic segmentation. *Neurocomputing*, 595:127913, 2024.

Wenli Liang, Caifeng Shan, Yuanjian Yang, and Jungong Han. Multi-branch differential bidirectional fusion network for rgb-t semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, 2025.

Jinfu Liu, Wujie Zhou, Yueli Cui, Lu Yu, and Ting Luo. Gcnet: Grid-like context-aware network for rgb-thermal semantic segmentation. *Neurocomputing*, 506:60–67, 2022.

Jinfu Liu, Hong Liu, Xia Li, Jiale Ren, and Xinhua Xu. Milnet: Multiplex interactive learning network for rgb-t semantic segmentation. *IEEE Transactions on Image Processing*, 2025.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3431–3440, 2015.

Ying Lv, Zhi Liu, and Gongyang Li. Context-aware interaction network for rgb-t semantic segmentation. *IEEE Transactions on Multimedia*, 26:6348–6360, 2024.

Hyeonwoo Noh, Seunghoon Hong, and Bohyung Han. Learning deconvolution network for semantic segmentation. In *Proceedings of the IEEE international conference on computer vision*, pp. 1520–1528, 2015.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pp. 234–241. Springer, 2015.

J-L Starck, Michael Elad, and David L Donoho. Image decomposition via the combination of sparse representations and a variational approach. *IEEE transactions on image processing*, 14(10):1570–1582, 2005.

Yuxiang Sun, Weixun Zuo, and Ming Liu. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. *IEEE Robotics and Automation Letters*, 4(3):2576–2583, 2019.

Yuxiang Sun, Weixun Zuo, Peng Yun, Hengli Wang, and Ming Liu. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. *IEEE Transactions on Automation Science and Engineering*, 18(3):1000–1011, 2021.

Xinlong Wang, Rufeng Zhang, Chunhua Shen, Tao Kong, and Lei Li. Dense contrastive learning for self-supervised visual pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3024–3033, 2021.

Yike Wang, Gongyang Li, and Zhi Liu. Sgfnet: Semantic-guided fusion network for rgb-thermal semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33 (12):7737–7748, 2023.

Larry Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer, 2004. doi: 10.1007/978-0-387-21736-9.

Kan Wu, Jinnian Zhang, Houwen Peng, Mengchen Liu, Bin Xiao, Jianlong Fu, and Lu Yuan. TinyViT: Fast pretraining distillation for small vision transformers. In *European Conference on Computer Vision*, 2022a.

Wei Wu, Tao Chu, and Qiong Liu. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. *Pattern Recognition*, 131:108881, 2022b.

Bin Xie, Jiale Cao, Jin Xie, Fahad Shahbaz Khan, and Yanwei Pang. Sed: A simple encoder-decoder for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3426–3436, 2024a.

Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. *Advances in neural information processing systems*, 34:12077–12090, 2021.

Zhaozhi Xie, Bochen Guan, Weihao Jiang, Muyang Yi, Yue Ding, Hongtao Lu, and Lei Zhang. Pa-sam: Prompt adapter sam for high-quality image segmentation. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pp. 1–6. IEEE, 2024b.

Chang Xu, Qingwu Li, Xiongbiao Jiang, Dabing Yu, and Yaqin Zhou. Dual-space graph-based interaction network for rgb-thermal semantic segmentation in electric power scene. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(4):1577–1592, 2023.

Fan Yang, Feng Shao, Baoyang Mu, Xiongli Chai, and Qiuping Jiang. Usformer: A u-shaped structure transformer for RGB-thermal semantic segmentation and traffic scene understanding. *IEEE Transactions on Intelligent Transportation Systems*, 27(1):623–638, 2026.

Yuhui Yuan, Xilin Chen, and Jingdong Wang. Object-contextual representations for semantic segmentation. In *European conference on computer vision*, pp. 173–190. Springer, 2020.

Chaoning Zhang, Dongshen Han, Yu Qiao, Jung Uk Kim, Sung-Ho Bae, Seungkyu Lee, and Choong Seon Hong. Faster segment anything: Towards lightweight sam for mobile applications. *arXiv preprint arXiv:2306.14289*, 2023a.

Jiaming Zhang, Huayao Liu, Kailun Yang, Xinxin Hu, Ruiping Liu, and Rainer Stiefelhagen. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. *IEEE Transactions on intelligent transportation systems*, 24(12):14679–14694, 2023b.

