Context-Aware Interaction Network for RGB-T Semantic Segmentation

Ying Lv , Zhi Liu , *Senior Member, IEEE*, and Gongyang Li

Abstract—RGB-T semantic segmentation is a key technique for autonomous driving scenes understanding. For the existing RGB-T semantic segmentation methods, however, the effective exploration of the complementary relationship between different modalities is not implemented in the information interaction between multiple levels. To address such an issue, the Context-Aware Interaction Network (CAINet) is proposed for RGB-T semantic segmentation, which constructs interaction space to exploit auxiliary tasks and global context for explicitly guided learning. Specifically, we propose a Context-Aware Complementary Reasoning (CACR) module aimed at establishing the complementary relationship between multimodal features with the long-term context in both spatial and channel dimensions. Further, considering the importance of global contextual and detailed information, we propose the Global Context Modeling (GCM) module and Detail Aggregation (DA) module, and we introduce specific auxiliary supervision to explicitly guide the context interaction and refine the segmentation map. Extensive experiments on two benchmark datasets of MFNet and PST900 demonstrate that the proposed CAINet achieves state-of-the-art performance.

Index Terms—Context-aware complementation, detail aggregation, global context, RGB-T semantic segmentation.

I. INTRODUCTION

S EMANTIC segmentation [1], [2], [3], [4], [5], [6], [7], [8], [9] is a fundamental computer vision task that assigns a class to each pixel in an image, and thus forms dense semantic regions and provides scene understanding for many real-world applications such as autonomous driving [10], [11], robotic manipulation [12], medical diagnosis [13], and virtual reality [14]. Recently, semantic segmentation performance has been improved

Ying Lv is 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, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China (e-mail: yinglv@shu.edu.cn).

Zhi Liu and Gongyang Li 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, School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China, and also with Wenzhou Institute of Shanghai University, Wenzhou 325000, China (e-mail: liuzhisjtu@163.com; ligongyang@shu.edu.cn).

The code is available at https://github.com/YingLv1106/CAINet.
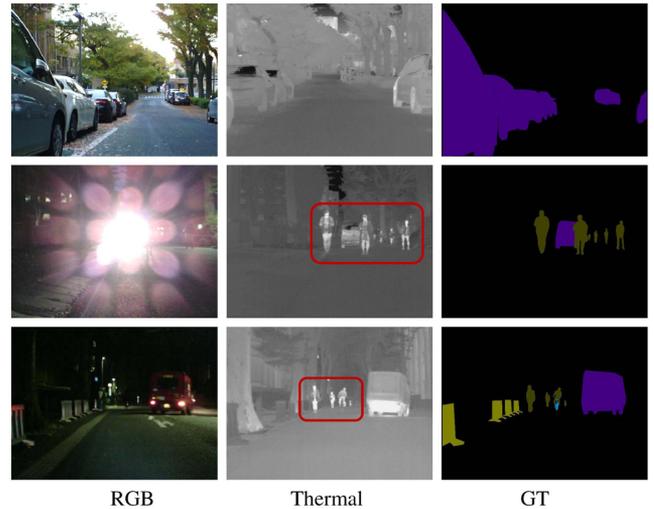
Digital Object Identifier 10.1109/TMM.2023.3349072

Fig. 1. Samples of RGB-T semantic segmentation from MFNet [18] dataset, which contains daytime, bright light and nighttime images from top to bottom.
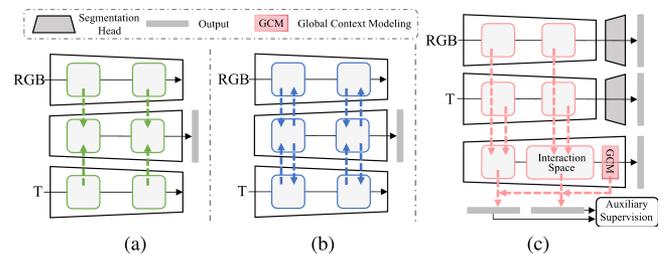

Fig. 2. Three typical categories of architectures in RGB-T semantic segmentation, including a) direct fusion [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], b) feedback fusion [31], [32], [33], and c) our proposed context-aware interaction fusion network.

significantly [4], [15], [16], [17]. However, unimodal methods perform poorly in realistic driving scenarios, e.g., cluttered backgrounds, distant small targets, and unfavorable lighting conditions (dazzle lamps or even darkness), as shown in Fig. 1. It is difficult to identify pedestrians and vehicles from RGB images, but they can be clearly identified from thermal images. RGB-based semantic segmentation is inadequate, and to overcome this problem, thermal images are introduced. However, this brings new challenges, such as the effective fusion of cross-modal information for RGB-T semantic segmentation. The mainstream RGB-T semantic segmentation methods can be divided into two categories (as shown in Fig. 2), i.e., direct fusion [18], [19], [20],

[21], [22], [23], [24], [25], [26], [27], [28], [29], [30] and feedback fusion [31], [32], [33]. The first category (Fig. 2(a)) consists of two backbones that are used to extract features separately. Then, specific modules are designed for direct fusion independent of the backbones. The second category (Fig. 2(b)) differs from the first one in that fusion features independent of the backbone are fed back to the respective backbone, completing the interaction between the encoder and the fusion layer at multiple levels, respectively. Both of the above fusion paradigms have greatly explored the fusion of features based on RGB and thermal modalities.

