CFEFNet: Cross-Modal Feature Enhancement and Fusion Network for RGB-T Semantic Segmentation

Ziyang Zhang, Lei Sun*
School of Systems Science and Engineering, Sun Yat-sen University, Guangzhou, China
zhangzy375@mail2.sysu.edu.cn, sunlei8@mail.sysu.edu.cn

***Abstract*—Semantic segmentation based on RGB images and thermal infrared (TIR) images can enhance the robustness of scene parsing and improve segmentation performance. However, existing RGB-thermal (RGB-T) segmentation methods still inadequately exploit the complementary information between modalities and exhibit limited performance in segmenting targets of diverse scales and shapes in complex scenes. Moreover, cross-attention mechanism incurs considerable computational overhead when modeling global dependencies. To address these issues, we propose a novel Cross-Modal Feature Enhancement and Fusion Network (CFEFNet). Specifically, we design a Bidirectional Feature Enhancement (BFE) module to bridge the gap between two modalities. Then, a Frequency-Domain Cross-Attention (FDCA) module is proposed to efficiently promote cross-modal information exchange. Furthermore, we introduce a Spatially-Aware Feature Integration (SAFI) module to capture multi-scale information and anisotropic spatial context. Experimental results on two public datasets MFNet and PST900 show that our proposed CFEFNet achieves state-of-the-art performance. The code is available at https://github.com/ziyangbest/CFEFNet.**

***Index Terms*—RGB-T semantic segmentation, Cross-attention mechanism, Frequency-domain, Feature fusion**

I. Introduction

Semantic segmentation is a fundamental computer vision task that assigns a class label to each pixel for scene understanding and holds significant application value in key areas such as autonomous driving [1] and robotic navigation [2]. With the rapid development of deep learning, semantic segmentation has achieved notable improvements [3]–[6]. However, RGB-based segmentation methods are adversely affected by lighting variations and exhibit severely compromised robustness in challenging environments like darkness or overexposure. Since TIR images are formed based on the object's own thermal radiation, they are inherently insensitive to lighting variations, thereby effectively overcoming the limitations of RGB images. As a result, recent studies [7], [8] have leveraged the combination of TIR and RGB images to improve semantic segmentation performance.

Most existing RGB-T segmentation methods typically employ an encoder-decoder architecture, integrating RGB and TIR information through feature-level fusion [9]–[11]. Simple operations like addition and channel concatenation [7], [12] often fail to capture complementary information from RGB and TIR images. To solve these problems, some effective fusion strategies such as attention-guided fusion and gate-based weighted mechanisms are employed to improve segmentation performance. ABMDRNet [13] designs an adaptive weighted bidirectional modality discrepancy reduction network to recalibrate features using a channel-weighted fusion module. MFTNet [14] applies a self-attention mechanism within the multispectral fusion encoder to address both the intra-spectral correlations and inter-spectral complementarity of RGB-T. AGFNet [15] explores the complementarity between the two-modal features through a gating mechanism, and further enhance the interaction of two-modal features using channel and spatial attention mechanisms.

Despite the improvements made, current methods often overlook the differential information between two modalities, fail to utilize the complementary features to effectively enhance feature representation, resulting in models overly relying on information from a single modality and leading to insufficient integration of multi-modal information. Furthermore, the cross-attention mechanism which is adopted by [16]–[19] achieves excellent performance in RGB-T semantic segmentation through the exchange of information across modalities. However, since it models global dependencies in the spatial domain, the quadratic growth in computation imposed by cross-attention mechanism limits its application to high-resolution image features. Therefore, effectively balancing the model's performance with its spatial and temporal complexity is a challenging task. In addition, some methods [20]–[22] capture multi-scale semantic and contextual information by applying dilated convolutions with different rates on deep features which have powerful representation for scene understanding. Although these methods exhibit strong performance in recognizing objects of varying scales, their reliance on square receptive fields limits the flexibility in adapting to various orientation changes in complex scenes. For instance, it is difficult to recognize banded objects, highlighting the need to further leverage the complementary information between two modalities from deep features to improve segmentation accuracy.

Motivated by the above observations, this paper proposes a cross-modal feature enhancement and fusion method. To bridge the modal gap, we propose a Bidirectional Feature Enhancement (BFE) module, which leverages the feature differences between RGB and TIR modalities. By converting these differential signals into dynamic guidance information for fusion, the module encourages the network to focus on

*Corresponding Author

the most complementary regions across modalities, thereby achieving bidirectional enhancement of cross-modal feature representations. Next, we propose a Frequency-Domain Cross-Attention (FDCA) module, which introduces the Fast Fourier Transform to balance the efficiency and performance of global dependencies modeling, enabling guided fusion of modality-specific details. Subsequently, a Spatially-Aware Feature Integration (SAFI) module is designed to simultaneously capture multi-scale information and anisotropic spatial context, significantly improving the segmentation accuracy of objects of varying scales and shapes in complex scenarios.

Our contributions can be summarized as follows:

- We propose a novel RGB-T semantic segmentation method that fully realizes cross-modal feature enhancement and fusion, while effectively balancing the performance and computational complexity of the model.
- We propose a BFE module that converts modal differences into dynamic signals to guide feature fusion, leveraging the complementary information between two modalities to enhance the feature representation capabilities of each modality.
- We propose a FDCA module to efficiently promote cross-modal information exchange and facilitate feature fusion. Aditionally, we propose a SAFI module to simultaneously capture multi-scale information and anisotropic spatial context.
- Experiments on the MFNet and PST900 datasets demonstrate that our proposed network achieves excellent segmentation performance compared with other state-of-the-art methods.