Qi Zhang, Jinwen Xing, Han Zhang, Ruixiang Yan, Mengyao Hao, Longquan Yan, and Zhaodi Li. Ig-sam: Implicitly guided sam for rgb-thermal semantic segmentation. *Expert Systems with Applications*, pp. 132804, 2026.

Qiang Zhang, Shenlu Zhao, Yongjiang Luo, Dingwen Zhang, Nianchang Huang, and Jungong Han. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2633–2642, 2021.

Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2881–2890, 2017.

Shenlu Zhao and Qiang Zhang. A feature divide-and-conquer network for rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(6):2892–2905, 2023.

Shenlu Zhao, Yichen Liu, Qiang Jiao, Qiang Zhang, and Jungong Han. Mitigating modality discrepancies for rgb-t semantic segmentation. *IEEE Transactions on Neural Networks and Learning Systems*, 35(7):9380–9394, 2023.

Shenlu Zhao, Jingyi Li, and Qiang Zhang. C^4net: Excavating cross-modal context-and content-complementarity for rgb-t semantic segmentation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(2):1347–1361, 2025.

Shenlu Zhao, Jingyi Li, Qiang Jiao, Qiang Zhang, and Jungong Han. Sar: semantic ambiguity relief network for rgb-t semantic segmentation. *Information Fusion*, pp. 104671, 2026.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip H. S. Torr, and Li Zhang. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6881–6890, 2021a.

Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 6881–6890, 2021b.

Wujie Zhou, Ying Lv, Jingsheng Lei, and Lu Yu. Embedded control gate fusion and attention residual learning for rgb–thermal urban scene parsing. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):4794–4803, 2023.

Wujie Zhou, Tingting Gong, and Weiqing Yan. Knowledge distillation segformer-based network for rgb-t semantic segmentation. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 55 (3):2170–2182, 2024.

Xiaofei Zhou, Xiaoling Wu, Liuxin Bao, Haibing Yin, Qiuping Jiang, and Jiyong Zhang. Agfnet: Adaptive gated fusion network for rgb-t semantic segmentation. *IEEE Transactions on Intelligent Transportation Systems*, 26(5):6477–6492, 2025.

A Appendix

A.1 Qualitative comparison on the PST900 dataset

Qualitative comparisons in Fig. 7 further support the quantitative results. In challenging underground scenes with low illumination, noise, and background interference, multiple competing methods produce incomplete regions, coarse predictions, or miss small targets. In contrast, BiSANet generates more complete and accurate predictions for all targets. When either modality is severely degraded, the hierarchical dual-modal encoder and bidirectional semantic adapters jointly exploit complementary cues from the other modality, while the Semantic Enhancement Decoder further refines object regions and boundaries. Consequently, BiSANet better preserves object integrity and boundary details, demonstrating strong robustness on PST900.

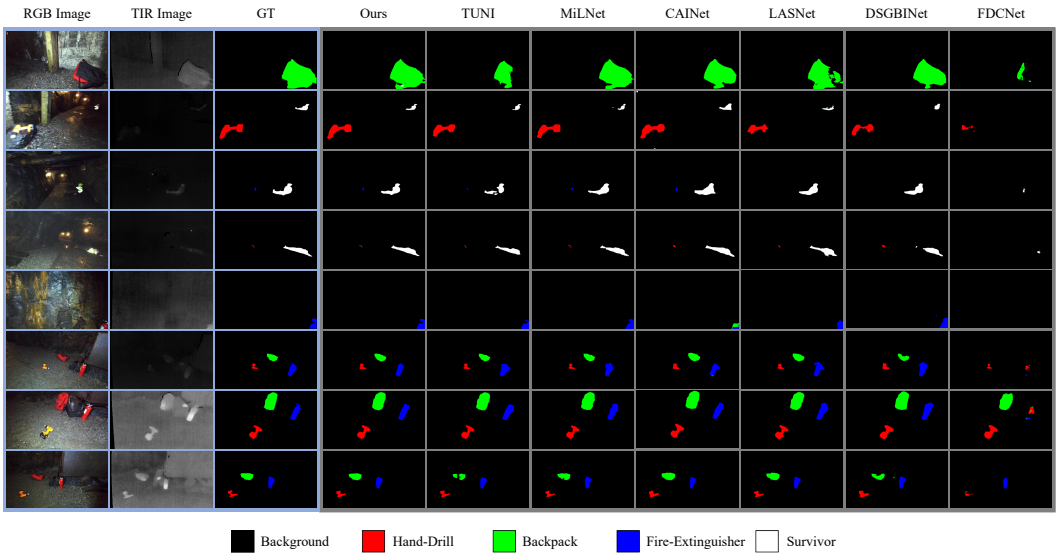


Figure 7: Qualitative comparison on the PST900 dataset.