However, both categories have their shortcomings. Most direct fusion models use the same module for feature fusion at different levels, which neglects the characteristics of different levels. Meanwhile, level-by-level fusion is not conducive to information interaction between levels. Although the feedback fusion models complete the feature interaction, the feedback of interactive information to the backbone leads to information bottleneck [34], which makes it impossible to recur shallow features to deep levels of the model to form high-level semantic information in the feature extraction stages. In addition, the effective exploration of complementary relationship between different modalities is not implemented for information interaction between multiple levels. Moreover, the supervision is only set at the end of the above two categories of models, and there is no explicit guiding supervision in the multi-level feature interaction thus making the representation of multi-level fused features ambiguous and uncertain.

To mitigate the above shortcomings, we propose a Context-Aware Interaction Network (CAINet) for RGB-T semantic segmentation, aiming to achieve an effective exploration of the complementary relationship between different modalities for the information interaction between multiple levels. We design a new fusion paradigm to implement CAINet by combining the advantages of both direct fusion and feedback fusion paradigms, as shown in Fig. 2(c). Since multiple low-level features can provide boundary detail features for refining the segmentation results, we also fully explore them in our CAINet. In addition, we adopt different auxiliary supervision in CAINet to improve the feature representation at multiple levels.

Specifically, we propose the Context-Aware Complementary Reasoning (CACR) module to exploit complementary relationship between multiple modalities and their long-range dependencies along spatial and channel dimensions. Then the proposed Global Context Modeling (GCM) module provides global context to guide multi-level feature interactions. The contextual complementary information and global guidance are obtained from the CACR and GCM modules at three high-level stages. In order to explicitly convey the complementary information of multiple modalities, we assign specific auxiliary supervision to multi-level features as shown in Fig. 2(c), which has the advantage of explicitly guiding each level of feature representation. Hence long-range dependencies and multimodal complementary information are fully utilized with the cooperation of CACR and GCM. Moreover, the Detail Aggregation (DA) module is proposed to explore boundary information and binary mask, which can solve some problems, i.e., binary mask supervision to alleviate the interference of cluttered background and boundary supervision to further improve performance by performing boundary detail refinement. Thanks to all this, our model achieves state-of-the-art performance on MFNet and PST900 datasets.

Our main contributions are summarized as follows:
- We fully explore the complementary relationship between high-level multimodal features and propose a novel context-aware interaction fusion paradigm to implement CAINet, which achieves state-of-the-art performance on MFNet and PST900 datasets.
- We propose the CACR module for establishing complementary relationship between multimodal features with long-term dependence in spatial and channel dimensions, GCM module to explore global context clues to provide global guidance for feature interactions, and DA module to aggregate detailed features to further promote segmentation performance.
- We introduce auxiliary supervision and residual learning into CAINet. The auxiliary supervision ensures explicit guidance step by step and the residual learning retains the high-level global context information.

The rest of this paper is organized as follows. Section II explains the related work on RGB and RGB-T semantic segmentation. The proposed CAINet is elaborated in Section III. Section IV presents the experimental results and analysis of ablation experiments. Finally, Section V concludes this paper.

II. RELATED WORK

A. RGB Semantic Segmentation

Since the introduction of the fully convolutional network (FCN) for semantic segmentation, various CNN-based semantic segmentation models have been proposed and achieved satisfactory performance. However, there are limitations in the receptive field. Many works attempted to alleviate this weakness. For instance, Chen et al. [35] proposed a deep convolutional network, which utilizes multiple sampling rates of dilated convolution in atrous spatial pyramid pooling. Similarly, Zhao et al. [36] proposed a pyramid scene parsing network that aggregates contextual information at different scales.

Sequentially, many semantic segmentation methods integrate attention mechanisms into the convolutional structure. Wang et al. [37] proposed the non-local network, which leverages weighted sum to aggregate the features of all points to obtain global features. Hu et al. [38] introduced a novel technique called SENet, which incorporates global context information to adjust channel dependencies and rescale channel weights. Woo et al. [39] proposed the convolution-based attention module, which splits the attention process into two independent parts, namely, the channel attention and spatial attention modules. Cao et al. [40] investigated a new global context modeling framework known as GCNet, which can build long-range dependencies while being computationally efficient. Fu et al. [41] presented a dual attention network to capture feature dependencies in both the spatial and channel dimensions.

Following the introduction of visual transformers, subsequent advancements in semantic segmentation have led to the emergence of transformer-based techniques. Zheng et al. [42] adopted an encoder with transformer structure to tackle semantic segmentation as a sequence-to-sequence prediction task. Xie et al. [4] introduced a hierarchical structure transformer encoder combining both local and global attention mechanisms to produce powerful representations. Cheng et al. [16] proposed a mask classification paradigm that leverages both instance segmentation and semantic segmentation tasks. Furthermore, Zhang et al. [43] transferred a similar mapping strategy between a set of learnable class tokens and spatial feature mappings to segmentation masks. Lastly, Kim et al. [44] presented a generalized sample-less learner for arbitrarily dense prediction tasks called Visual Token Matching (VTM). Overall, these unimodal segmentation methods have demonstrated their efficacy in generating powerful representations for semantic segmentation. However, in situations with reduced visibility or obstacles such as fog, rain, snow, or dust caused by adverse weather conditions, as well as in low-light or no-light conditions like nighttime or inside tunnels, visible light imaging systems may struggle to effectively detect objects against the background. This is where the supplementation of infrared technology becomes necessary.