II. Related Works

A. RGB Semantic Segmentation

Early semantic segmentation networks heavily relied on manually crafted simple features, which had limited representational capability and were unable to fully capture the complex semantic information in images. The emergence of convolutional neural networks has driven numerous breakthroughs in the field of semantic segmentation. FCN [3] is the first end-to-end fully convolutional semantic segmentation network, laying the foundation for modern semantic segmentation. Subsequently, many studies have made significant progress. UNet [4] achieves efficient multi-scale feature fusion through its unique symmetric encoder-decoder architecture and cross-layer skip connections. DeepLab [5] uses dilated convolution to capture contextual information at different scales. SegFormer [6] further unifies a hierarchical Transformer encoder with a lightweight multilayer perceptron (MLP) decoder, achieving powerful segmentation performance.

B. RGB-T Semantic Segmentation

RGB images make it difficult to accurately segment target contours and objects, while TIR images have an inherent advantage in perceiving targets based on radiation differences. Therefore, it is essential to integrate the complementary information of both to enhance segmentation performance.

In recent years, RGB-T semantic segmentation methods based on encoder-decoder architectures have developed rapidly. MFNet [7] pioneeringly combined RGB and TIR modalities for this task and released the MFNet dataset. On this basis, a series of fusion strategies have been proposed: RTFNet [12] uses simple element-wise addition to fuse multi-modal features. To further optimize the fusion process, EGFNet [23] introduces edge-aware guidance and deep supervision with semantic boundary maps, significantly improving the segmentation accuracy of object boundaries. SGFNet [24] introduces a semantic guidance head from the TIR branch coupled with a multi-modal coordination and enhancement unit in the RGB branch to progressively fuse and semantically enhance multi-scale features. CMX [16] achieves comprehensive feature calibration through channel and spatial rectification, and enhances global information exchange using a cross-attention mechanism. SCRNet [22] focuses on using global contextual information to correct semantic conflicts. It decouples and enhances complementary information through similarity measurement and channel attention. 3CNet [25] adopts a "correction-then-fusion" strategy that first aligns cross-modal features via orthogonal and spatial attention.

Although these methods have advanced RGB-T segmentation technology, there are still issues that need to be addressed in complementary information mining, global dependencies modeling and robustness to object scale and shape variations.

III. Method

A. Overall Architecture

Figure 1 shows the overall architecture of our proposed method. Paired RGB and TIR images are fed into two parallel encoder branches to extract their hierarchical features, yielding multi-scale representations denoted as $\{F_R^i\}_{i=1}^4$ and $\{F_T^i\}_{i=1}^4$, where the feature resolutions correspond to $\{1/4, 1/8, 1/16, 1/32\}$ of the input size. The BFE module is proposed to enhance the feature representation of each modality by leveraging cross-modal complementary information. The FDCA module efficiently promotes cross-modal information exchange between two modalities. The SAFI module is designed to capture multi-scale information and anisotropic spatial context. Finally, the fused features from different levels are fed into the MLP decoder to generate the segmentation map.

B. Bidirectional Feature Enhancement Module

RGB images generally perform better in terms of color and texture, while TIR images have advantages in temperature and thermal radiation. By fully leveraging the complementary information of the two modalities, it is possible to enhance feature fusion and improve the model's robustness to environmental changes. Therefore, we design a Bidirectional Feature Enhancement (BFE) module, as shown in Fig. 2. Concretely, for a given pair of RGB-T features F_R^i and F_T^i, We first apply both global average pooling (GAP) and global max pooling (GMP) to the feature difference, thus comprehensively capturing differential information between

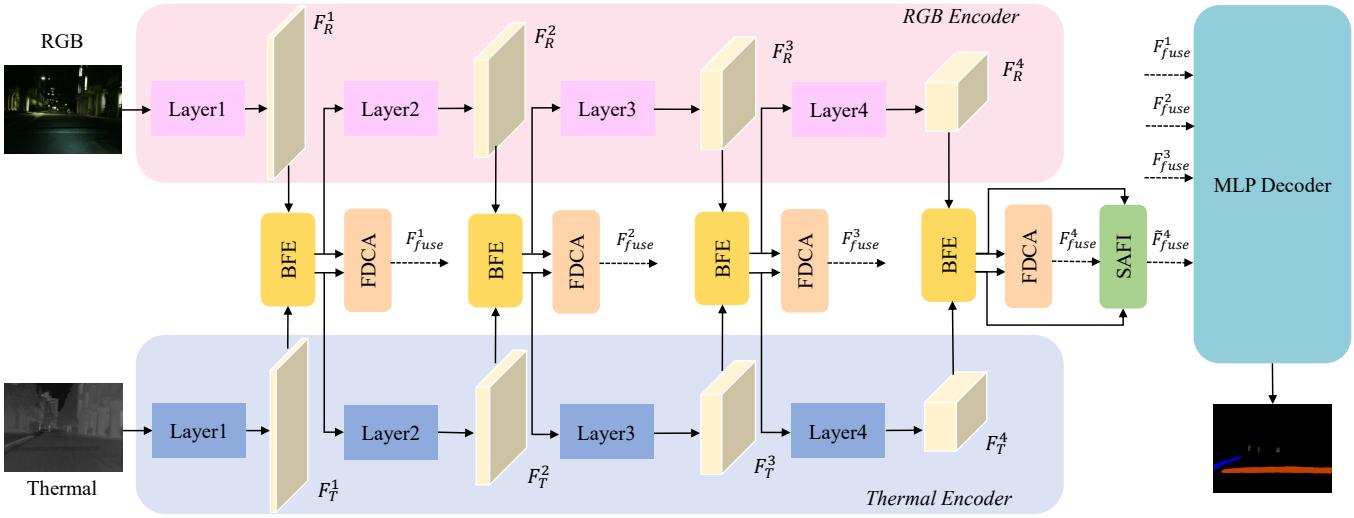

Fig. 1. The overall architecture of our proposed CFEFNet. It consists of a dual-branch encoder, an MLP decoder and three key modules: Bidirectional Feature Enhancement (BFE) module, Frequency-Domain Cross-Attention (FDCA) module and Spatially-Aware Feature Integration (SAFI) module.

