SGFNet: Semantic-Guided Fusion Network for RGB-Thermal Semantic Segmentation

Yike WangLi, Gongyang Li, and Zhi Liu, *Senior Member, IEEE*

***Abstract*— Recently, semantic segmentation based on RGB and thermal infrared (TIR) images has become a research hotspot because of its stability in the weak light environment. However, most of the current methods ignore the differences between the two modalities of data and do not use semantic information in multi-modal fusion. In this paper, we propose a novel Semantic-Guided Fusion Network (SGFNet) for RGB-Thermal semantic segmentation, which makes full use of semantic information in the multi-modal fusion. Our SGFNet consists of an asymmetric encoder with TIR branch and RGB branch and a decoder. We concentrate on enhancing the multi-modal feature representation in the encoder with a pattern of fusion and enhancement. Specifically, considering that TIR images are stable under weak light conditions, we first propose a Semantic Guidance Head to extract semantic information in the TIR branch. In the RGB branch, we propose a Multi-modal Coordination and Distillation Unit to fuse multi-modal features first. Then, we propose a Cross-level and Semantic-guided Enhancement Unit to enhance the fused features with cross-level information and semantic information. We arrange these two units at all stages of the RGB branch to generate features with strong representation abilities at different levels. For the decoder, to obtain large receptive fields and fine edges, we improve the Lawin ASPP decoder by introducing edge information extracted from the low-level features, proposing the edge-aware Lawin ASPP decoder. With our encoder and decoder working together, our SGFNet can identify objects accurately and segment objects finely. Extensive experiments on the MFNet dataset demonstrate the superior performance of the proposed SGFNet compared with state-of-the-art methods. The code and results of our method are available at https://github.com/kw717/SGFNet.**

***Index Terms*— RGB-thermal semantic segmentation, asymmetric encoder, semantic guidance, multi-label learning.**

I. Introduction

SEMANTIC segmentation [1], [2], [3] has made great progress in the past decade and has been widely used in automatic driving [4], [5] and medical analysis [6].

Manuscript received 8 December 2022; revised 4 March 2023; accepted 28 May 2023. Date of publication 30 May 2023; date of current version 7 December 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62171269 and in part by the China Postdoctoral Science Foundation under Grant 2022M722037. This article was recommended by Associate Editor Z. Ding. *(Yike Wang and Gongyang Li contributed equally to this work.) (Corresponding author: Zhi Liu.)*

The authors are with the Key Laboratory of Specialty Fiber Optics and Optical Access Networks, Joint International Research Laboratory of Specialty Fiber Optics and Advanced Communication, Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China, and also with the School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: shuwangyike@shu.edu.cn; ligongyang@shu.edu.cn; liuzhisjtu@163.com).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2023.3281419.

Digital Object Identifier 10.1109/TCSVT.2023.3281419

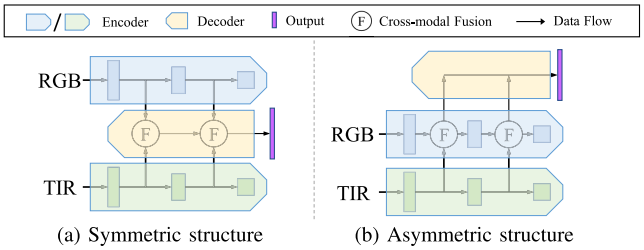

Fig. 1. Two structures for RGB-T semantic segmentation models, including (a) symmetric structure [10], [11], [12] and (b) asymmetric structure [12], [13], [14].

Due to the physical limitations of image sensors, most unimodal semantic segmentation models are limited in some application scenarios, such as low light environment, fog, and smoke. Therefore, multi-modal semantic segmentation has received much attention in recent years. Thermal infrared (TIR) image reflects the information of the surface temperature of objects, and can highlight the vehicles and pedestrians whose temperature is higher than the surrounding environment. Researchers introduce TIR images as supplements to improve the ability of models to handle the challenging urban driving scenes, which inspired the rise of RGB-thermal (RGB-T) semantic segmentation [7], [8], [9].

In general, existing RGB-T semantic segmentation models can be divided into two structures, *i.e.*, symmetric structure [10], [11], [12] and asymmetric structure [12], [13], [14], as shown in Fig. 1. The symmetric structure (Fig. 1(a)) employs two independent encoders on RGB and TIR images for feature extraction. Then it applies the same operations to RGB and TIR features for multi-modal fusion, and performs such multi-modal fusion level by level in the decoder. The asymmetric structure (Fig. 1(b)) also uses two encoders. Different from the former structure, it performs multi-modal fusion in the encoder, that is, it first employs an encoder on the TIR image, and then fuses TIR and RGB features in the RGB encoder.

However, both structures have their shortcomings. Most symmetric structure-based models operate on the two modalities in the same way, ignoring the differences between different modalities. At the same time, the level by level fusion is not conducive to the information interaction between layers in the feature extraction. Although asymmetric structure-based models treat RGB abd TIR features differently, most of them adopt the fusion strategy of element-wise summation. This fusion strategy without screening may introduce noise and is rough. In addition, most asymmetric structure-based

Fig. 2. Three typical scenes in RGB-T data. The first two scenes are nighttime scene with various unfavorable illumination conditions, and the last one is daytime scene. Please note the difference between the RGB image and the TIR image in the red box.

models do not design the network structure according to the characteristics of data in different modalities.

Reasonable use of the characteristics of data in different modalities is important for RGB-T semantic segmentation. In Fig. 2, we show three typical scenes in RGB-T data to illustrate this. Comparing TIR images with RGB images, in the first two nighttime scenes with various lighting conditions, we can find that TIR images can stably reflect objects and objects in RGB images are overwhelmed by light noise. In the last daytime scene with good lighting conditions, we can find that RGB images contain more appearance information, such as clear edges, than TIR images. Therefore, we believe that TIR images contain sufficient semantic information and may eliminate or reduce the interference of light noise in RGB images, and RGB images can also effectively supplement the details of objects in TIR images.

Motivated by the above observations, in this paper, we propose a novel *Semantic-Guided Fusion Network* (SGFNet) to make full use of the characteristics of data in different modalities for RGB-T semantic segmentation. Our SGFNet has the asymmetric structure. But different from previous methods, we focus on utilizing the stable semantic information of TIR images in the multi-modal fusion first, and then we focus on enhancing the fused features to further improve the representation ability and eliminate the interference in RGB features.

Concretely, following the asymmetric structure, given the RGB and TIR image pair, we first use an encoder to generate the feature embeddings from the TIR image. For these multi-level TIR features, we propose a Semantic Guidance Head (SGH) to integrate them all together to extract the stable semantic information. Next, we apply another encoder to the RGB image to generate the feature embeddings. Notably, we propose a Multi-modal Coordination and Distillation Unit (MCDU) to fuse the TIR and RGB features. Also, we propose a Cross-level and Semantic-guided Enhancement Unit (CSEU) to enhance the fused feature with the help of cross-level information and semantic information. We arrange the above two units in all levels of the RGB encoder to enhance the multi-modal fusion and eliminate interference in RGB features level by level. Finally, with the powerful features generated from different levels in the RGB encoder, we propose the edge-aware Lawin ASPP decoder for fine detail inference of objects by introducing an auxiliary edge supervision. In this way, our proposed SGFNet can generate satisfactory segmentation maps and outperform state-of-the-art methods on the MFNet dataset.

In summary, our main contributions are threefold:

- We make full use of the characteristics of data in different modalities, and propose a novel SGFNet for RGB-T semantic segmentation to utilize semantic information extracted from TIR images in the multi-modal fusion.
- We propose MCDU and CSEU to conduct the multi-modal fusion under semantic guidance to eliminate interference and generate features with strong representation, following the pattern of fusion and enhancement. MCDU fuses RGB features and TIR features to distillate the informative features. Then, CSEU utilizes the semantic information of TIR features and the enhanced features of previous level to further enhance the fused feature.
- We propose the edge-aware Lawin ASPP decoder to enhance the edge extraction and representation, which makes the generated segmentation maps have fine edges and structures.

II. Related Work

In this section, we review literature about RGB semantic segmentation and RGB-T semantic segmentation.

A. RGB Semantic Segmentation

In recent years, the RGB semantic segmentation based on convolutional neural network has attracted the attention of many researchers. Long et al. [1] proposed the pioneering end-to-end Fully Convolutional Network (FCN), which takes input of arbitrary size and produces output of corresponding size. Badrinarayanan et al. [2] introduced an encoder-decoder structure in SegNet, which learns image features through an encoder and segments objects through a decoder. Ronneberger et al. proposed a U-shaped U-Net [6] for medical image segmentation, which is similar to the encoder-decoder structure.