B. RGB-T Semantic Segmentation

Although the RGB semantic segmentation methods show superior performance, they still perform poorly when encountering scenes with extreme lighting conditions, as depicted in Fig. 1. So many researchers additionally introduce thermal images to improve the segmentation performance. Ha et al. [18] proposed MFNet as a pioneer of RGB-T semantic segmentation, which uses a two-stream structure for feature extraction and then fuses these features using concatenation. RTFNet [19] and FuseSeg [22] adopted a comparable architecture as MFNet, albeit fusing features through an element-wise summation. Shivakumar et al. [20] proposed a two-stage structure and fused the output of the first stage with thermal and color images. However, these methods almost used simple fusion strategies, i.e., the element-wise summation and concatenation are used to capture cross-modal features, which may lead to information redundancy by ignoring the differences between cross-modal information.

Many studies have attempted to design specialized feature fusion operations. Guo et al. [21] proposed a multi-stage skip connection between the encoder and decoder. Zhang et al. [23] introduced a strategy of bridging-then-fusing to reduce multimodal differences and achieved multi-scale contextual feature fusion. Lan et al. [45] introduced a two-stage multimodal multi-stage network (MMNet) to extract features from different modalities separately and the second stage performed feature fusion and refinement. Wang et al. [46] proposed a double fusion embedded learning approach to force each task to adapt to the feature representation of the other tasks. Deng et al. [25] introduced channel and spatial attention modules in the encoder to enhance the feature fusion. Zhou et al. [24], [26], [27], [30], [47], [48] fused cross-modal features in a specially designed module

and binary boundary-assisted supervision for feature refinement. Yi et al. [49] proposed an attentional feature fusion network of channel coordinates to obtain channel and coordinate correlations between multimodal features. Xu et al. [50] introduced graph convolution to fuse multimodal high-level features from a global and semantic perspective and also uses multitasking supervision. Zhao et al. [51] performed unimodal feature extraction separately, spatial attention in shallow layer, shared convolutional blocks in the high layer, and channel attention to fuse them. Frigo et al. [52] explored match correlation and confidence weights to explicitly estimate feature matches for decoders by computing feature correlations. And confidence weights are used for multimodal feature fusion. Wang et al. [53], [54] proposed modality recovery methods that effectively predict the distribution-consistent space of missing modalities, achieving consistency in multimodal distributions and semantic disambiguation.

Fu et al. [55] presented the attention mechanism to fuse multilevel features and added the extracted global information to the decoding stage. Feng et al. [56] introduced a teacher-student distillation model that used a boundary-assisted teacher model to learn cross-modal features and a student model to learn only thermal image features. Cai et al. [57] proposed a two-branch network enhanced by boundary refinement and guided foreground background features and fused with high-level features to preserve semantic information. Zhou et al. [29] proposed an embedded control gate fusion paradigm, and developed an attention residual learning strategy to ensure effective instruction of high-level features. Li et al. [28] treated multi-level features differently, i.e., the high-, middle- and low-level semantic features were used to localize, activate target regions, and refine boundary, respectively.

Above mentioned feature fusion method may ignore the interaction of hierarchical features and cannot sufficiently explore multimodal complementary information. In recent years, some interaction fusion methods have been proposed. Zhou et al. [58] proposed a transformer-based framework to handle the correlation and complementation between multimodal, and later return the fused features to their respective backbone networks. This method differs from the feature fusion method mentioned above as shown in Fig. 2(a), which feeds the fused features to their respective backbones. Wu et al. [32] proposed a complementary-aware cross-modal feature fusion network (CCFFNet) to select complementary information from multimodal features and fused them through a channel weighting mechanism. However, feature fusion in spatial details is not sufficient. Liu et al. [33] proposed a vision-transformer-based cross-modal feature rectify fusion network that exploited global max pooling in terms of channel and spatial interaction, but the pooling layer may lead to loss of spatially detailed features. The complementary relationship between the two modalities may not be well-exploited by these models.