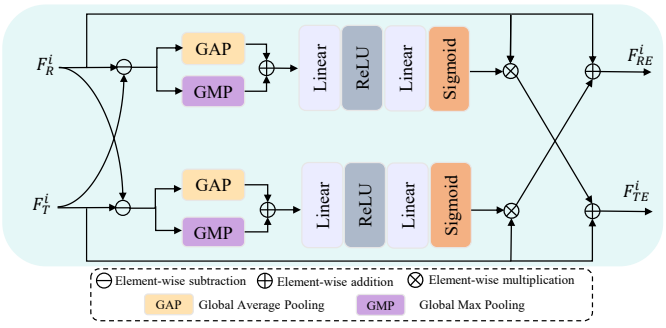

Fig. 2. The structure of the proposed BFE module.

modalities, and then process the combined output through a multilayer perceptron (MLP) and a sigmoid activation function to obtain the modality-specific weights. The process can be formulated as:

$$W_R^i = \sigma(\text{MLP}(\text{GAP}(F_R^i - F_T^i) + \text{GMP}(F_R^i - F_T^i))), \quad (1)$$

$$W_T^i = \sigma(\text{MLP}(\text{GAP}(F_T^i - F_R^i) + \text{GMP}(F_T^i - F_R^i))), \quad (2)$$

where σ represents the sigmoid activation function, MLP is a Multilayer Perceptron which contains linear mapping and ReLU activation function. Then, we obtain the enhanced features F_{RE}^i and F_{TE}^i by separately adding the weighted features from the other modality to the original ones, the procedure is as follows:

$$F_{RE}^i = F_R^i + W_T^i \otimes F_T^i, \quad (3)$$

$$F_{TE}^i = F_T^i + W_R^i \otimes F_R^i, \quad (4)$$

where $\otimes$ denotes element-wise multiplication.

C. Frequency-Domain Cross-Attention Module

Cross-attention mechanism enables the RGB modality to actively retrieve thermal signatures from the TIR modality to enhance performance under harsh lighting conditions, while the TIR modality can acquire color and texture details from the RGB modality to improve contour recognition. To effectively

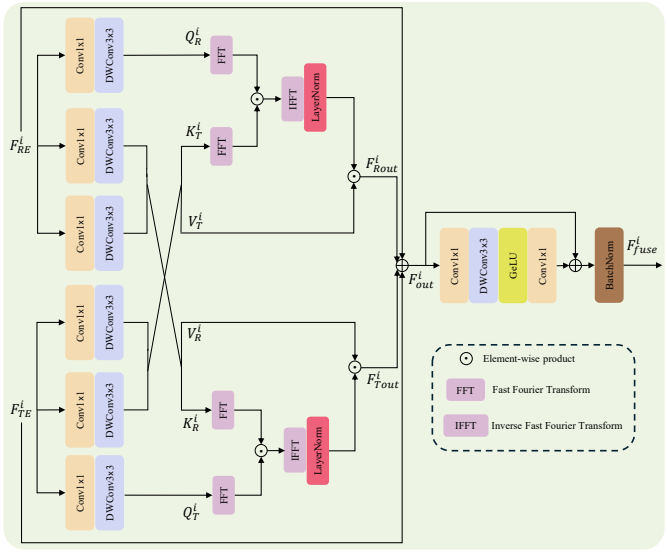

Fig. 3. The structure of the proposed FDCA module.

promote cross-modal information exchange and reduce computational complexity, inspired by the previous research [26], we propose a Frequency-Domain Cross-Attention (FDCA) module that transform tokens to the frequency domain via Fast Fourier transform (FFT) to compute dot-product attention. According to the convolution theorem, convolution in the spatial domain is equivalent to pointwise multiplication in the frequency domain. Thus, the space and time complexity can be reduced to $O(N)$ and $O(N \log N)$, respectively, with N representing the total number of pixels. As depicted in Fig. 3, to be specific, given the enhanced features F_{RE}^i and F_{TE}^i,

we initially apply the 1×1 convolution and 3×3 depthwise convolution to generate query Q, key K, and value V for each modality separately. Next, the query Q of one modality and the key K of the other modality are projected into the frequency domain via the Fast Fourier Transform (FFT), where their correlations are estimated through element-wise product and then transformed back into the spatial domain via the Inverse Fast Fourier Transform (IFFT) to obtain the attention weights. The weights are then used to perform weighted aggregation of the value V from the same modality as K via element-wise product, resulting the attended features F^i_{Rout} and F^i_{Tout}. The above implementations are described as follows:

$$Q^i_R, K^i_R, V^i_R = \text{DWConv}_3\left(\text{Conv}_1(F^i_{RE})\right), \tag{5}$$

$$Q^i_T, K^i_T, V^i_T = \text{DWConv}_3\left(\text{Conv}_1(F^i_{TE})\right), \tag{6}$$

$$F^i_{Rout} = \left(\text{LN}\left(\mathcal{F}^{-1}\left(\mathcal{F}(Q^i_R)\overline{\mathcal{F}(K^i_T)}\right)\right)V^i_T\right), \tag{7}$$