Inspired by the above works, RGB semantic segmentation has made great progress. Furthermore, researchers found that the larger receptive field is the key to obtain better segmentation results. Typically, Chen et al. [15] proposed the Atrous Spatial Pyramid Pooling (ASPP) in DeeplabV2, using atrous convolution to achieve larger receptive fields. Zhao et al. [16] proposed the Pyramid Scene Parsing Network (PSPNet), which captures different-region-based global context information by several pooling layers. Yu et al. [17] proposed BiSeNet to preserve spatial information with a spatial path and acquired sufficient large receptive field information with a contextual path.

In addition, self-attention mechanism [18], [19] has been widely used in the RGB semantic segmentation. Fu et al. [20] proposed Dual Attention Network (DANet) to build two

typical attention modules, *i.e.*, spatial attention and channel attention, via self-attention. Inspired by the recently developed vision transformer, Strudel et al. [21] built the Segmenter on Vision Transformer [19]. Xie et al. [22] improved the transformer decoder and proposed a lightweight decoder composed of multilayer perceptrons for hierarchical transformers [23], [24]. Yan et al. [25] introduced the large window attention which allows the local window to query a larger area of context window with only a little computation overhead, and further proposed a decoder with large window attention spatial pyramid pooling (Lawin ASPP).

By increasing the receptive field, researchers can successfully segment objects of different sizes with the extracted global information. The introduction of transformer further improves the performance of semantic segmentation. However, semantic segmentation using only RGB images cannot achieve good performance in poor lighting scenes. Using additional modalities other than RGB images can solve this problem, so researchers propose multi-modal semantic segmentation.

B. RGB-T Semantic Segmentation

With the development of imaging technology and the popularization of sensors, many researches introduce TIR images to solve the challenging scenes in RGB semantic segmentation. As mentioned above, RGB-T semantic segmentation models can be divided into two structures. The first one is the symmetric structure. Ha et al. [10] proposed the first symmetric structure-based RGB-T semantic segmentation model, named MFNet. MFNet used two identical independent encoders to extract features, and used the same strategy to fuse features of different modalities in the decoder. Following MFNet, there are many symmetric structure-based models being proposed. Zhang et al. [11] proposed ABMDRNet to reduce differences between two modalities by a sub-network during the RGB-thermal fusion so as to better fuse the features of two modalities by the same operation. Zhou et al. [27] proposed GMNet that divides features into two levels and fuses them by two different modules, and uses multilabel supervision in the training stage. In [28], Zhou et al. separated the multi-modal feature fusion from the encoder and decoder, and extracted global information from the high-level fused features via the parallel convolution structure to assist the decoding process.

The other one is the asymmetric structure. The main characteristic of this structure is that different modalities are treated differently, and the multi-modal feature fusion is performed in the encoder. For example, Sun et al. [13] proposed RTFNet with a two-encoders-one-decoder structure. Different from the symmetric structure, RTFNet fused TIR features extracted from the TIR encoder into the RGB encoder between the stages of encoder by element-wise addition. Based on the structure of RTFNet, Deng et al. [29] proposed FEANet which adds the feature-enhanced attention module before the element-wise addition. These models often adopt the element-wise addition for multi-modal feature fusion, which is simple and direct, but do not take into account the elimination of light noise.

Most of the above methods do not use semantic information, while some methods use the global semantic information in the decoding phase rather than the multi-modal fusion phase. In addition, most methods do not effectively utilize the differences between data in different modalities. In this paper, we make full use of the characteristics of data in different modalities, proposing SGFNet. Our SGFNet is based on the asymmetric structure, and improves the multi-modal fusion of the asymmetric structure by utilizing semantic information.

III. Proposed Method

In this section, we describe our SGFNet in deteil. First, the overview of our SGFNet is presented in Sec. III-A. Then, in Sec. III-B and Sec. III-C, we show the details of our MCDU and CSEU, respectively. In Sec. III-D, our edge-aware Lawin ASPP decoder is described. Finally, in Sec. III-E, we introduce the loss function.

A. Network Overview

In Fig. 3, we show the overall architecture of our proposed SGFNet, which is based on the asymmetric structure. We adopt an asymmetric encoder to extract features. The asymmetric encoder includes the TIR branch and the RGB branch, whose backbones are ResNet50 [26]. The structure of each block of these two branches is the same except that MCDU and CSEU are placed between blocks of the RGB branch. The five convolution blocks in the RGB branch and the TIR brach are denoted as R_n and T_n ($n = 1, 2, 3, 4, 5$), respectively. The extracted features of R_n and T_n are denoted as $\{\boldsymbol{F}_n^r, \boldsymbol{F}_n^t\} \in \mathbb{R}^{c_n \times h_n \times w_n}$.

In the asymmetric encoder, the TIR branch first extracts $\boldsymbol{F}_n^t$ from a TIR image, and then extracts the semantic information $\boldsymbol{S} \in \mathbb{R}^{c_1 \times h_1 \times w_1}$ from the above multi-level $\boldsymbol{F}_n^t$ in the SGH. As SGH shown in the upper right corner of Fig. 3, we reduce the channel number of $\boldsymbol{F}_5^t$ by convolution operation and upsample it, then add it to $\boldsymbol{F}_4^t$. The generated features are subjected to the same operation until it is added to $\boldsymbol{F}_1^t$. We take the output features as semantic information $\boldsymbol{S}$. To enhance the accuracy of semantic information $\boldsymbol{S}$, we impose a semantic supervision to SGH. We perform the feature enhancement on $\boldsymbol{F}_n^t$ through the channel attention and spatial attention [30], generating the enhanced TIR features $\hat{\boldsymbol{F}}_n^t \in \mathbb{R}^{c_n \times h_n \times w_n}$. Starting from R_1, we adopt the MCDU to fuse RGB features with the enhanced TIR features and semantic information. Specifically, we fuse $\boldsymbol{F}_1^r$ with $\hat{\boldsymbol{F}}_1^t$, generating $\hat{\boldsymbol{F}}_1^r$. We input $\hat{\boldsymbol{F}}_1^r$ into R_2, and then obtain $\boldsymbol{F}_2^r$. Different from the first stage, in the second stage, we not only fuse $\boldsymbol{F}_2^r$ with $\hat{\boldsymbol{F}}_2^t$ in MCDU, but also enhance the fused features by $\boldsymbol{S}$ and the cross-level $\hat{\boldsymbol{F}}_1^r$ in CSEU, generating $\hat{\boldsymbol{F}}_2^r$. The operations in the other three stages are the same as those in the second stage. In this way, we can obtain five-level $\hat{\boldsymbol{F}}_n^r \in \mathbb{R}^{c_n \times h_n \times w_n}$ with powerful multi-modal representation abilities.

We propose the edge-aware Lawin ASPP decoder [25] in our SGFNet. Different from the original Lawin ASPP decoder, we introduce the edge information into our decoder. Concretely, we perform the edge supervision to the low-level $\hat{\boldsymbol{F}}_1^r$ and $\hat{\boldsymbol{F}}_2^r$ to extract edge details, thus complementing the