Table 7: Ablation study of different loss combinations on MFNet.

No.	$\mathcal{L}_{\text{CE}}$	$\mathcal{L}_{\text{Lov}}$	$\mathcal{L}_{\text{bin}}$	$\mathcal{L}_{\text{edge}}$	mAcc	mIoU
1	✓				72.6	58.6
2	✓	✓			71.8	59.8
3	✓	✓	✓		71.7	59.9
4	✓	✓		✓	72.0	60.7
5	✓	✓	✓	✓	73.2	61.2

A.2 Ablation study of different loss combinations

Table 7 evaluates the contribution of the four training objectives. Using only the weighted cross-entropy loss $\mathcal{L}_{\text{CE}}$ yields 72.6% mAcc and 58.6% mIoU. Adding the Lovász-Softmax loss $\mathcal{L}_{\text{Lov}}$ improves mIoU by 1.2 percentage points to 59.8%, indicating better region-level overlap despite a slight change in mAcc. Further introducing foreground–background supervision $\mathcal{L}_{\text{bin}}$ increases mIoU to 59.9%, while boundary supervision $\mathcal{L}_{\text{edge}}$ provides a larger gain to 60.7% by emphasizing semantic transitions, which is beneficial for small objects and thin structures.

Combining all four losses achieves the best performance of 73.2% mAcc and 61.2% mIoU, improving over cross-entropy alone by 0.6 and 2.6 percentage points, respectively. These results demonstrate their complementary roles: $\mathcal{L}_{\text{CE}}$ supports category discrimination, $\mathcal{L}_{\text{Lov}}$ improves region overlap, $\mathcal{L}_{\text{bin}}$ encourages foreground completeness, and $\mathcal{L}_{\text{edge}}$ strengthens boundary localization. The two auxiliary heads are used only during training and therefore introduce no additional inference-time computation.

A.3 Effectiveness of the HDE design

Table 8 evaluates the two key designs of the Hierarchical Dual-Modal Encoder. Removing spatial gating reduces mAcc from 73.2% to 71.9% and mIoU from 61.2% to 60.4%, showing its importance in selecting useful low-level cross-modal differences. Removing adaptive weighting reduces mIoU to 59.8%, indicating that selectively combining modality-specific, discrepancy, and mixed representations is important for reliable high-level fusion. The full HDE achieves the best performance, confirming the effectiveness of both level-specific designs.

Table 8: Ablation study of the HDE design on MFNet.

Variant	mAcc	mIoU
w/o spatial gating	71.9	60.4
w/o adaptive weighting	73.2	59.8
Full HDE	73.2	61.2