$$F^i_{Tout} = \left(\text{LN}\left(\mathcal{F}^{-1}\left(\mathcal{F}(Q^i_T)\overline{\mathcal{F}(K^i_R)}\right)\right)V^i_R\right), \tag{8}$$

where Conv_1, DWConv_3, and LN represent the 1×1 convolution, 3×3 depthwise convolution and layer normalization. $\mathcal{F}$, $\overline{\mathcal{F}}$ and $\mathcal{F}^{-1}$ denote the fast Fourier transform, the conjugate transpose operation and inverse fast Fourier transform. Finally, we add the attended features and the original features to produce the feature F^i_{out}:

$$F^i_{out} = F^i_{Rout} + F^i_{Tout} + F^i_{RE} + F^i_{TE}. \tag{9}$$

Finally, a residual block is incorporated to refine feature representation, improve the training efficiency and obtain the fused feature F^i_{fuse}, the process can be described as follows:

$$F^i_{fuse} = \text{BN}((\text{Conv}_1(\sigma_2(\text{DWConv}_3(\text{Conv}_1(F^i_{\text{out}}))))) + F^i_{\text{out}}), \tag{10}$$

where BN represents batch normalization, σ_2 is the GELU activation function.

D. Spatially-Aware Feature Integration Module

Informed by the atrous spatial pyramid pooling (ASPP) structure [5], which is widely adopted to capture multi-scale information, and recognizing that strip pooling [27] effectively models long-range dependencies crucial for anisotropic spatial context, we further introduce a Spatially-Aware Feature Integration (SAFI) module. The SAFI module is designed to utilize the rich semantic information from deep features to improve segmentation performance of targets with varying scales and shapes in complex scenes. As illustrated in Fig. 4, the input features F^4_{RE} and F^4_{TE} are element-wise multiplied and added, then they are concatenated and passed through a 1×1 convolutional layer to obtain the feature F^4_E as follows:

$$F^4_E = \sigma_3(\text{BN}(\text{Conv}_1(Cat((F^4_{RE} + F^4_{TE}), (F^4_{RE} \otimes F^4_{TE}))))), \tag{11}$$

where σ_3 is the ReLU activation function. Next, F^4_E is fed into two parallel processes. For multi-scale information capture, we concatenate the outputs of four parallel dilated convolutions

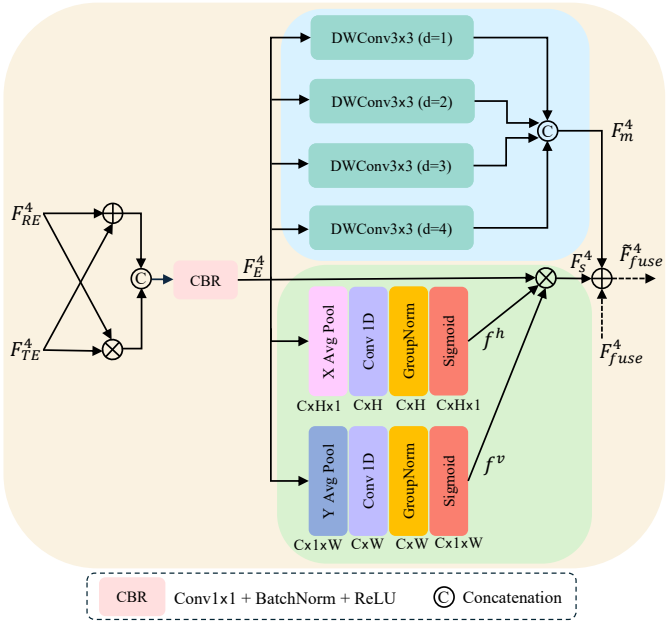

Fig. 4. The structure of the proposed SAFI module. "XAvgPool" and "YAvg Pool" refer to 1D horizontal global pooling and 1D vertical global pooling, respectively.

with dilation rates of 1, 2, 3 and 4, respectively, the operation can be defined as:

$$F^4_m = Cat(\text{DWConv}^d_3(F^4_E)), d = 1, 2, 3, 4, \tag{12}$$

where $\text{DWConv}^d_3(\cdot)$ represents the depthwise separable dilated convolution with kernel size 3, d is the dilated ratio. For anisotropic spatial context capture, inspired by the ELA module [28], we first decompose the input feature $F^4_E \in \mathbb{R}^{C \times H \times W}$ along the horizontal and vertical directions, obtaining $F^h \in \mathbb{R}^{C \times H \times 1}$ and $F^v \in \mathbb{R}^{C \times 1 \times W}$. Subsequently, We apply 1D convolution with a kernel size of 7 to each of them to further enhance the spatial information representation capacity, followed by group normalization and a sigmoid activation function, which yields the weights $f^h \in \mathbb{R}^{C \times H \times 1}$ and $f^v \in \mathbb{R}^{C \times 1 \times W}$. These weights are then multiplied by F^4_E to produce the output F^4_s, the above steps can be expressed as follows:

$$F^h_{c,i} = \frac{1}{W} \sum_{0 \le j < W} F_{c,i,j}, \tag{13}$$

$$F^v_{c,j} = \frac{1}{H} \sum_{0 \le i < H} F_{c,i,j}, \tag{14}$$

$$f^h = \sigma\left(\text{GN}\left(\text{Conv 1D}_7\left(F^h_{c,i}\right)\right)\right), \tag{15}$$

$$f^v = \sigma\left(\text{GN}\left(\text{Conv 1D}_7\left(F^v_{c,j}\right)\right)\right), \tag{16}$$

$$F^4_s = {F_E}^4 \otimes f^h \otimes f^v, \tag{17}$$

where c, i, and j denote the channel, height, width indices, respectively.