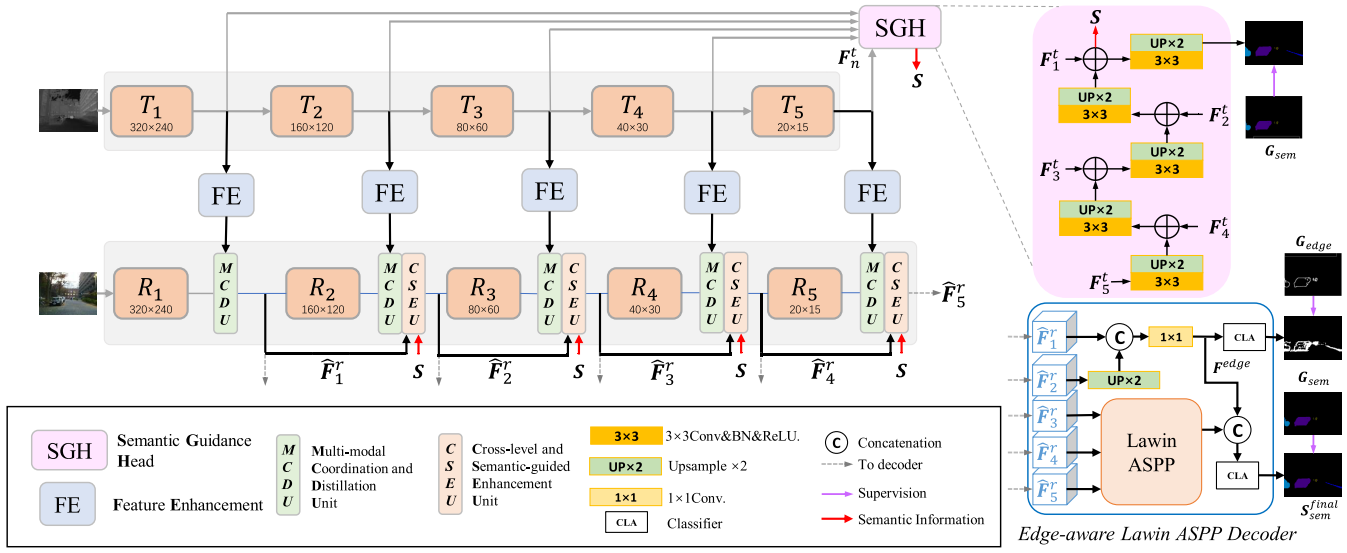

Fig. 3. The overall framework of our SGFNet, which is based on the asymmetric structure. The asymmetric encoder consists of the TIR branch and the RGB branch, whose backbones are ResNet50 [26]. SGFNet uses Multi-modal Coordination and Distillation Unit (MCDU) to fuse multi-modal features, and place it between each two convolution blocks of the RGB branch. Notably, SGFNet includes two special modules for generating and utilizing semantic information, *i.e.*, Semantic Guidance Head (SGH) and Cross-level and Semantic-guided Enhancement Unit (CSEU). SGH aggregates the thermal features $\boldsymbol{F}_n^t$ generated from the TIR branch to produce semantic information $\boldsymbol{S}$, and CSEU uses the semantic information to enhance multi-modal fusion. Based on the extracted fused features $\hat{\boldsymbol{F}}_n^r$ of the RGB branch, we segment objects in the edge-aware Lawin ASPP decoder.

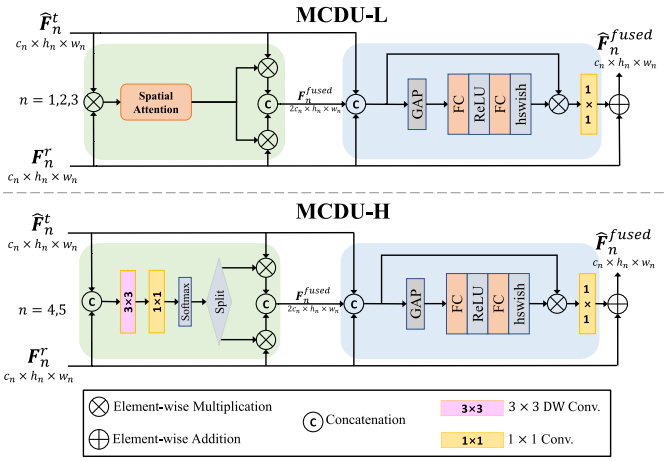

Fig. 4. Illustration of two variants of the Multi-modal Coordination and Distillation Unit. The upper one is for low-level features, and the lower one is for high-level features. The green part is the coordination part, and the blue part is the distillation part.

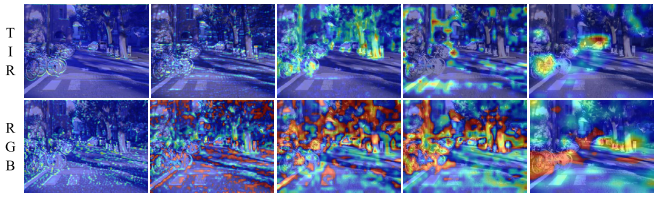

Fig. 5. Feature visualization of five-level TIR and RGB features, sorting from low level to high level. Here, we superimpose TIR and RGB features on the RGB image.

high-level information with the large receptive field. Based on the remaining high-level features of the encoder and the edge detail information, we infer all kinds of objects in the edge-aware Lawin ASPP decoder.

B. Multi-Modal Coordination and Distillation Unit

Multi-modal Coordination and Distillation Unit (MCDU) is responsible for multi-modal fusion between the stages of the RGB branch. As shown in Fig. 4, MCDU consists of two parts, that is, the coordination part (*i.e.*, the green box in Fig. 4) fuses multi-modal features in the spatial dimension and the distillation part (*i.e.*, the blue box in Fig. 4) purifies the original features and the fused features in the channel dimension. In order to complement the information of different modalities better, we design two variants of MCDU to adapt to different levels of features, *i.e.*, MCDU-L and MCDU-H, whose coordination parts are different.

1) Coordination Part: As the five-level TIR and RGB features visualized in Fig. 5, we observe that the activation regions of low-level features are widely distributed, while those of high-level features are more concentrated. Based on the above characteristics of features at different levels, we design two specific coordination parts in MCDU-L and MCDU-H, as shown in Fig. 4.

For the coordination part of MCDU-L, we focus on identifying the common activation regions in two modalities of chaotic distribution. Therefore, we propose MCDU-L for three low-level features ($n = 1, 2, 3$) to coordinate features of two modalities and activate potential targets of both modalities. We first multiply $\boldsymbol{F}_n^r$ with $\hat{\boldsymbol{F}}_n^t$ to locate the common activation regions. Then, we perform the spatial attention operation [30] to the multiplied features to obtain a shared attention map. Finally, we separately weight $\boldsymbol{F}_n^r$ and $\hat{\boldsymbol{F}}_n^t$ with the above shared attention map and concatenate them to get the fused

features $\boldsymbol{F}_n^{fused} \in \mathbb{R}^{2c_n \times h_n \times w_n}$, which can be computed as follows:

$$\boldsymbol{F}_n^{fused} = \mathrm{Cat}(\hat{\boldsymbol{F}}_n^t \otimes \mathrm{SA}(\hat{\boldsymbol{F}}_n^t \otimes \boldsymbol{F}_n^r), \boldsymbol{F}_n^r \otimes \mathrm{SA}(\hat{\boldsymbol{F}}_n^t \otimes \boldsymbol{F}_n^r)), \tag{1}$$

where $\mathrm{SA}(\cdot)$ is the spatial attention operation, $\otimes$ is the element-wise multiplication and $\mathrm{Cat}(\cdot)$ is the concatenation operation.

Since the visible objects in two modalities are different, the object location and semantic information in the high-level feature maps ($n = 4, 5$) are different. Thus, for the coordination part of MCDU-H, we focus on extracting the important complementary information of different concentrated regions in two modalities, which is different from MCDU-L. Concretely, in MCDU-H, we first concatenate $\hat{\boldsymbol{F}}_n^t$ and $\boldsymbol{F}_n^r$, and then adopt a depth-wise separable convolutional layer and a 1×1 convolutional layer to obtain feature maps with two channels. We generate two attention maps from the features maps through the Softmax activation function and splitting along the channel dimension by index slicing of tensor. In this way, the model can adaptively allocate attention to multi-modal features in a pixel-by-pixel manner, learning the informative features. After that, we adopt these two attention maps to modulate $\hat{\boldsymbol{F}}_n^t$ and $\boldsymbol{F}_n^r$, respectively, and integrate them through concatenation, getting the fused feature $\boldsymbol{F}_n^{fused} \in \mathbb{R}^{2c_n \times h_n \times w_n}$.

2) Distillation Part: The coordination part can make full use the characteristics of different modalities, but will bring noise when the multi-modal fusion is not ideal, which is ignored by most methods [13], [29]. Thus, we propose the distillation part in MCDU to purify the original features and the fused features in the channel dimension, extract effective information and reduce information redundancy.

As the blue box shown in Fig. 4, the distillation part of MCDU-L and MCDU-H is the same, and its inputs are $\boldsymbol{F}_n^{fused}$, $\hat{\boldsymbol{F}}_n^t$ and $\boldsymbol{F}_n^r$. We first concatenate $\boldsymbol{F}_n^{fused}$, $\hat{\boldsymbol{F}}_n^t$ and $\boldsymbol{F}_n^r$ as $\boldsymbol{F}_n^{Re} \in \mathbb{R}^{4c_n \times h_n \times w_n}$. Then we re-weight $\boldsymbol{F}_n^{Re}$ in the channel dimension, that is, we successively perform global average pooling operation and two fully connected layers on $\boldsymbol{F}_n^{Re}$ to get a channel-wise attention map, and adopt this map to modulate $\boldsymbol{F}_n^{Re}$ through the element-wise multiplication. We additionally add the original $\boldsymbol{F}_n^r$ to the modulated features by a short path, getting $\hat{\boldsymbol{F}}_n^{fused} \in \mathbb{R}^{c_n \times h_n \times w_n}$. In this way, we can distillate the valuable features from the fused features.