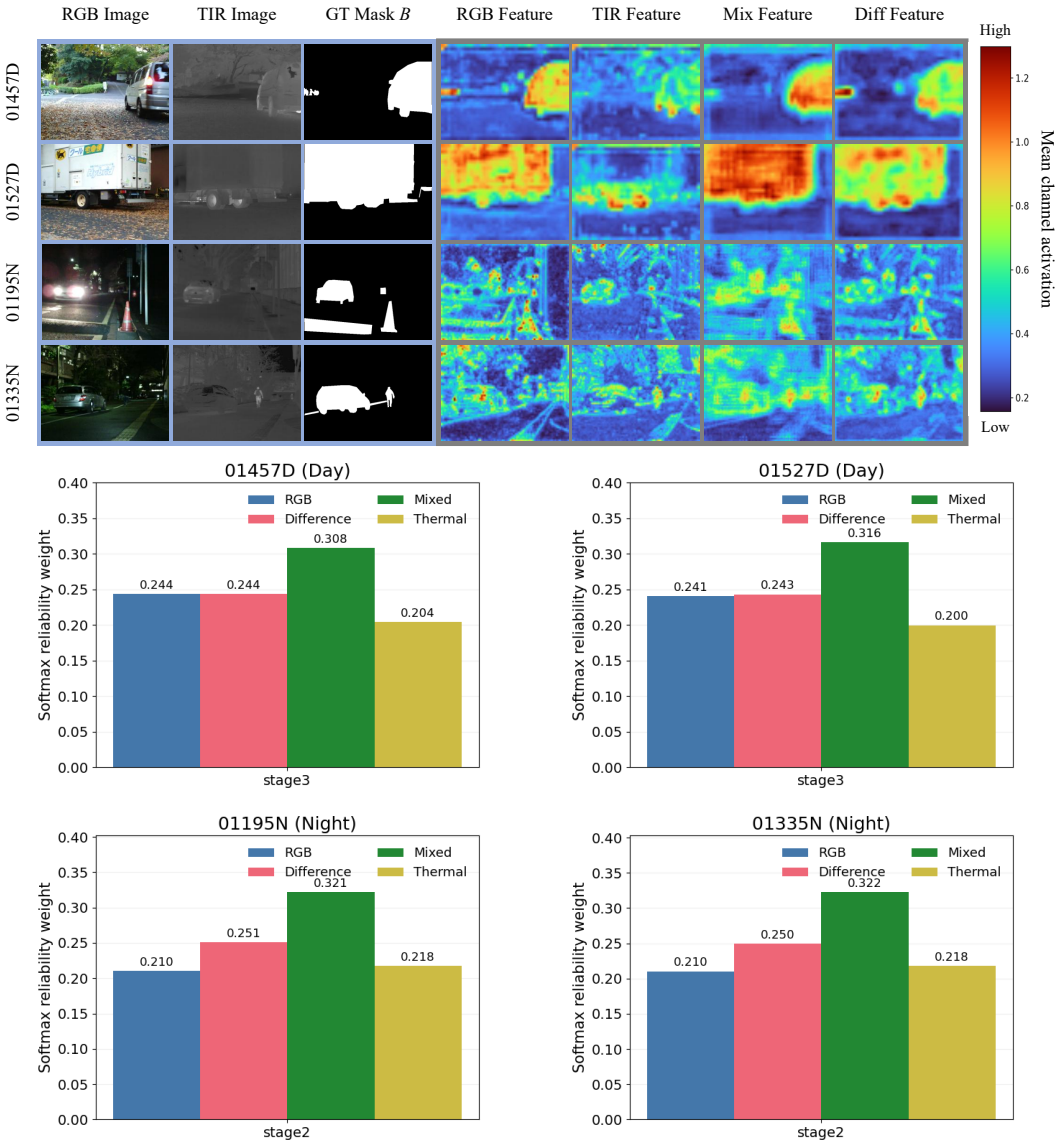


Figure 8: Visualization of adaptive feature weighting in HDE. The upper part shows mean channel activations of the four candidate features, and the lower part shows their average Softmax reliability weights for daytime and nighttime examples.

To further examine the adaptive weighting mechanism, Fig. 8 visualizes the four candidate representations and their reliability weights under different illumination conditions. The mixed feature consistently receives the highest weight because it integrates complementary information from both modalities. In daytime scenes, RGB and difference features generally receive higher weights than

Table 9: Comparison of unidirectional and bidirectional adaptation under day and night conditions on MFNet.

Variant	Day		Night	
	mAcc	mIoU	mAcc	mIoU
Extractor only	73.5	54.1	67.9	58.3
Injector only	75.2	53.2	71.1	59.5
Full Adapter	73.4	53.7	72.0	61.3
Gain vs. Extractor	−0.1	−0.4	+4.1	+3.0
Gain vs. Injector	−1.8	+0.5	+0.9	+1.8

Table 10: Ablation study of the SED design on MFNet.

Variant	mAcc	mIoU
w/o ASPP	72.8	60.6
w/o low-level concatenation	72.3	60.4
Full SED	73.2	61.2

the thermal feature, consistent with their more complete target responses. For example, the bicycle in the first example and part of the car in the second are weakly represented in the thermal stream.

The weighting behavior changes accordingly at night. The thermal feature receives a slightly higher weight than the RGB feature, while the difference feature is also assigned a relatively large weight because it captures complementary fine-grained cues. This is particularly evident in the third example, where the thermal feature weakly represents the color cone while the RGB feature provides insufficient car information, and in the fourth example, where the car response is weak in the thermal feature while the person is poorly represented in RGB. These observations further demonstrate that HDE adaptively adjusts the contributions of different feature representations according to their scene-dependent reliability.

A.4 Effectiveness of the BSAs under Day and Night Conditions

Table 9 further compares unidirectional and bidirectional semantic adaptation under different illumination conditions. During daytime, the three variants exhibit comparable mIoU, with the Extractor-only variant slightly higher than the full Adapter. This suggests that when RGB appearance is sufficiently informative, direct transfer of RGB-pretrained semantic priors already provides effective guidance, leaving less room for additional target-domain adaptation.