Then, the outputs of the two processes are added to obtain the final fused feature $\tilde{F}^4_{fuse}$, it can be expressed in the following formula:

$$\tilde{F}^4_{fuse} = F^4_s + F^4_m + F^4_{fuse}. \tag{18}$$

E. Loss Function

Weighted cross-entropy loss is employed in our proposed model, it is expressed by

$$L_{wce} = -\frac{1}{N}\sum_{i=1}^{N} w_i(y_i \log(\hat{y}_i)), \tag{19}$$

where N is the total number of pixels of the input image, y_i is the ground truth for the ith pixel, w_i denotes the class weight for the ith pixel and $\hat{y}_i$ represents the prediction for the ith pixel. Moreover, we introduce the Lovász-softmax loss [29], it can be represented as:

$$L_{lovasz} = \frac{1}{|C|}\sum_{c\in C}\overline{\Delta_{J_c}}(m(c)),$$
$$m(c) = \begin{cases} 1-\hat{y}_i(c) & \text{if } c = y(c) \\ \hat{y}_i(c) & \text{otherwise,} \end{cases} \tag{20}$$

where $|C|$ denotes the class number, $\overline{\Delta_{J_c}}$ defines the Lovász extension of the Jaccard index for class c. Then we combine it and the weighted cross-entropy loss function, the total loss is calculated by

$$L_{total} = L_{wce} + L_{lovasz}. \tag{21}$$

IV. Experiments

A. Experiment Setup

1) Datasets: Two widely used datasets, MFNet [7] and PST900 [8] are employed to train and evaluate our proposed network. The MFNet dataset consists of 1569 RGB and TIR urban scene image pairs, including 820 pairs of daytime images and 749 pairs of nighttime images. It comprises images with a resolution of 480 $\times$ 640 pixels and each labeled at the pixel level for nine distinct classes: Car, Curve, Person, Guardrail, Car Stop, Bike, Bump, Color Cone, and Background. We adhere to the same data partitioning strategy as employed in previous works [16], [25]. The PST900 dataset contains 894 pairs of RGB and TIR images with pixel-level labels for five classes: Hand-Drill, Survivor, Fire-Extinguisher, Backpack, and Background. Additionally, PST900 dataset is divided into the training dataset with 597 image pairs and testing dataset with 288 image pairs, all images are in a 720 $\times$ 1280 pixel resolution.

2) Implementation Details: The proposed model is implemented on a single NVIDIA V100 GPU using PyTorch framework. We use the Mix Transformer encoder (MiT-B2) as the feature extraction backbone, which is pre-trained on the ImageNet [30] dataset. The model is trained for 300 epochs using the AdamW optimizer with an initial learning rate of 6e-5 and a weight decay of 1e-2. The batch size is set to 4 for the MFNet dataset and 2 for the PST900 dataset. We employ a poly learning rate schedule with 10 warm-up epochs and apply data augmentation via random flipping and cropping.

3) Evaluation Metrics: In the experiments, we adopt two commonly used metrics for semantic segmentation evaluation: mean Accuracy (mAcc) and mean Intersection over Union (mIoU). mAcc measures the average classification accuracy across all classes, while mIoU quantifies the average overlap between predictions and ground truth masks for each class. They are calculated as follows:

$$\text{mAcc} = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k} p_{ij}}, \tag{22}$$

$$\text{mIoU} = \frac{1}{k}\sum_{i=1}^{k}\frac{p_{ii}}{\sum_{j=1}^{k}(p_{ij}+p_{ji}) - p_{ii}}, \tag{23}$$

where k is the number of object classes. p_{ii} represents the number of correctly classified pixels for class i, p_{ij} represents the number of pixels for class i misclassified as class j, and p_{ji} represents the number of pixels for class j misclassified as class i.

B. Comparison with State-of-the-art Methods

1) Quantitative Results on MFNet Dataset: We evaluate the proposed CFEFNet by conducting a comprehensive comparison with 14 existing state-of-the-art methods in RGB-T semantic segmentation. The quantitative results on the MFNet dataset are presented in Table I. Our method demonstrates a significant improvement over previous SOTA methods, achieving 59.8% mIoU and 76.4% mAcc. Specifically, our CFEFNet outperforms the second-best MCLNet by margins of 0.4% in mIoU and 2.2% in mAcc. Besides, our method achieves competitive segmentation performance across all categories, especially for the categories of "Car Stop" and "Guardrail", which mainly exhibit strip-like and banded structures. It obtains the highest IoU in both categories, with scores of 45.1% and 20.6%, respectively.

2) Quantitative Results on PST900 Dataset: To further validate the effectiveness and robustness of our CFEFNet, we conduct quantitative analysis on the PST900 dataset to verify its performance in dark and complex underground environments. As presented in Table II, the experimental results show that our CFEFNet achieves the highest mIoU of 87.1% and attains a competitive second place of 94.0% in mAcc among all compared methods, demonstrating consistent advantages across all categories. Notably, it ranks first in both Acc and IoU for the "backpack" and "fire-extinguisher" categories. Overall, these results reveal that our CFEFNet exhibits strong generalization ability.