C. Cross-Level and Semantic-Guided Enhancement Unit

In the full-time environment targeted by RGB-T semantic segmentation, there are RGB images with extreme lighting conditions, which makes $\hat{\boldsymbol{F}}_n^{fused}$ generated from MCDU contain some light noise. As shown in Fig. 2, TIR images are stable and not disturbed by light conditions. Therefore, to solve this problem, we propose the Cross-level and Semantic-guided Enhancement Unit (CSEU), which extracts effective semantic information from TIR images to locate correct objects and eliminate noise disturbance in $\hat{\boldsymbol{F}}_n^{fused}$. In addition, we introduce the enhanced features of the previous level to facilitate cross-level interaction in CSEU.

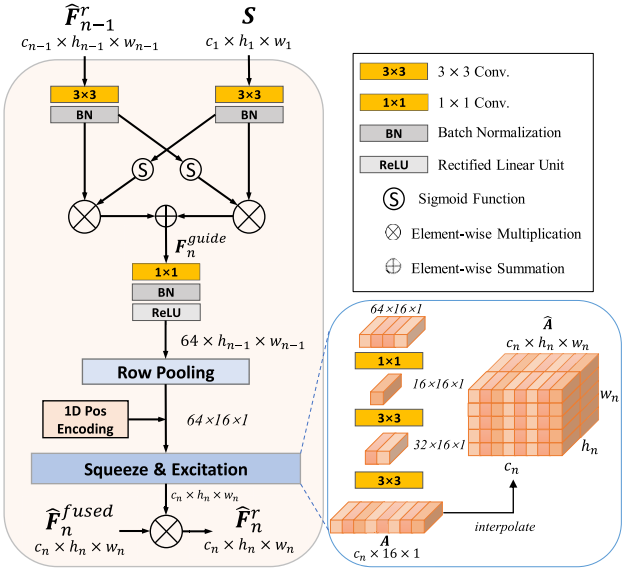

Fig. 6. Illustration of the Cross-level and Semantic-guided Enhancement Unit.

As shown in Fig. 6, the inputs of CSEU are $\hat{\boldsymbol{F}}_{n-1}^r$, $\boldsymbol{S}$, and $\hat{\boldsymbol{F}}_n^{fused}$. We use $\hat{\boldsymbol{F}}_{n-1}^r$ and $\boldsymbol{S}$ to guide the enhancement of $\hat{\boldsymbol{F}}_n^{fused}$. Since $\hat{\boldsymbol{F}}_{n-1}^r$ contains more texture and appearance information and $\boldsymbol{S}$ contains stable semantic information, we blend them with each other to generate guidance features that contain both appearance and semantics. Concretely, we first normalize the dimensions of $\hat{\boldsymbol{F}}_{n-1}^r$ and $\boldsymbol{S}$ to $64 \times h_{n-1} \times w_{n-1}$ by the convolutional layer for saving computing costs. Then, we perform the mutual modulation on these two normalized features, that is, modulating $\hat{\boldsymbol{F}}_{n-1}^r$ by $\boldsymbol{S}$ and modulating $\boldsymbol{S}$ by $\hat{\boldsymbol{F}}_{n-1}^r$. Finally, we add both modulated features to achieve the initial guidance features $\boldsymbol{F}_n^{guide} \in \mathbb{R}^{64 \times h_{n-1} \times w_{n-1}}$.

In addition, in the city scenes, the distribution of objects in height is significantly different [31]. Therefore, we introduce the vertical position prior information into $\boldsymbol{F}_n^{guide}$ to model the correlation of object positions, that is, we adjust the weight of each row in every channel according to the height-wise semantic information of $\boldsymbol{F}_n^{guide}$. Specifically, we first extract the height-wise information by a row-wise average pooling operation and compress the height from h_{n-1} to 16. Then, we add the sinusoidal positional encoding [18] to model the relative vertical position relationship. After that, we use a Squeeze&Excitation [32] operation to model interdependence between different vertical positions and channels, so as to get an attention map $\boldsymbol{A} \in \mathbb{R}^{c_n \times 16 \times 1}$. Here, we replace fully connected layers in Squeeze&Excitation with 3×3 convolutional layers to consider the relationship between adjacent rows better. Finally, we restore $\boldsymbol{A}$ to $\hat{\boldsymbol{A}} \in \mathbb{R}^{c_n \times h_n \times w_n}$ by the interpolation, and adopt it to effectively enhance $\hat{\boldsymbol{F}}_n^{fused}$, generating $\hat{\boldsymbol{F}}_n^r \in \mathbb{R}^{c_n \times h_n \times w_n}$. In this way, we enhance feature representation channel-wise and height-wise through semantic-guided attention.

In Fig. 7, we visualize $\hat{\boldsymbol{F}}_4^{fuse}$ and $\hat{\boldsymbol{F}}_4^r$ to illustrate the effectiveness of our CSEU. We observe that CSEU makes the

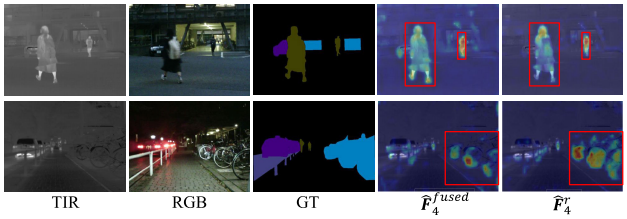

Fig. 7. Feature visualization of $\hat{\boldsymbol{F}}_4^{fuse}$ and $\hat{\boldsymbol{F}}_4^r$ in CSEU. Please note the features of pedestrians and bikes in the red boxes.

activated area more focused on the objects and eliminates the surrounding interference. In other words, CSEU enhances the representation of each category, making our model more robust for segmentation of different categories of objects.

D. Edge-Aware Lawin ASPP Decoder

As a dense prediction task, semantic segmentation needs to process objects of different sizes in a single image simultaneously, so multi-scale information interaction is one of the keys to improve performance. There are many related works exploring how to perform multi-scale information interaction, including using multi-scale receptive fields [16], [43] and increasing receptive fields [15], [44], [45]. We adopt the Lawin ASPP [25] as our decoder, which introduces multi-scale features through a large-window attention mechanism. Moreover, we improve the original Lawin ASPP decoder to the edge-aware Lawin ASPP decoder for clearly and finely segmenting objects.

The original Lawin ASPP decoder consists of a Lawin ASPP and a classifier. Its inputs are features at four different levels. The three relatively higher-level features are first integrated through the Lawin ASPP with multi-scale receptive fields. Then the classifier concatenates the lowest-level features and the integrated features for segmentation. As shown in the bottom right corner of Fig. 3, our edge-aware Lawin ASPP decoder receives the features at five levels, *i.e.*, $\hat{\boldsymbol{F}}_n^r$ $(n = 1, 2, 3, 4, 5)$. Here, we follow the original Lawin ASPP to fuse $\hat{\boldsymbol{F}}_3^r$, $\hat{\boldsymbol{F}}_4^r$, and $\hat{\boldsymbol{F}}_5^r$ through the Lawin ASPP, and get $\boldsymbol{F}^{high}$. Differently, since the low-level features, *i.e.*, $\hat{\boldsymbol{F}}_1^r$ and $\hat{\boldsymbol{F}}_2^r$, contain detailed texture information, we extract the edge features $\boldsymbol{F}^{edge}$ through the edge supervision to facilitate the segmentation of object structure. Then, based on $\boldsymbol{F}^{high}$ and $\boldsymbol{F}^{edge}$, the classifier generates the segmentation map $\boldsymbol{S}_{sem}^{final}$. The edge-aware Lawin ASPP decoder can be expressed as:

$$\begin{cases} \boldsymbol{F}^{high} = \text{LawinASPP}(\hat{\boldsymbol{F}}_3^r, \hat{\boldsymbol{F}}_4^r, \hat{\boldsymbol{F}}_5^r), \\ \boldsymbol{F}^{edge} = \text{Conv}(\text{Cat}(\hat{\boldsymbol{F}}_1^r, \text{UP}(\hat{\boldsymbol{F}}_2^r))), \\ \boldsymbol{S}_{sem}^{final} = \text{CLA}\big(\text{Cat}(\boldsymbol{F}^{high}, \boldsymbol{F}^{edge}\big), \end{cases} \tag{2}$$

where LawinASPP$(\cdot)$ is the Lawin ASPP, UP$(\cdot)$ is the up-sampling operation and CLA$(\cdot)$ is the classifier.

In this way, our SGFNet can segment objects with clear boundaries.