In contrast, bidirectional adaptation shows a clear advantage at night. Compared with the Extractor-only variant, the full Adapter improves nighttime mAcc and mIoU by 4.1 and 3.0 percentage points, respectively. It also outperforms the Injector-only variant by 0.9 points in mAcc and 1.8 points in mIoU. Under degraded illumination, RGB-pretrained semantics are more likely to be mismatched with RGB-T representations, while thermal cues provide complementary target-domain information. Injecting these cues into the semantic pathway before transferring the adapted priors back to the RGB-T branch therefore yields more reliable semantic correction. These results further support the benefit of bidirectional adaptation, particularly under challenging nighttime conditions.

Fig. 9 provides qualitative support for the above results. In daytime scenes, the three variants produce similar predictions. At night, the difference becomes more pronounced: targets that are weak or nearly invisible in RGB, such as distant bicycles and parts of cars, are better preserved by the full Adapter, consistent with its larger nighttime gains.

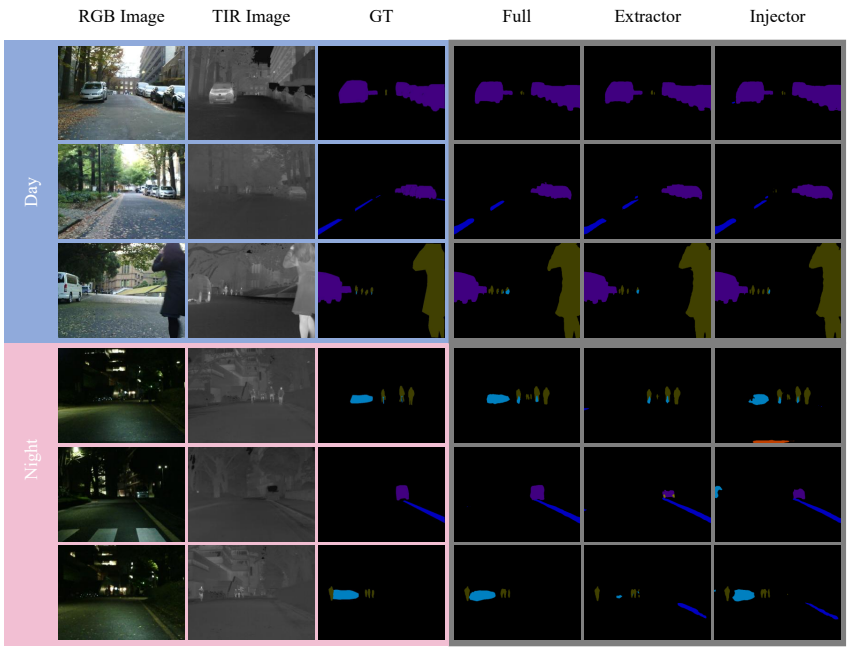


Figure 9: Qualitative comparison of Adapter variants under daytime and nighttime conditions on MFNet. Full denotes the bidirectional Adapter, while Extractor and Injector denote the corresponding unidirectional variants.

A.5 Effectiveness of the SED design

Table 10 examines the two key designs of the Semantic Enhancement Decoder. Removing ASPP reduces mAcc from 73.2% to 72.8% and mIoU from 61.2% to 60.6%. In the full decoder, ASPP applies parallel dilated convolutions to the adapted high-level semantic feature, capturing multi-scale context and improving recognition when local appearance is weak or ambiguous.

Removing the low-level feature concatenation from the boundary head causes a larger decrease to 72.3% mAcc and 60.4% mIoU. While intermediate and high-level features provide strong semantics, their reduced spatial resolution limits fine-boundary recovery. The low-level feature F_1^{RT} preserves local structures and spatial transitions, providing complementary detail for small objects, thin structures, and adjacent categories.

Fig. 10 further illustrates the fine-grained segmentation capability of the model. For regular-sized objects, BiSANet better preserves detailed boundaries, such as vehicle-wheel structures, and produces contours closer to the ground truth. For distant small targets, including persons, cars, and bicycles, it also yields more complete and accurate predictions with fewer missed or fragmented regions than the compared methods. These results further support the role of SED in combining low-level spatial details with high-level semantic information for improved boundary recovery and small-object segmentation.

A.6 Efficiency Analysis

Table 11 compares BiSANet with existing RGB-T segmentation methods in terms of accuracy and computational efficiency. The default ViT-L configuration achieves the highest mIoU of 61.2%, outperforming the second-best USformer by 0.9 percentage points, but runs at 11.4 FPS. Replacing ViT-L with ViT-B reduces the model size to 200.3 M parameters and increases the speed to 19.0 FPS, while mIoU decreases by only 0.9 percentage points to 60.3%.