3) Qualitative Analysis: We also provide a qualitative visualization on the MFNet dataset using some representative daytime and nighttime scenes as examples, as presented in Fig. 5. The first three rows show that our method achieves superior segmentation performance in dark night scenarios where the foreground and background are difficult to distinguish. For instance, in the third row, our method can accurately segment the small distant categories of "Person" and "Bike", while competing methods struggle to detect them. As shown in the last three rows, our method maintains robust performance

TABLE I
QUANTITATIVE RESULTS 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		
$MFNet_{17}$ [7]	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
$RTFNet_{19}$ [12]	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
$ABMDRNet_{21}$ [13]	94.3	84.8	90.0	69.6	75.7	60.3	64.0	45.1	44.1	33.1	31.0	5.1	61.7	47.4	66.2	50.0	69.5	54.8
$EGFNet_{22}$ [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
$SGFNet_{23}$ [24]	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	51.1	73.4	55.0	73.6	57.6
CMX_{23} [16]	-	89.4	-	74.8	-	64.7	-	47.3	-	30.1	-	8.1	-	52.4	-	59.4	-	58.2
$IGFNet_{23}$ [17]	93.2	88.0	83.4	74.0	71.8	62.7	67.6	48.2	45.4	36.0	**68.5**	14.2	58.8	52.4	68.3	57.5	72.9	59.0
$MDBFNet_{24}$ [10]	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
$DHFNet_{24}$ [9]	93.8	88.5	81.3	71.7	76.8	63.9	66.5	47.0	51.3	39.7	46.7	10.6	60.6	52.0	66.9	59.2	71.4	59.0
$DCIT_{24}$ [18]	-	**89.5**	-	75.2	-	**66.4**	-	47.2	-	35.8	-	10.1	-	49.9	-	**59.6**	-	59.1
$SCRNet_{25}$ [22]	95.2	87.9	90.3	74.1	73.4	62.7	63.3	47.4	41.5	36.2	49.8	7.5	60.6	53.2	66.6	54.4	71.1	58.0
$AGFNet_{25}$ [15]	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
$SICFNet_{25}$ [11]	95.2	89.4	89.0	74.9	76.7	64.4	64.6	**49.6**	44.2	34.7	42.3	7.5	63.7	56.5	76.7	57.8	72.4	59.1
$MCLNet_{25}$ [19]	94.3	88.6	86.2	74.0	79.4	63.3	73.3	48.0	**55.7**	42.4	40.5	11.4	60.4	53.5	**79.2**	55.4	74.2	59.4
Ours	95.5	87.9	**91.8**	73.0	**83.5**	64.2	**78.8**	44.6	54.9	**45.1**	51.6	**20.6**	57.8	52.0	75.1	53.2	**76.4**	**59.8**

TABLE II
QUANTITATIVE RESULTS ON THE PST900 DATASET.

Methods	Background		Hand-Drill		Backpack		Fire-Extinguisher		Survivor		mAcc	mIoU
	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$RTFNet_{19}$ [12]	**99.8**	99.0	7.8	7.1	80.0	74.2	62.4	51.9	78.5	70.1	65.7	60.5
$PSTNet_{20}$ [8]	-	98.9	-	53.6	-	69.2	-	70.1	-	50.0	-	68.4
$ABMDRNet_{21}$ [13]	**99.8**	99.0	81.2	61.5	71.4	67.9	76.5	66.2	66.4	62.0	79.1	71.3
$EGFNet_{22}$ [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
$SGFNet_{23}$ [24]	**99.8**	99.4	94.0	76.7	90.4	89.4	75.6	85.4	82.7	76.7	91.2	82.8
$MDBFNet_{24}$ [10]	**99.8**	**99.6**	95.1	84.4	92.7	88.7	94.5	76.1	**89.1**	**80.7**	**94.2**	85.9
$3CNet_{25}$ [25]	-	99.4	-	84.3	-	83.2	-	68.7	-	70.1	-	81.6
$AGFNet_{25}$ [15]	**99.8**	99.5	94.0	80.3	90.7	85.3	91.0	80.2	83.5	78.7	91.8	84.8
$MCLNet_{25}$ [19]	**99.8**	**99.6**	95.6	**88.3**	93.9	90.2	93.3	78.3	83.0	77.6	93.1	86.8
Ours	**99.8**	**99.6**	95.1	80.1	**95.2**	**90.9**	**97.1**	**87.9**	82.8	77.1	94.0	**87.1**

in daytime environments. Consider the complex scenario in the fifth row which contains multiple and sparsely distributed objects, our method correctly identify all targets, while other methods produce incomplete predicted masks. In general, compared with the ground truth, our method obtains more precise predictions in various scenarios, demonstrating the advantages in fusing cross-modal features to boost segmentation performance.

4) Computational Complexity Analysis: To evaluate the performance and efficiency of the proposed model, we compare our CFEFNet with other methods that employ different backbones including ResNet and MiT in terms of parameters (Params) and computational complexity (FLOPs). As shown in Table III, with 55.3M parameters and 62.7G FLOPs, CFEFNet is more computationally efficient. Compared to CMX, IGFNet, DCIT and MCLNet, which also use the same Transformer-based backbone and incorporate the cross-attention mechanism, our model achieves higher mIoU and mAcc while maintaining the lowest computational cost as well as competitive parameters. To summarize, CFEFNet strikes a superior balance between segmentation accuracy and computational complexity, outperforming existing methods in both aspects.

TABLE III
MODEL COMPLEXITY ANALYSIS ON THE MFNET DATASET WITH PREVIOUS SOTA METHODS.