E. Loss Function

An effective loss function can improve the network performance. As shown in Fig. 3, both SGH and decoder have supervisions, so our overall loss function consists of two parts. The first part is a semantic supervision $\mathbb{L}_{\text{sem}}^{\text{sgh}}$. We adopt the widely used weighted cross-entropy loss for it. In the second part, we added two supervisions in the training process for our decoder, *i.e.*, edge supervision $\mathbb{L}_{\text{edge}}^{\text{D}}$ and semantic supervision $\mathbb{L}_{\text{sem}}^{\text{D}}$. Notably, following [27], we extract the edges of all classes except the background class from semantic labels to form the edge label. Specifically, we move a fixed-size window (5×5) over the semantic label, if the pixels in the window do not have the same semantic class, we regard the center pixel of the window as the edge. Then, we set the pixel values on the edges to one and others to zero, composing the edge label. We adopt the classic binary cross-entropy loss for $\mathbb{L}_{\text{edge}}^{\text{D}}$. For $\mathbb{L}_{\text{sem}}^{\text{D}}$, we use a hybrid loss function, including the weighted cross-entropy loss and the Lovasz-softmax loss, which is expressed as follows:

$$\mathbb{L}_{\text{sem}}^{\text{D}} = \big(\ell_{wbce}(\boldsymbol{S}_{sem}^{final}, \boldsymbol{G}_{sem}) + \ell_{Lovasz}(\boldsymbol{S}_{sem}^{final}, \boldsymbol{G}_{sem})\big)/2, \tag{3}$$

where $\ell_{wbce}(\cdot)$ is the weighted cross-entropy loss, $\ell_{Lovasz}(\cdot)$ is the Lovasz-softmax loss, and $\boldsymbol{G}_{sem}$ is the GT of semantic segmentation. Therefore, the total loss function can be expressed as:

$$\mathbb{L}_{\text{total}} = \alpha\mathbb{L}_{\text{sem}}^{\text{sgh}} + \beta(\mathbb{L}_{\text{sem}}^{\text{D}} + \mathbb{L}_{\text{edge}}^{\text{D}}). \tag{4}$$

We empirically set $\{\alpha, \beta\}$ as $\{1, 2\}$ to make the network pay more attention to the final segmentation map and improve the accuracy of edge extraction, which is beneficial to train our SGFNet.

IV. Experiments

A. Experiments Setup

1) Datasets: We use the public MFNet dataset [10] and the PST900 dataset [38] to train and evaluate the proposed SGFNet.

MFNet dataset contains 1569 RGB-T day and night city scene images. The images are captured by InfRec R500 in both visible and thermal infrared spectrum (814μm) with different lenses and sensors, and contain eight classes of common objects in urban driving scene, *i.e.*, car, person, bike, curve, car stop, guardrail, color cone and bump. Both RGB and TIR images have 480×640 pixels of spatial resolution. The dataset is divided into three parts: the training set (784 pairs of images), the validation set (393 pairs of images) and the test set (392 pairs of images).

PST900 dataset contains 894 synchronized and calibrated RGB and thermal image pairs with per pixel human annotations across four distinct classes, *i.e.*, hand-drill, backpack, fire-extinguisher and survivor. The spatial resolution of both RGB and TIR images is 1280×720. The dataset is divided into two parts: the training set (only 597 pairs of images, not 606 pairs) and the test set (288 pairs of images).

2) Implementation Details: We train and test our proposed SGFNet on an NVIDIA GTX 3090 GPU (24GB RAM) based on the PyTorch [46] platform. We choose MFNet dataset for training and testing, and follow the splitting manner given in [10]. For training, we do not crop the RGB image and

TABLE I

QUANTITATIVE COMPARISONS (%) ON THE TEST SET OF MFNET DATASET. THE VALUE 0.0 REPRESENTS THAT THERE ARE NO TRUE POSITIVES. '-' MEANS THAT THE AUTHORS DO NOT PROVIDE THE CORRESPONDING RESULTS. THE TOP THREE RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN **RED**, **BLUE** AND **GREEN**

Methods	Type	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		
DANet$_{19}$ [20]	RGB	89.8	84.5	65.2	55.0	76.5	62.6	44.2	33.4	32.7	27.4	2.8	0.9	46.6	41.9	56.0	44.5	57.0	49.7
DANet$_{19}$ [20]	RGB-T	91.3	71.3	82.7	48.1	**79.2**	51.8	48.0	30.2	25.5	18.2	5.2	0.7	47.6	30.3	19.9	18.8	55.2	41.3
HRNet$_{19}$ [33]	RGB	91.1	84.9	66.6	55.4	76.6	60.3	42.6	33.3	37.9	28.3	11.5	2.5	44.8	40.3	62.6	46.9	59.2	49.9
HRNet$_{19}$ [33]	RGB-T	90.8	86.9	75.1	67.3	70.2	59.2	39.1	35.3	28.0	23.1	12.1	1.7	50.4	46.6	55.8	47.3	57.9	51.7
FuseNet$_{17}$ [34]	RGB-D	81.0	75.6	75.2	66.3	64.5	51.9	51.0	37.8	28.7	15.0	0.0	0.0	31.1	21.4	51.9	45.0	52.4	45.6
D-CNN$_{18}$ [35]	RGB-D	85.2	77.0	61.7	53.4	76.0	56.5	40.2	30.9	9.9	29.3	22.8	6.4	32.9	30.1	36.5	32.3	55.1	46.1
ACNet$_{19}$ [36]	RGB-D	93.7	79.4	86.8	64.7	77.8	52.7	57.2	32.9	**51.5**	28.4	7.0	0.8	57.5	16.9	49.8	44.4	64.3	46.3
SA-Gate$_{20}$ [37]	RGB-D	86.0	73.8	80.8	59.2	69.4	51.3	56.7	38.4	24.7	19.3	0.0	0.0	56.9	24.5	52.1	48.8	58.3	45.8
MFNet$_{17}$ [10]	RGB-T	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
RTFNet50$_{19}$ [13]	RGB-T	91.3	86.3	78.2	67.8	71.5	58.2	**69.8**	43.7	32.1	24.3	13.4	3.6	40.4	26.0	73.5	**57.2**	62.2	51.7
RTFNet152$_{19}$ [13]	RGB-T	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	74.4	**55.7**	63.1	53.2
PSTNet$_{20}$ [38]	RGB-T	-	76.8	-	52.6	-	55.3	-	29.6	-	25.1	-	**15.1**	-	39.4	-	45.0	-	48.4
MMNet$_{21}$ [39]	RGB-T	-	83.9	-	69.3	-	59.0	-	43.2	-	24.7	-	4.6	-	42.2	-	50.7	62.7	52.8
MLFNet$_{21}$ [12]	RGB-T	-	82.3	-	68.1	-	**67.3**	-	27.3	-	30.4	-	**15.7**	-	**55.6**	-	40.1	-	53.8
FuseSeg$_{21}$ [14]	RGB-T	93.1	**87.9**	81.4	71.7	78.5	**64.6**	68.4	44.8	29.1	22.7	**63.7**	6.4	55.8	46.9	66.4	47.9	70.6	54.5
ABMDRNet$_{21}$ [11]	RGB-T	**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
FEANet$_{21}$ [29]	RGB-T	93.3	**87.8**	82.7	71.1	76.7	61.1	65.5	**46.5**	26.6	22.1	**70.8**	6.6	**66.6**	**55.3**	**77.3**	48.9	**73.2**	**55.3**
EGFNet$_{22}$ [28]	RGB-T	**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
LASNet$_{22}$ [40]	RGB-T	**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
GCNet$_{22}$ [41]	RGB-T	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**
CCFFNet50$_{22}$ [42]	RGB-T	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	-	**57.4**
SGFNet (Ours)	RGB-T	**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**

TIR image for network input. We use the color jitter and the random horizontal flipping, scaling and cropping for data augmentation. The batch size is set to 4. The Ranger optimizer is used to optimise the training process. The initial learning rate and weight decay are set to $5e^{-5}$ and $5e^{-4}$, respectively. The parameters of the backbones are initialized by the pre-trained ResNet50 model [26]. We train the proposed SGFNet in an end-to-end manner for 250 epochs.

3) Evaluation Metrics: We use two evaluation metrics of mean accuracy (mAcc) and mean intersection over union (mIoU) to evaluate our SGFNet and other methods.

B. Comparison With State-of-the-Art Methods

We compare our SGFNet with state-of-the-art RGB/RGB-D/RGB-T semantic segmentation methods. For a fair comparison, we retrain the first two categories of methods with their default parameter settings on the same training set as ours, and then test the retrained models on the same testing set as ours to obtain their segmentation maps. We obtain the segmentation maps of the last category of methods through the public benchmarks or codes.