Aiming to effectively explore multimodal complementary information, we combine the benefits of two fusion paradigms, i.e., direct fusion and feedback fusion, we propose a novel fusion paradigm that builds an interaction space for reasoning about complementary relationship between different
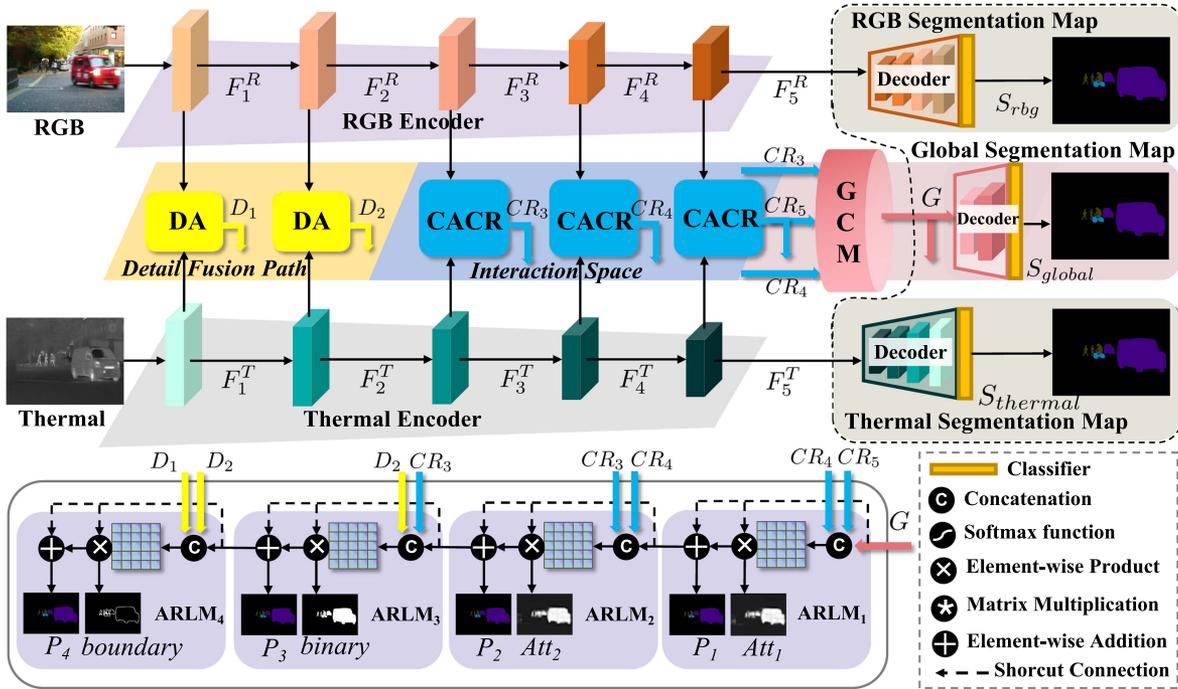
Fig. 3. The overview of proposed CAINet. Specifically, CAINet consists of six components including RGB and thermal encoders, interaction space reasoning, global context modeling, three-decoder supervision, detailed feature fusion, and residual learning of multiple auxiliary tasks supervision. The residual learning [29] (ARLM) module is to assist context-aware complementary reasoning (CACR) and global context modeling (GCM) to implement multimodal feature interaction; the detail aggregation (DA) module refines the final segmentation map. During inference, we can remove the other supervised branches and retain the final predictive semantic segmentation map P_4, suggesting that the performance enhancement comes with no added inference cost.

modalities at multi-level stages. And we propose CACR, GCM, and DA for exploring complementary relationship, harvesting global contextual cue, and refining segmentation results, respectively. Furthermore, auxiliary supervisions assist high-level contextual information to guide feature interaction in an explicit manner. Based on the above, we propose the Context-Aware Interaction Network, i.e., CAINet, for RGB-T semantic segmentation.

III. PROPOSED METHOD

The main pipeline of our Context-Aware Interaction Network (CAINet) can be divided into five stages: RGB and thermal stream encoder-decoder, context-aware complementary reasoning (Section III-B), global context modeling (Section III-C), detail aggregation (Section III-D) and residual learning with auxiliary supervision (Section III-E). The overall framework is shown in Fig. 3.

A. Architecture Overview

We employ two MobileNet-V2 [59] networks pretrained on ImageNet [60] as the backbone. As suggested in ECGFNet [29], we remove the last down-sampling operations, hence enlarging the resolution of the feature map, so the output stride becomes 16 instead of 32. Let the multistage RGB and thermal features obtained from the i-th ($i \in \{1, 2, 3, 4, 5\}$) block of MobileNet-V2 be F_i^R and F_i^T, respectively. They are followed by the decoder

layers and independently supervised to obtain the coarse segmentation maps denoted as S_{rgb} and $S_{thermal}$, respectively. To establish a multimodal complementary relationship for RGB and thermal features with long-term dependencies in both spatial and channel dimensions, we use the CACR module at three levels (i.e., F_{3-5}^R and F_{3-5}^T). Additionally, we introduce the GCM module to explore global context clues and provide global guidance for feature interactions. Following the GCM module is a decoder to generate a global coarse segmentation map S_{global}. The DA module aggregates detailed features and further improves segmentation performance. We employ the Attention Residual Learning Module (ARLM) [29] to establish a upsample stream with auxiliary supervision and target supervision running in parallel. This upsample stream consists of four stages, where the auxiliary supervision is carried out sequentially for attention, binary, and boundary supervision. In contrast, the target supervision exclusively focuses on semantic segmentation supervision. Ultimately, through residual learning, we obtain progressively refined semantic segmentation maps. The predicted segmentation map P_4 from the fourth stage serves as the final segmentation result.