Backbone	Method	Params(M)	FLOPs(G)	mAcc	mIoU
ResNet	RTFNet	254.5	337.4	63.1	53.2
	EGFNet	62.8	201.3	72.7	54.8
	SGFNet	125.2	143.7	73.6	57.6
	MDBFNet	97.9	249.3	75.3	57.8
	AGFNet	72.2	219.1	73.2	58.6
MiT	CMX(B2)	66.6	66.9	-	58.2
	IGFNet(B2)	67.4	69.2	72.9	59.0
	DCIT(B2)	78.4	184.2	-	59.1
	MCLNet(B2)	**34.1**	82.6	74.2	59.4
	Ours	55.3	**62.7**	**76.4**	**59.8**

C. Ablation Study

In this section, we conduct ablation experiments on the MFNet dataset to evaluate the contribution of each component to the overall performance. The quantitative results are shown in Table IV. The baseline model consists of the dual-branch MiT encoder and the MLP decoder, with feature fusion performed by element-wise addition. When we add the BFE module, there is a noticeable improvement of 1.4%

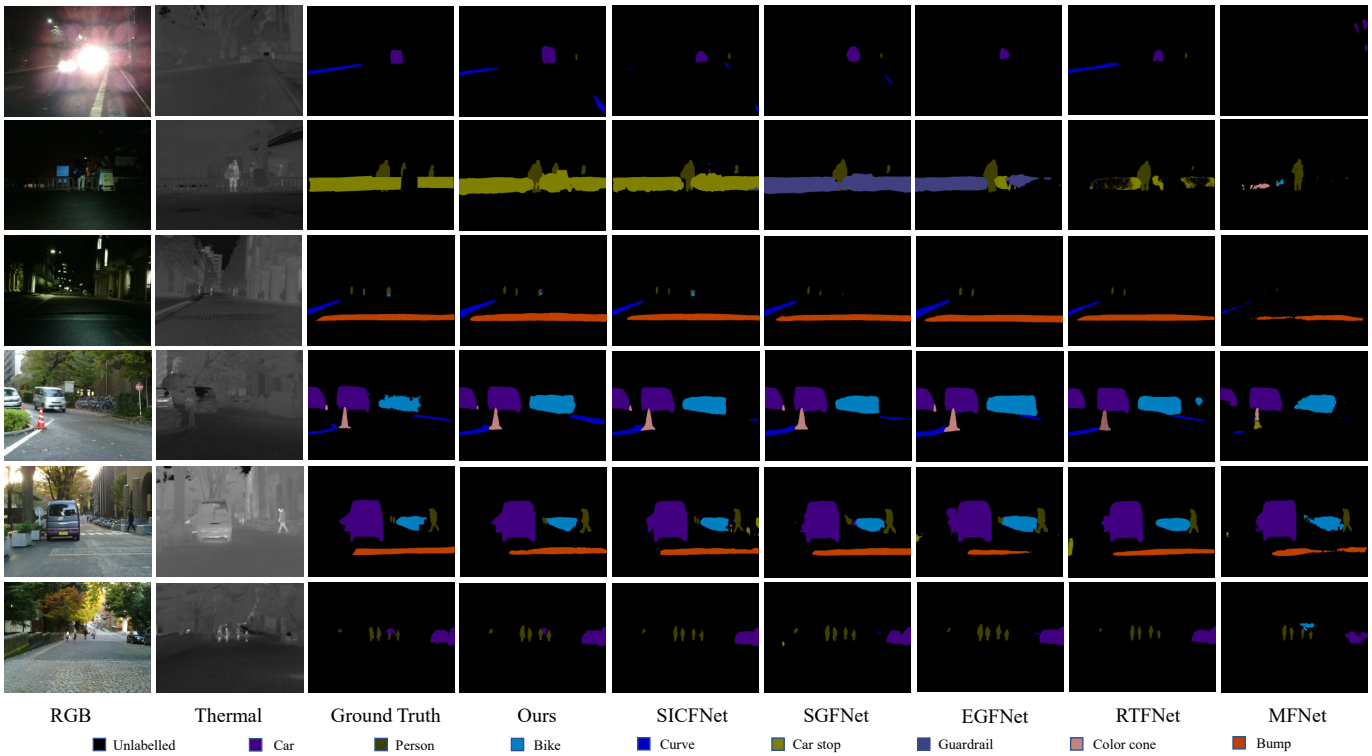

Fig. 5. Qualitative comparison of various methods on the MFNet dataset, covering both daytime and nighttime conditions.

mIoU and 4.0% mAcc, demonstrating the effectiveness of our proposed BFE module in utilizing complementary information between two modalities. Incorporating only the FDCA module brings an improvement of 1.8% mIoU and 4.1% mAcc with merely an increase of 1.4M parameters and 2.4G FLOPs compared to the baseline model, this indicates the effectiveness of the FDCA module in reducing computational complexity and facilitating global information exchange. When both BFE and FDCA modules are integrated, they yield a marked improvement of 2.4% mIoU and 4.7% mAcc. When we further introduce the SAFI module, which effectively captures multi-scale information and anisotropic spatial context, it leads to an additional gain of 0.6% mIoU and 3.4% mAcc. The perfect cooperation of the three modules results in our excellent complete model and achieves the best performance.

TABLE IV
ABLATION RESULTS OF DIFFERENT COMPONENTS.

BFE	FDCA	SAFI	Params(M)	FLOPs(G)	mAcc	mIoU
✗	✗	✗	50.0	60.0	67.9	56.8
✓	✗	✗	53.1	60.0	71.9	58.2
✗	✓	✗	51.4	62.4	72.0	58.6
✓	✓	✗	54.4	62.5	72.6	59.2
✓	✓	✓	55.3	62.7	**76.4**	**59.8**

V. CONCLUSION

This study presents a novel Cross-Modal Feature Enhancement and Fusion network for RGB-T Semantic Segmentation. We use the BFE module to harness the differential information between two modalities and achieve joint enhancement of cross-modal features. Then, we propose the FDCA module to promote cross-modal information exchange between two modalities and reduce the computational complexity of the cross-attention mechanism. Moreover, the SAFI module is employed to improve the segmentation accuracy for objects of varied scales and shapes. Evaluation results on two benchmark datasets show that our CFEFNet outperforms existing state-of-the-art methods in both segmentation performance and model efficiency. For future work, we plan to extend our approach to other modalities beyond RGB-Thermal, such as RGB-Depth and RGB-Event.