1) Comparison on the MFNet Dataset: We compare our SGFNet with three categories of 18 state-of-the-art methods on the MFNet dataset. The first category is the RGB semantic segmentation method, including DANet [20] and HRNet [33]. We also modify these RGB semantic segmentation methods by adding an extra input channel for thermal images to convert them into RGB-T version. The second category is the RGB-D semantic segmentation method, including FuseNet [34], D-CNN [35], ACNet [36] and SA-Gate [37]. We replace the input depth images by thermal images. The last category is the RGB-T semantic segmentation method, including MFNet [10], RTFNet [13], PSTNet [38], MMNet [39], MLFNet [12], FuseSeg [14], ABMDRNet [11], FEANet [29], EGFNet [28], LASNet [40], GCNet [41] and CCFFNet50 [42] (the backbone of CCFFNet50 is ResNet50, and the author only provide mIoU of this version).

We list the quantitative comparison results of our SGFNet and all compared methods in Tab. I. Overall, our method shows excellent performance in two comprehensive evaluation metrics compared to the 18 semantic segmentation methods on MFNet dataset. Among the eight categories, our method achieves the best performance in three categories,*i.e.*, cars,

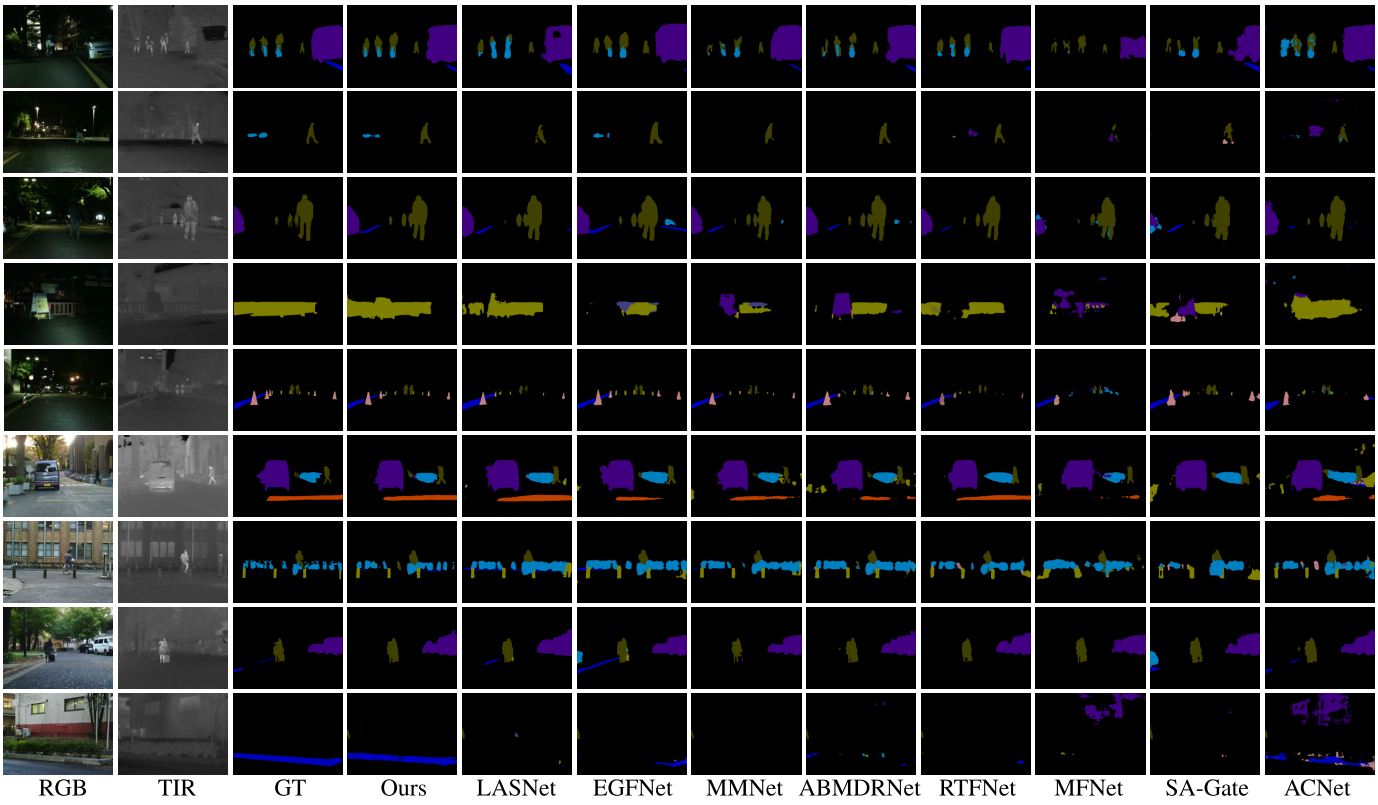

Fig. 8. Visual comparisons of our method and eight classic state-of-the-art methods in nighttime (the first five scenes) and daytime (the last four scenes) on the test set of MFNet.

pedestrians and color cone. Cars and pedestrians are the most concerned objects in driving scenes, accounting for 52.7% and 15.2% of non-background labels, respectively in the MFNet dataset. Our method performs best in these two categories, which proves that our method is more suitable for driving scenes.

In Tab. II, we show the performance of some comparison methods in the daytime and nighttime scenes of the MFNet dataset. Our method has advantages in both scenes. The excellent performance in daytime scene proves that our method can make full use of multiple modalities and fuse the information of vivid RGB images. On the other hand, the excellent performance in nighttime scene shows that the semantic guidance of our method is effective and our method can accurately segment objects under poor lighting conditions with its help.

In Fig. 8, we visualize the segmentation maps of our method and the other eight state-of-the-art methods on nine scenes, including five nighttime scenes (top) and four daytime scenes (bottom). Obviously, our segmentation maps are closer to GTs and are the best of all methods. In the five nighttime scenes, we use the stable semantic information provided by the TIR image to accurately segment the objects in the RGB image. At the same time, with the help of multi-modal fusion, our model can also handle small objects well. In the four daytime cases, effective multi-modal fusion makes unobvious objects in the TIR image to be segmented successfully, and the semantic information makes the segmentation maps free of noise. It is worth noting that most comparison methods mix other types of wrong prediction results around the edges of objects, or make wrong predictions in the background region. In contrast, our method generates stable prediction

TABLE II

QUANTITATIVE COMPARISON (%) ON THE TEST SET OF MFNET DATASET IN DAYTIME AND NIGHTTIME. THE BEST RESULT IN EACH COLUMN IS HIGHLIGHTED IN **RED**

Methods	Type	Daytime		Nighttime	
		Acc	IoU	Acc	IoU
$DANet_{19}$ [20]	RGB	61.0	46.3	52.6	47.0
$DANet_{19}$ [20]	RGB-T	50.9	37.5	52.4	40.1
$HRNet_{19}$ [33]	RGB	64.7	46.7	54.0	47.3
$HRNet_{19}$ [33]	RGB-T	54.4	46.1	55.1	50.7
$FuseNet_{17}$ [34]	RGB-D	49.5	41.0	48.9	43.9
$D\text{-}CNN_{18}$ [35]	RGB-D	50.6	42.4	50.7	43.2
$ACNet_{19}$ [36]	RGB-D	60.7	41.6	63.9	47.4
$SA\text{-}Gate_{20}$ [37]	RGB-D	49.3	37.9	56.9	45.6
$MFNet_{17}$ [10]	RGB-T	42.6	36.1	48.9	43.9
$RTFNet50_{19}$ [13]	RGB-T	57.3	44.4	59.4	52.0
$RTFNet152_{19}$ [13]	RGB-T	60.0	45.8	60.7	54.8
$MLFNet_{21}$ [12]	RGB-T	-	45.6	-	54.9
$FuseSeg_{21}$ [14]	RGB-T	62.1	47.8	67.3	54.6
$ABMDRNet_{21}$ [11]	RGB-T	58.4	46.7	68.3	55.5
SGFNet (Ours)	RGB-T	**70.7**	**49.2**	**72.2**	**58.4**

TABLE III
QUANTITATIVE COMPARISONS (%) ON THE TEST SET OF PST900 DATASET. '-' MEANS THAT THE AUTHORS DO NOT PROVIDE THE CORRESPONDING RESULTS. THE TOP TWO RESULTS IN EACH COLUMN ARE HIGHLIGHTED IN **RED** AND **BLUE**