B. Context-Aware Complementary Reasoning

There exists a complementary relationship between multimodal semantic regions, and how to effectively tap this complementary relationship is crucial for multimodal information
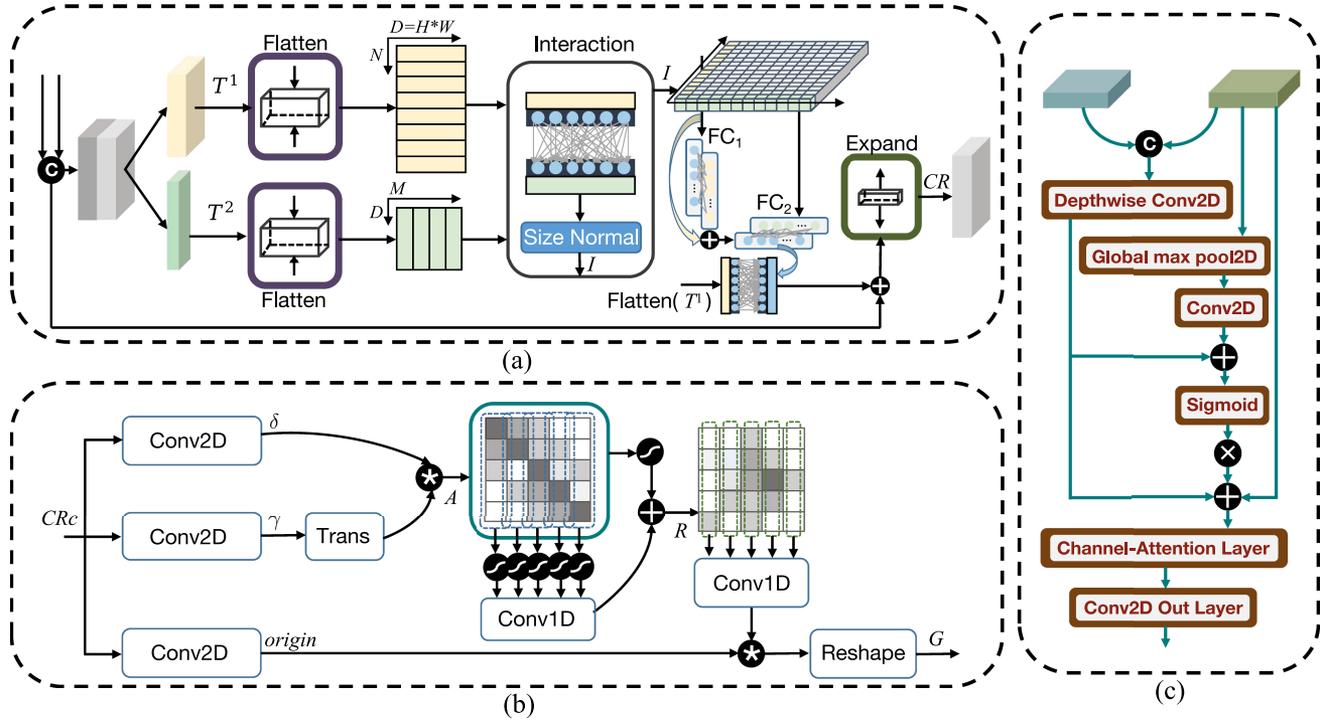
Fig. 4. Illustration of a) the context-aware complementary reasoning (CACR) module, b) the global context modeling (GCM) module, where δ, γ, and $origin$ represent feature maps from different convolution layers and c) the detail aggregation (DA) module.

fusion [50]. We propose the CACR module to construct an interaction space, which reasons about the complementary relationship on a long range of the interaction space, including the spatial and channel dimensions. Specifically, the CACR module views the inputs as a set of features (i.e., pixels) with the corresponding feature dimensions (e.g., spatial structure, color, surface temperature, and local and global texture), which enable the exchange of long-range contexts, enhancing multimodal features at the local level and global level simultaneously. We apply the CACR module in three high-level stages of our model, i.e., $\{F_{3-5}^R, F_{3-5}^T\}$, and their outputs are CR_3, CR_4, and CR_5, respectively, as shown in Fig. 3.

The detailed structure of the CACR module is shown in Fig. 4(a). The inputs to the CACR module are the high-level features $F_3^{R/T}$, $F_4^{R/T}$, and $F_5^{R/T}$. Given an input feature $\mathbf{X} \in \mathbb{R}^{C \times H \times W}$ converted to a collection of features $\mathbf{P} \in \mathbb{R}^{C \times D}$, where $D = H \times W$ is the number of features, and each channel contain information. The image of $C \times H \times W$ can only provide two-dimensional information of objects and cannot be used for geometry analysis directly. In contrast, the features $C \times D$ represent the spatial and attribute information of the feature [61], so it can be easily used for geometry analysis and processing. For context interaction space transformation, we let the model learn a transfer function $\boldsymbol{f}(\cdot)$:

$$T(X) = \boldsymbol{f}(X; \omega) = X\omega, \tag{1}$$

where ω is the learnable convolution layer, two different dimensional features are then obtained, which are $T^1 \in \mathbb{R}^{N \times H \times W}$ and

$T^2 \in \mathbb{R}^{M \times H \times W}$. We set $N = C$ and $M = C/2$. We aim to build a similarity correlation between the regions of two features and further complement multimodal features in the interaction space based on regional correlations:

$$I = \frac{[\text{Flatten}(T^2)]^T * \text{Flatten}(T^1)}{N \times M}, \tag{2}$$

where $\text{Flatten}(\cdot)$ denotes that map these features to feature set space and $*$ denotes the matrix multiplication. We get the similarity correlation features $I \in \mathbb{R}^{M \times N}$ of the interaction space, which contains the multimodal feature description of each pixel and can capture the relationship between any regions, and then reason about the complete multimodal feature map. In practice, 'Size Normal' as shown in Fig. 4(a) is to simply divide the feature map by its size $N \times M$ to control the numerical amplitude of the feature map.

In particular, the similarity correlation I is then adaptively assigned to each feature of the region in the interaction space. Fig. 4(a) shows the implementation of this process and the inference along two different dimensions through two fully connected layers. By doing so, the features can communicate with each other based on complementary relationship between multimodal features, which allows the exchange of long-range contexts. Therefore, the complementary reasoning CR is given as follows:

$$CR' = \sigma((\text{FC}_2(\sigma(\text{FC}_1(I) + I)))), \tag{3}$$

$$CR = \text{Expand}(X + \text{Flatten}(T^1)^T * CR'), \tag{4}$$

where σ denotes the ReLU activation function, FC_1 represents the fully connected layer along the N dimension, and FC_2 denotes the M dimension. Expand$(\cdot)$ represents to extend these features to restore the H and W dimensions. Here, we follow similar correlations to reason about complementarity and apply two fully connected layers to assign feature dimensions (from the interaction space dimension to the original dimension). After this, we obtain three high-level complementary features CR_3, CR_4 and CR_5. Specifically, CACR module is different from convolutional blocks and vision transformers [62], where the former ones perform a sliding window on the image to extract features from local regions, and the latter ones treat an image as a sequence of patches and extract features via self-attention mechanism [63] in a global range, as well as different from the GCN [64] module with edge and adjacency matrix. The CACR module establishes the similarity correlation between the multimodal feature regions and is a novel approach that constructs an interaction space to reason about the complementary relationship, allowing it to capture both local and global contextual information in a flexible and adaptive way.