REFERENCES

[1] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, "Deep Multi-Modal Object Detection and Semantic Segmentation for Autonomous Driving: Datasets, Methods, and Challenges," *IEEE Transactions on Intelligent Transportation Systems*, vol. 22, no. 3, pp. 1341–1360, 2020.

[2] T. Guan, D. Kothandaraman, R. Chandra, A. J. Sathyamoorthy, K. Weerakoon, and D. Manocha, "GA-Nav: Efficient Terrain Segmentation for Robot Navigation in Unstructured Outdoor Environments," *IEEE Robotics and Automation Letters*, vol. 7, no. 3, pp. 8138–8145, 2022.

[3] J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 3431–3440.

[4] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[5] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 4, pp. 834–848, 2018.

[6] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," *Advances in neural information processing systems*, vol. 34, pp. 12 077–12 090, 2021.

[7] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. 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)*, 2017, pp. 5108–5115.

[8] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, "PST900: RGB-Thermal Calibration, Dataset and Segmentation Network," in *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 2020, pp. 9441–9447.

[9] H. Chen, Z. Wang, H. Qin, and X. Mu, "DHFNet: Decoupled Hierarchical Fusion Network for RGB-T dense prediction tasks," *Neurocomputing*, vol. 583, p. 127594, 2024.

[10] W. Liang, C. Shan, Y. Yang, and J. Han, "Multi-Branch Differential Bidirectional Fusion Network for RGB-T Semantic Segmentation," *IEEE Transactions on Intelligent Vehicles*, vol. 10, no. 4, pp. 2362–2372, 2025.

[11] B. Zhang, Z. Li, F. Sun, Z. Li, X. Dong, X. Zhao, and Y. Zhang, "SICFNet: Shared Information Interaction and Complementary Feature Fusion Network for RGB-T traffic scene parsing," *Expert Systems with Applications*, vol. 276, p. 127171, 2025.

[12] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-Thermal Fusion Network for Semantic Segmentation of Urban Scenes," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2576–2583, 2019.

[13] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "ABMDRNet: Adaptive-weighted Bi-directional Modality Difference Reduction Network for RGB-T Semantic Segmentation," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 2633–2642.

[14] H. Zhou, C. Tian, Z. Zhang, Q. Huo, Y. Xie, and Z. Li, "Multispectral Fusion Transformer Network for RGB-Thermal Urban Scene Semantic Segmentation," *IEEE Geoscience and Remote Sensing Letters*, vol. 19, pp. 1–5, 2022.

[15] X. Zhou, X. Wu, L. Bao, H. Yin, Q. Jiang, and J. Zhang, "AGFNet: Adaptive Gated Fusion Network for RGB-T Semantic Segmentation," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 5, pp. 6477–6492, 2025.

[16] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, "CMX: Cross-Modal Fusion for RGB-X Semantic Segmentation With Transformers," *IEEE Transactions on Intelligent Transportation Systems*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[17] H. Li and Y. Sun, "IGFNet: Illumination-Guided Fusion Network for Semantic Scene Understanding using RGB-Thermal Images," in *2023 IEEE International Conference on Robotics and Biomimetics (ROBIO)*. IEEE, 2023, pp. 1–6.

[18] K. Dai and S. Chen, "Dual-branch deep cross-modal interaction network for semantic segmentation with thermal images," *Engineering Applications of Artificial Intelligence*, vol. 135, p. 108820, 2024.

[19] T. Lu, H. Wu, W. Fu, L. Fang, and S. Li, "Bridging RGB-T Image Fusion and Semantic Segmentation via Multi-Task Collaborative Learning," *Information Fusion*, p. 103988, 2025.

[20] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, "DHFNet: dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation," *The Visual Computer*, vol. 40, no. 1, pp. 169–179, 2024.

[21] Z. Shen, J. Wang, Y. Weng, Z. Pan, Y. Li, and J. Wang, "ECFNet: Efficient cross-layer fusion network for real time RGB-Thermal urban scene parsing," *Digital Signal Processing*, vol. 151, p. 104579, 2024.

[22] S. Zhao, Z. Jin, Q. Jiao, Q. Zhang, and J. Han, "Resolving semantic conflicts in RGB-T semantic segmentation," *Pattern Recognition*, vol. 162, p. 111398, 2025.

[23] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 36, no. 3, 2022, pp. 3571–3579.

[24] Y. Wang, G. Li, and Z. Liu, "SGFNet: Semantic-Guided Fusion Network for RGB-Thermal Semantic Segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 12, pp. 7737–7748, 2023.

[25] Y. Guo, Z. Chen, X. Rong, C. Liu, L. Song, and Y. Li, "3CNet: Cross-modal cooperative correction network for RGB-T semantic segmentation," *Image and Vision Computing*, p. 105638, 2025.

[26] L. Kong, J. Dong, J. Ge, M. Li, and J. Pan, "Efficient Frequency Domain-based Transformers for High-Quality Image Deblurring," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 5886–5895.

[27] Q. Hou, L. Zhang, M.-M. Cheng, and J. Feng, "Strip Pooling: Rethinking Spatial Pooling for Scene Parsing," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 4002–4011.

[28] W. Xu, Y. Wan, and W. Zhao, "ELA: efficient location attention for deep convolution neural networks," *Journal of Real-Time Image Processing*, vol. 22, no. 4, pp. 1–14, 2025.

[29] M. Berman, A. R. Triki, and M. B. Blaschko, "The Lovasz-Softmax Loss: A Tractable Surrogate for the Optimization of the Intersection-Over-Union Measure in Neural Networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4413–4421.

[30] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.