Methods	Type	Background		Hand-Drill		Backpack		Fire-Extinguisher		Survivor		mAcc	mIoU
		Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
$ACNet_{19}$ [36]	RGB-D	99.8	99.3	53.6	51.5	85.6	83.2	84.9	56.0	69.1	65.2	78.7	71.8
$SA\text{-}Gate_{20}$ [37]	RGB-D	99.7	99.3	89.9	**81.0**	89.0	79.8	80.7	73.0	64.2	62.2	84.7	79.1
$MFNet_{17}$ [10]	RGB-T	-	98.6	-	41.1	-	64.3	-	60.3	-	20.7	-	57.0
$PSTNet_{20}$ [38]	RGB-T	-	98.9	-	53.6	-	69.2	-	70.1	-	50.0	-	68.4
$EGFNet_{22}$ [28]	RGB-T	99.5	99.3	**98.0**	64.7	**94.2**	83.1	**95.2**	71.3	**83.3**	74.3	**94.0**	78.5
$MTANet_{22}$ [47]	RGB-T	-	99.3	-	62.1	-	**87.5**	-	65.0	-	**79.1**	-	78.6
$LASNet_{22}$ [40]	RGB-T	99.8	**99.5**	92.4	77.8	**90.8**	**86.5**	**91.8**	**82.8**	**83.4**	75.5	**91.6**	**84.4**
$CCFFNet50_{22}$ [42]	RGB-T	**99.9**	99.4	89.7	**82.8**	77.5	75.8	87.6	**79.9**	79.7	72.7	86.9	82.1
SGFNet (Ours)	RGB-T	**99.8**	**99.4**	**94.0**	76.7	90.4	85.4	89.4	75.6	82.7	**76.7**	91.2	**82.8**

TABLE IV
COMPARISON OF PARAMETERS AND FLOPS

Methods	Input size	Parameter (M)	FLOPs (GMac)	mIoU (%)
$RTFNet50_{19}$ [13]	640 × 480	185.24	245.71	51.7
$RTFNet152_{19}$ [13]	640 × 480	254.51	337.04	53.2
$PSTNet_{20}$ [38]	640 × 480	20.38	129.37	48.4
$MMNet_{21}$ [39]	640 × 480	23.90	95.00	52.8
$ABMDRNet_{21}$ [11]	640 × 480	64.60	194.33	54.8
$CCFFNet50_{22}$ [42]	640 × 480	71.07	345.40	57.4
$EGFNet_{22}$ [28]	640 × 480	28.18	109.15	54.8
$LASNet_{22}$ [40]	640 × 480	93.58	233.81	54.9
Ours	640 × 480	125.25	143.73	57.6

results through the semantic information to avoid the above problems.

2) Comparison on the PST900 Dataset: We also compare our SGFNet with 8 state-of-the-art multi-modal methods on the PST900 dataset, such as RGB-D semantic segmentation methods (ACNet [36] and SA-Gate [37]) and RGB-T semantic segmentation methods (MFNet [10], PSTNet [38] EGFNet [28], MTANet [47], LASNet [40] and CCFFNet50 [42]).

The quantitative performance of our method and all compared methods show that our method achieves competitive performance on the PST900 dataset, as shown in Tab. III. This indicates that our method can also effectively handle complex scenes and low-contrast images in underground environments. Specifically, our method ranks second on mIoU. This demonstrates the effectiveness and robustness of our SGFNet on different datasets of different scenes.

3) Comparison of Model Complexity: Parameters and FLOPs are important metrics to evaluate the model complexity and efficiency. In Tab. IV, we report the parameters and FLOPs of our model and some compared methods. Compared with RTFNet [13], which is also based on the asymmetric structure, our model significantly reduces the amount of parameters and calculations. Compared with symmetric-structure LASNet [40], our model also achieves the leading mIoU while reducing a large amount of computation. This demonstrates that our SGFNet achieves a suitable trade-off between the model complexity and the accuracy.

C. Ablation Studies

We present ablation studies on MFNet dataset to demonstrate the effectiveness of components of our SGFNet. We conduct ablation studies in the following three parts: 1) the effectiveness of MCDU and CSEU in SGFNet, 2) the importance of each part in MCDU and CSEU, and 3) the validity of the edge information in our decoder.

1) The Effectiveness of MCDU and CSEU in SGFNet: We provide three variants to evaluate the individual effectiveness of MCDU and CSEU: 1) Baseline + FE, which replaces all MCDU and CSEU with the element-wise summation, 2) FE + MCDU, and 3) FE + MCDU + CSEU. The experimental results are presented in Tab. V. Baseline achieves 54.6% on mIoU, which is 3.0% lower than our full SGFNet, demonstrating that our two modules can indeed improve the effectiveness of multi-modal fusion. With the help of MCDU

TABLE V

QUANTITATIVE RESULTS (%) OF ASSESSING THE EFFECTIVENESS OF MCDU AND CSEU IN SGFNET. THE BEST ONE IS **RED**

No.	Baseline	FE	MCDU	CSEU	MFNet [10] mIoU
1	✓	✓			54.6
2	✓	✓	✓		56.3
3	✓	✓		✓	56.3
4	✓				53.1
5	✓		✓		55.9
6	✓		✓	✓	56.8
7	✓	✓	✓	✓	**57.6**

Fig. 9. Some segmentation results of the variant without CSEU (No.2) and the original model with CSEU (No.7). Please note the light noise in the red boxes and the noise suppression of CSEU.

or CSEU, FE + MCDU and FE + CSEU improve the performance compared to Baseline, respectively.

Moreover, we also provide another three variants without feature enhancement on thermal features to further evaluate the individual effectiveness of MCDU and CSEU: 4) Baseline, which removes FE, MCDU and CSEU, 5) Baseline + MCDU, and 6) Baseline + MCDU + CSEU. According to the experimental results, the removal of FE reduces the overall performance of our model, but the addition of MCDU and CSEU significantly improves the performance of our model. This further shows the effectiveness of MCDU and CSEU.

Furthermore, We have added some segmentation results of the variant without CSEU (No.2) and the original model with CSEU (No.7) in Fig. 9 to demonstrate the effectiveness of CSEU in locating correct objects and eliminating noise disturbance in low-light conditions by introducing semantic information of TIR images. We can observe that our CSEU can effectively suppress the light noise and enhance the target region in the thermal images, which leads to more accurate and consistent segmentation.

2) The Importance of Each Part in MCDU and CSEU: First, to measure the effectiveness of special design for different levels in MCDU, we provide two variants in the upper part of Tab. VI: MCDU-L is used in all levels (*i.e.*, *all low*) and MCDU-H is used in all levels (*i.e.*, *all high*). The experimental results show that it is insufficient to adopt the same coordination part for features of different levels and our special design is necessary. Then, to verify the effect of each part in MCDU, we provide two variants: removing the coordination parts of MCDU (*i.e.*, *w/o CP*) and removing the distillation parts of MCDU (*i.e.*, *w/o DP*). The experimental results show that both parts in MCDU are indispensable. Furthermore, to verify that the structure in MCDU-H can extract important complementary information of different concentrated regions in two modalities, we provide a variant that removes the complementary attention, *i.e.*, *w/o CA*. The result reported in Tab. VI indicates that the structure in MCDU-H can capture important complementary information in two modalities, which can improve the feature extraction and fusion.

TABLE VI

QUANTITATIVE RESULTS (%) OF ASSESSING THE IMPORTANCE OF EACH PART IN MCDU AND CSEU. THE BEST ONE IS **RED**

Aspects	Models	MFNet [10] mIoU
	SGFNet (Ours)	**57.6**
MCDU	*all low*	56.8
	all high	56.8
	w/o CP	55.6
	w/o DP	55.5
	w/o CA	55.1
CSEU	*w/o SG*	56.5
	w/o UF	54.6
	w/o MM	56.3
	w/o PE	56.0

To assess the effectiveness of each component of CSEU, we provide four variants of CSEU in the bottom part of Tab. VI: 1) removing the semantic information, *i.e.*, *w/o SG*, 2) removing the upper-level features, *i.e.*, *w/o UF*, 3) removing the mutual modulation of semantic information and upper-level features, *i.e.*, *w/o MM*, and 4) removing the positional encoding, *i.e.*, *w/o PE*. Based on the upper-level features, the introduction of semantic information makes mIoU of *w/o SG* increase by 1.1%. On the other hand, semantic information needs to be supplemented by multi-modal features of details. The information interaction of semantic information and the upper-level features can highlight the key regions of the upper-level features and the key of semantic information, making mIoU increase by 1.3%. The positional encoding can make better use of the prior information of the vertical location distribution of objects, making mIoU increase by 1.6%.

3) The Validity of the Edge Information in Our Decoder and the Effectiveness of Our Encoder: In order to verify the validity of edge information in our edge-aware Lawin ASPP and the performance of our encoder, we design several variants. The first variant, *i.e.*, *w/o edge*, which removes the edge supervision to restore our decoder to the original one, is used to validate the edge information in our decoder. The experimental results reported in Tab. VII show that the segmentation accuracy can be effectively improved by introducing edge information into our decoder.