C. Global Context Modeling

Many efforts try to overcome the limitation of local convolution operators by introducing a global context, such as global pooling of SENet [38], ASPP [65], and self-attention mechanism [63]. Although we also incorporate a global paradigm, in the proposed approach, we go one step further and perform higher-level global reasoning on the complementary relationship between multi-level features in the GCM module as shown in Fig. 4(b). The GCM module yields the global context modeling feature G, which will be used as a global guide for joint auxiliary supervision with the ARLM.

Auxiliary and target tasks require global contextual information to provide object location cues, as stated in GLCNet [66], global information is essential to enhance semantic prediction. The key is to improve the semantic representation of each channel using global context modeling. To account for this, As shown in Fig. 4(b), we first aggregate three high-level complementary features by channel concatenation to produce semantic guidance CR_c. Then, we use RM to represent the relationship modeling, which can be obtained as follows:

$$A = [\text{Conv2D}(CRc)]^T * \text{Conv2D}(CRc), \qquad (5)$$

$$RM = \text{Conv1D}\left(\beta(A)\right) + \frac{\exp(A)}{\sum_j^c \sum_i^c \exp(A)}, \qquad (6)$$

where $Conv2D$ denotes 2D convolution layer with a 1×1 kernel and stride 1, β represents the Softmax function along with channel dimension, and $Conv1D$ denotes the 1D convolution with a 1×1 kernel and stride 1. Further, global context features G can be obtained as follows:

$$G = \text{Reshape}\left[\text{Conv2D}_3(\mathbf{CRc})\right]^T * \text{Conv1D}_2(\mathbf{RM}), \quad (7)$$

where $Reshape$ represents feature dimension reshaping, $Conv2D_3$ denotes a 2D convolution with a 1×1 kernel and stride 1, $Conv1D_2$ denotes the 1D convolution with a 1×1 kernel and stride 1. And G as the global guidance, combines

with residual learning of ARLM to associate auxiliary and target supervision. In addition, G passes through the decoder to get a global coarse segmentation map after the interaction, which may lose detail information, so next we will construct the DA module for feature map refinement.

D. Detail Aggregation Module

The features play different roles at different levels in the convolutional neural network (CNN). The low-level ones (e.g., $F_1^{R/T}$ and $F_2^{R/T}$) contain boundary information and abundant texture. Inspired by this observation, we propose a simple yet effective DA fusion scheme to well utilize multimodal features in uniting the ARLM module for auxiliary boundary detail supervision, further boosting the performance of our CAINet.

As shown in Fig. 4(c), we combine the RGB and thermal features F_i^R and F_i^T ($i = 1, 2$) by concatenation along the channel direction in the AD module. In order to maintain the performance while greatly reducing computational costs, we use depthwise separable convolution to get the combined features F_i^c. For the thermal features, we utilize a global max pooling layer and convolution with 7×7 kernel size to improve the extraction of spatial information, and here the extracted thermal features are denoted as T_i^s. Finally, the detail aggregation features D_i are obtained by a channel attention layer calculated as follows:

$$D_i = \text{CA}\left(F_i^c \left(\sigma\left(F_i^c + T_i^s\right)\right) + F_i^T\right), \qquad (8)$$

where CA denotes the channel attention layer. After we obtain the detail aggregation features D_i, and through subsequent ARLM steps, we obtained the boundary and binary map, refining the final semantic segmentation result.

E. Residual Learning With Auxiliary and Target Tasks

In order to explicitly convey the correlation between multiple modalities and target regions, we adopt the idea of residual learning to achieve cross-modal feature interaction in a collaborative manner with the CACR and GCM modules. Residual learning refers to the use of global context features G to gradually guide the interactions of multi-level complementary features, and is performed step-by-step with multi-level interaction features and global information. As shown at the bottom of Fig. 3, the ARLM has three inputs, e.g., the global context features G, and the complementary features CR_5 and CR_4 for $ARLM_1$, after which the predicted segmentation map P_1 and the predicted attention map Att_1 are obtained. After the same step-by-step calculation as above, predicted results can be obtained from residual learning, which are P_2, Att_2, P_3, the predicted binary map $binary$, P_4, and the predicted boundary map $boundary$.

In particular, in auxiliary supervision, the purpose of using attention maps for supervision is to explicitly reason about complementary relationship between multimodal features. The ground truth of attention map Q is obtained by using corrosion expansion and Gaussian fuzzy operation on the ground truth of

binary map.[1] And we take the sum of mean square error and the modified linear correlation coefficient metric [66] as the loss function:

$$L_{Att_i} = \frac{1}{N} \sum_{i=1}^{n} \|Att_i - Q\|^2 - \frac{\text{cov}(Att_i, Q)}{\sqrt{\text{var}(Att_i)}\sqrt{\text{var}(Q)}}, \quad (9)$$

where cov and var represent covariance and variance, respectively.