A more compact variant adopts the ViT-T encoder from MobileSAM (Zhang et al., 2023a), which replaces the original SAM image encoder with TinyViT (Wu et al., 2022a). This configuration further reduces the model size to 109.0 M parameters and reaches 20.1 FPS, while retaining 60.0% mIoU.



Figure 10: Qualitative comparison of fine-boundary and small-object segmentation on MFNet.

Table 11: Comparison of segmentation accuracy and computational efficiency on MFNet. Computational efficiency is measured using 480×640 RGB-T inputs on an NVIDIA RTX 4080 GPU.

Methods	FLOPs(G)↓	Params(M)↓	FPS↑	mAcc (%)↑	mIoU (%)↑
EGFNet_{23} (Dong et al., 2023)	201.3	62.8	20.1	72.7	54.8
LASNet_{23} (Li et al., 2023)	233.8	93.6	41.9	75.4	54.9
CMX-B4_{23} (Zhang et al., 2023b)	134.3	139.9	26.9	–	59.7
SGFNet_{23} (Wang et al., 2023)	143.7	125.3	13.9	73.6	57.6
MDBFNet_{25} (Liang et al., 2025)	249.3	97.9	–	75.3	57.8
AGFNet_{25} (Zhou et al., 2025)	219.1	72.2	21.2	74.5	58.6
MiLNet_{25} (Liu et al., 2025)	136.3	92.3	34.8	73.2	60.1
SAR_{26} (Zhao et al., 2026)	160.2	242.9	–	72.5	59.4
USformer_{26} (Yang et al., 2026)	147.7	261.2	–	72.5	60.3
Ours (SAM ViT-L)	556.4	419.8	11.4	73.2	61.2
Ours (SAM ViT-B)	178.3	200.3	19.0	72.1	60.3
Ours (MobileSAM ViT-T)	181.7	109.0	20.1	71.4	60.0

Overall, BiSANet can flexibly balance segmentation accuracy and efficiency by selecting different semantic encoders: ViT-L favors maximum accuracy, while ViT-B and ViT-T provide substantially lighter and faster alternatives with only modest performance degradation.

A.7 Failure Cases Study

Although BiSANet achieves strong overall performance, Fig. 11 shows several challenging cases. In the first two rows, the Bike and Car occupy only a few pixels and have weak contrast with the

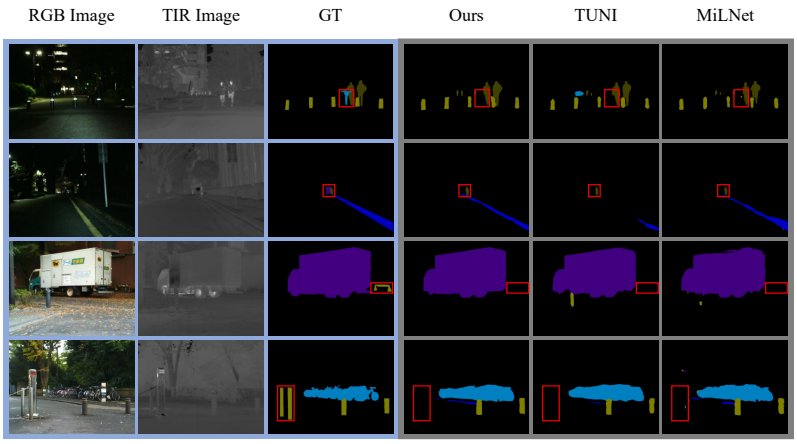


Figure 11: Representative failure cases of TUNI, MiLNet, and BiSANet. Red boxes highlight challenging or incorrectly segmented regions.

surroundings. Their limited spatial cues can be further weakened by repeated downsampling, leading to partial or complete misses. Similar errors also occur in competing methods, suggesting that extremely small objects under poor illumination remain a general challenge for RGB-T segmentation.

The third and fourth rows show failures on the *Car Stop* category. In one case, the target has an atypical appearance and lies close to a much larger vehicle, whose dominant response can suppress the small neighboring region. In the other, weak RGB and thermal contrast makes the target difficult to distinguish from roadside structures or background. These cases indicate that BiSANet remains limited when targets are extremely small, rare, or ambiguous in both modalities. Future work may explore higher-resolution and scale-aware features, hard-example mining, modality-degradation augmentation, and uncertainty-aware fusion to improve robustness to such challenging cases.