To further validate the effectiveness of our encoder, we conduct experiments with other decoders, such as PSPNet [16],

TABLE VII
QUANTITATIVE RESULTS (%) OF ASSESSING THE VARIANTS OF OUR DECODER. THE BEST ONE IS **RED**

Models	MFNet [10]
	mIoU
SGFNet (Ours)	**57.6**
w/o edge	55.1
PSP	54.4
Deeplabv3	56.1
Uper	56.1

DeepLabv3 [48] and UPerNet [49]. These variants achieve competitive results, which demonstrates the robustness and effectiveness of our encoder.

V. CONCLUSION

In this paper, we propose a novel SGFNet to enhance the multi-modal feature representations in the asymmetric encoder. In the TIR branch of our SGFNet, we take full advantage of the stability of TIR images, and extract semantic information for enhancing multi-modal fusion. In the RGB branch of our SGFNet, we adopt two specific manners to fuse multi-modal information of different levels, and then distillate the fused features in MCDU. After the multi-modal fusion in MCDU, we enhance the fused features with the help of cross-level information and semantic information in CSEU. In addition, we improve the Lawin ASPP decoder by introducing edge information extracted from low-level features, forming our edge-aware Lawin ASPP decoder. Extensive experiments on the MFNet dataset prove that our SGFNet achieves competitive performance compared to 18 state-of-the-art methods.

REFERENCES

[1] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[2] V. Badrinarayanan, A. Kendall, and R. Cipolla, "SegNet: A deep convolutional encoder–decoder architecture for image segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 12, pp. 2481–2495, Dec. 2017.

[3] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 1520–1528.

[4] Z. Song, L. Zhao, and J. Zhou, "Learning hybrid semantic affinity for point cloud segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 7, pp. 4599–4612, Jul. 2022.

[5] Y. Tian and S. Zhu, "Partial domain adaptation on semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3798–3809, Jun. 2022.

[6] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, Oct. 2015, pp. 234–241.

[7] Y. Zhu, C. Li, J. Tang, B. Luo, and L. Wang, "RGBT tracking by trident fusion network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 2, pp. 579–592, Feb. 2022.

[8] J. Tang, D. Fan, X. Wang, Z. Tu, and C. Li, "RGBT salient object detection: Benchmark and a novel cooperative ranking approach," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 12, pp. 4421–4433, Dec. 2020.

[9] W. Zhou, Q. Guo, J. Lei, L. Yu, and J. Hwang, "ECFFNet: Effective and consistent feature fusion network for RGB-T salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1224–1235, Mar. 2022.

[10] 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 *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2017, pp. 5108–5115.

[11] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 2633–2642.

[12] Z. Guo, X. Li, Q. Xu, and Z. Sun, "Robust semantic segmentation based on RGB-thermal in variable lighting scenes," *Measurement*, vol. 186, Dec. 2021, Art. no. 110176.

[13] Y. Sun, W. Zuo, and M. Liu, "RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes," *IEEE Robot. Autom. Lett.*, vol. 4, no. 3, pp. 2576–2583, Jul. 2019.

[14] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "FuseSeg: Semantic segmentation of urban scenes based on RGB and thermal data fusion," *IEEE Trans. Autom. Sci. Eng.*, vol. 18, no. 3, pp. 1000–1011, Jul. 2021.

[15] L. 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 Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 4, pp. 834–848, Apr. 2018.

[16] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, "Pyramid scene parsing network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6230–6239.

[17] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, "BiSeNet: Bilateral segmentation network for real-time semantic segmentation," in *Proc. ECCV*, 2018, pp. 325–341.

[18] A. Vaswani et al., "Attention is all you need," in *Proc. NIPS*, vol. 30, 2017, pp. 1–11.

[19] A. Dosovitskiy et al., "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. ICLR*, 2021, pp. 1–22.

[20] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang, "A single stream network for robust and real-time RGB-D salient object detection," in *Proc. ECCV*, Aug. 2020, pp. 646–662.

[21] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7242–7252.

[22] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. NIPS*, Dec. 2021, pp. 12077–12090.

[23] Z. Liu et al., "Swin Transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.

[24] W. Wang et al., "PVT v2: Improved baselines with pyramid vision transformer," *Comput. Vis. Media*, vol. 8, no. 3, pp. 415–424, Sep. 2022.

[25] H. Yan, C. Zhang, and M. Wu, "Lawin transformer: Improving semantic segmentation transformer with multi-scale representations via large window attention," 2022, *arXiv:2201.01615*.

[26] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[27] W. Zhou, J. Liu, J. Lei, L. Yu, and J. Hwang, "GMNet: Graded-feature multilabel-learning network for RGB-thermal urban scene semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 7790–7802, 2021.

[28] W. Zhou, S. Dong, C. Xu, and Q. Yaguan, "Edge-aware guidance fusion network for RGB-thermal scene parsing," in *Proc. AAAI*, Feb. 2022, pp. 3571–3579.

[29] F. Deng et al., "FEANet: Feature-enhanced attention network for RGB-thermal real-time semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2021, pp. 4467–4473.

[30] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2018, pp. 3–19.

[31] S. Choi, J. T. Kim, and J. Choo, "Cars can't fly up in the sky: Improving urban-scene segmentation via height-driven attention networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9370–9380.

[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.

[33] K. Sun, B. Xiao, D. Liu, and J. Wang, "Deep high-resolution representation learning for human pose estimation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5686–5696.
[34] C. Hazirbas, L. Ma, C. Domokos, and D. Cremers, "FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Nov. 2016, pp. 213–228.
[35] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. ECCV*, Sep. 2018, pp. 144–161.
[36] X. Hu, K. Yang, L. Fei, and K. Wang, "ACNET: Attention based network to exploit complementary features for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1440–1444.
[37] X. Chen et al., "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for RGB-D semantic segmentation," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Aug. 2020, pp. 561–577.
[38] 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 *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2020, pp. 9441–9447.
[39] X. Lan, X. Gu, and X. Gu, "MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation," *Int. J. Speech Technol.*, vol. 52, no. 5, pp. 5817–5829, Mar. 2022.
[40] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "RGB-T semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2023.
[41] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "GCNet: Grid-like context-aware network for RGB-thermal semantic segmentation," *Neurocomputing*, vol. 506, pp. 60–67, Sep. 2022.
[42] W. Wu, T. Chu, and Q. Liu, "Complementarity-aware cross-modal feature fusion network for RGB-T semantic segmentation," *Pattern Recognit.*, vol. 131, Nov. 2022, Art. no. 108881.
[43] L. Chen, Y. Yang, J. Wang, W. Xu, and A. L. Yuille, "Attention to scale: Scale-aware semantic image segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3640–3649.
[44] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," in *Proc. Int. Conf. Learn. Representat.*, Dec. 2015, pp. 1–14.
[45] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun, "Large kernel matters—Improve semantic segmentation by global convolutional network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1743–1751.
[46] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Adv. Neural Inf. Process. Syst.*, Dec. 2019, pp. 8024–8035.
[47] W. Zhou, S. Dong, J. Lei, and L. Yu, "MTANet: Multitask-aware network with hierarchical multimodal fusion for RGB-T urban scene understanding," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 48–58, Jan. 2023.
[48] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[49] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2018, pp. 418–434.

Yike Wang received the B.E. degree from Shanghai University, Shanghai, China, in 2021, where he is currently pursuing the M.E. degree with the School of Communication and Information Engineering. His research interests include image processing and semantic segmentation.

Gongyang Li received the Ph.D. degree from Shanghai University, Shanghai, China, in 2022. From July 2021 to June 2022, he was a Visiting Ph.D. Student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He is currently a Post-Doctoral Fellow with the School of Communication and Information Engineering, Shanghai University. His research interests include image/video object segmentation, semantic segmentation, and saliency detection.

Zhi Liu (Senior Member, IEEE) received the B.E. and M.E. degrees from Tianjin University, Tianjin, China, in 1999 and 2002, respectively, and the Ph.D. degree from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, Shanghai, China, in 2005. He is currently a Professor with the School of Communication and Information Engineering, Shanghai University, Shanghai. From August 2012 to August 2014, he was a Visiting Researcher with the SIROCCO Team, IRISA/INRIA-Rennes, France, with the support by EU FP7 Marie Curie Actions. He has published more than 200 refereed technical papers in international journals and conferences. His research interests include image/video processing, machine learning, computer vision, and multimedia communication. He is an Area Editor of *Signal Processing: Image Communication*. He served as a Guest Editor for the Special Issue on "Recent Advances in Saliency Models, Applications and Evaluations" in *Signal Processing: Image Communication*.