For P_1 and P_2, we utilize Lovász-Softmax [67] to respectively calculate their losses as L_{seg1} and L_{seg2}, which has the perceptual quality and proportional invariance, and gives small objects proper correlation and false negative counts compared to the cross-entropy loss. The two semantic segmentation losses are L_{seg1} and L_{seg2} as follows:

$$L_{seg1} = \frac{1}{|C|} \sum_{c \in C} \overline{\Delta_{J_c}}(m(c)),$$

$$m(c) = \begin{cases} 1 - P_1(c) & \text{if } c = Y(c) \\ P_1(c) & \text{otherwise}, \end{cases} \quad (10)$$

where Y denotes the semantic segmentation label, $|C|$ represents the class number, ΔJc defines the Lovász extension of the Jaccard index. L_{seg2} is similar to L_{seg1}. For P_3 and P_4, we use the weighted cross-entropy function to calculate their losses, denoted as L_{seg3} and L_{seg4} with the following function:

$$L_{seg3} = -\frac{1}{N} \sum_{i=1}^{N} \omega \left(Y^i \log(P_3^i) \right), \quad (11)$$

L_{seg4} is similar to L_{seg3}. For these four semantic segmentation losses, we add them together to obtain the target loss L_{target}. For $binary$ and $boundary$,[2] we use the weighted binary cross-entropy loss function to calculate the binary loss and boundary loss, donated as L_{binary} and $L_{boundary}$, respectively.

$$L_{\text{binary}} = -\frac{1}{N} \sum_{i=1}^{N} \omega(Y_{bi}^i \log(binary^i) + (1 - Y_{bi}^i) \log(1 - binary^i)), \quad (12)$$

where N denotes the total number of input pixels, ω denotes the class weights computed following [68], and Y_{bi} denotes the ground truth of binary map. $L_{boundary}$ is similar to L_{binary}. Furthermore, for the coarse segmentation maps S_{rgb}, $S_{thermal}$, and S_{global} produced by the three decoders of the backbone network, we adopt the ordinary cross-entropy loss, and define the sum of these three losses as $L_{decoder}$. To sum up, the final loss for auxiliary and target supervision is calculated as follows:

$$L_{total} = L_{target} + L_{Att1} + L_{Att2} + L_{binary}$$
$$+ L_{boundary} + L_{decoder}. \quad (13)$$

[1]The ground truth of binary map is obtained by setting the non-zero elements of the label values in the semantic segmentation map to 1.

[2]The ground truth of boundary map is obtained by performing Canny edge detection on the ground truth of binary map.

IV. EXPERIMENTS

A. Experimental Protocol

1) Datasets: We train and evaluate the proposed CAINet on two RGB-T semantic segmentation datasets: MFNet [18] and PST900 [20]. The MFNet dataset contains 1569 RGB-T image pairs (820 taken during the day and 749 at night) and has 9 categories, including background, car, person, bike, curve, car stop, guardrail, color cone, and bump. The resolution size of the image pairs is 480×640 pixels. We divide the dataset into three parts: training set, validation set, and test set, using the same splitting scheme as proposed in MFNet. The PST900 dataset contains 894 aligned RGB and thermal image pairs taken by the Stereolabs ZED Mini stereo camera and the FLIR Boson 320 camera, respectively. This dataset contains five categories: background, fire extinguisher, backpack, hand drill, and survivor. The resolution of the RGB and thermal images is 1280×720. This dataset is divided into two parts, the training set and the test set, with 597 pairs constituting the training and 288 pairs constituting the test set.

2) Evaluation Metrics: We used two metrics to quantitatively evaluate the performance of semantic segmentation, which are the mean accuracy (mAcc) and the mean intersection over union (mIoU). The mean accuracy is calculated as the ratio of the number of correct pixels for each class to the number of all predicted pixels for that class, and then the average value is calculated by summation. The mean intersection over union is the average result of summing the intersection of the true values of the predictions for each class and the ratio of the sums. These measures are defined as follows:

$$mAcc = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij}}, \quad (14)$$

$$mIoU = \frac{1}{k+1} \sum_{i=0}^{k} \frac{p_{ii}}{\sum_{j=0}^{k} p_{ij} + \sum_{j=0}^{k} p_{ji} - p_{ii}}, \quad (15)$$

where k is the number of classes (including unlabeled) and class $id = 0$ indicates "unlabeled", P_{ij} is the number of pixels belonging to class i and is predicted as class j.

3) Implementation Details: We train our model by loading a pre-trained MobileNet-V2 [59] model on the ImageNet [60] dataset. The RGB branch and the thermal image branch are first trained separately. Then their model parameters are loaded and the GCM decoder branch is trained. Finally, the trained basic model parameters are loaded and the whole model is trained together with the residual learning branch. The batch size of each iteration is 8 images. The learning rate is initialized as 5×10^{-4}. The network parameters are learned by backpropagating the model using the Adam optimization algorithm. We use an early termination algorithm to prevent network overfitting. The experiments are performed on the publicly available Pytorch 1.12.0 [69] framework, using a workstation with a TITAN RTX GPU (24 GB RAM).

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

1) Evaluation on MFNet Dataset: We compare our CAINet with nineteen state-of-the-art methods, most of which are

TABLE I
QUANTITATIVE COMPARISONS (%) ON THE TEST SET OF MFNET DATASET

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		
MFNet[18]	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]	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]	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	74.7	55.7	63.1	53.2
PSTNet[20]	RGB-T	-	76.8	-	52.6	-	55.3	-	29.6	-	25.1	-	15.1	-	39.4	-	45.0	-	48.4
MLFNet[21]	RGB-T	-	82.3	-	68.1	-	67.3	-	27.3	-	30.4	-	15.7	-	55.6	-	40.1	-	53.8
FuseSeg[22]	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[23]	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
MMNet[45]	RGB-T	-	83.9	-	69.3	-	59.0	-	43.2	-	24.7	-	4.6	-	42.2	-	50.7	62.7	52.8
EGFNet[24]	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
FEANet[25]	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
MTANet[26]	RGB-T	95.8	88.1	90.9	71.5	80.3	60.7	75.3	40.9	62.8	38.9	38.7	13.7	63.8	45.9	70.8	47.2	75.2	56.1
MFFENet161[30]	RGB-T	93.1	88.2	83.2	74.1	77.1	62.9	67.2	46.2	52.3	37.1	65.0	7.6	58.5	52.4	73.4	47.4	74.3	57.1
GMNet[27]	RGB-T	94.1	86.5	83.0	73.1	76.9	61.7	59.7	44.0	55.0	42.3	71.2	14.5	54.7	48.7	73.1	47.4	74.1	57.3
CCFFNet101[32]	RGB-T	94.5	89.6	83.6	74.2	73.2	63.1	67.2	50.5	38.7	31.9	30.6	4.8	55.2	49.7	72.9	56.3	68.3	57.6
CCAFFMNet[49]	RGB-T	95.2	89.1	85.9	72.5	82.3	67.5	71.8	46.3	32.5	25.2	56.8	17.3	58.3	50.6	76.6	58.3	72.9	58.2
DSGBINet[50]	RGB-T	95.2	87.4	89.2	69.5	85.2	64.7	66.0	46.3	56.7	43.4	7.8	3.3	82.0	61.7	72.8	48.9	72.6	58.1
CMX-SegF-B2[33]	RGB-T	-	89.4	-	74.8	-	64.7	-	47.3	-	30.1	-	8.1	-	52.4	-	59.4	-	58.2
FDCNet[51]	RGB-T	94.1	87.5	91.4	72.4	78.1	61.7	70.1	43.8	34.4	27.2	61.5	7.3	64.0	52.0	74.5	56.6	74.1	56.3
ECGFNet[29]	RGB-T	89.4	83.5	85.2	72.1	72.9	61.6	62.8	40.5	44.8	30.8	45.2	11.1	57.2	49.7	65.1	50.9	69.1	55.3
LASNet[28]	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
CAINet (Ours)	RGB-T	93.0	88.5	74.6	66.3	85.2	68.7	65.9	55.4	34.7	31.5	65.6	9.0	55.6	48.9	85.0	60.7	73.2	58.6

The value 0.0 represents that there are no true positives. '-' means that the authors do not provide the corresponding results. The top two results in each column are highlighted in red and blue.

TABLE II
COMPARISON OF THE NUMBER OF PARAMETERS AND EXECUTION EFFICIENCY OF EACH METHOD

Methods	Input size	FLOPs/G↓	Params/M↓	mAcc	mIoU
RTFNet50 [19]	640×480	245.71	185.24	62.2	51.7
RTFNet152 [19]	640×480	337.04	254.51	63.1	53.2
PSTNet [20]	640×480	129.37	20.38	-	48.4
FuseSeg [22]	640×480	193.40	141.52	70.6	54.5
ABMDRNet [23]	640×480	194.33	64.60	69.5	54.8
EGFNet [24]	640×480	201.29	62.82	72.3	54.8
MTANet [26]	640×480	264.69	121.58	75.2	56.1
LASNet [28]	640×480	233.81	93.58	**75.4**	54.9
Ours	640×480	**123.62**	**12.16**	73.2	**58.6**

average performance. Specifically, our proposed CAINet achieves the highest performance with 58.6% in mIoU, outperforming the second-best methods by 0.4% (i.e., CCAFFMNet and CMX-SegF-B2 with 58.2% in mIoU). Among all categories, our method performs very well in the bike and bump, outperforming the second-best method by 2.9% in mAcc and 1.2% in mIoU in the bike, and 7.7% in mAcc and in 1.3% mIoU in the bump, respectively.

Of course, our model also has a slight limitation in that for small and obscured objects, our method will segment the contiguous morphology of the target, but compared to other methods, most do not miss detection, as shown in Fig. 5. Based on the visualizations, it can be inferred that the model proposed in this study outperforms the other methods, despite the possibility of some roadblocks being undetected. Overall, the CAINet model demonstrated favorable results in scenarios of high background lighting, as thermal imaging provides a reliable representation of heat that remains consistent across diverse lighting conditions.

2) Evaluation on the PST900 Dataset: We also compare the proposed CAINet on the PST900 dataset with ten state-of-the-art methods. They are all RGB-T semantic segmentation methods, including MFNet [18], PSTNet [20], MFFENet [30], EGFNet [24], MTANet [26], GMNet [27], CCFFNet [32], DS-GBINet [50], FDCNet [51], and LASNet [28]. The quantitative performance of our proposed method and the compared

recently proposed RGB-T semantic segmentation methods, including MFNet [18], two versions of RTFNet [19], PSTNet [20], MLFNet [21], FuseSeg [22], ABMDRNet [23], MMNet [45], EGFNet [24], FEANet [25], MTANet [26], MFFENet [30], GM-Net [27], CCFFNet [32], CCAFFMNet [49], DSGBINet [50], CMX-SegF-B2 [33], FDCNet [51], and ECGFNet [29], and LASNet [28]. The quantitative results on MFNet dataset are shown in Table I, including eight categories and overall
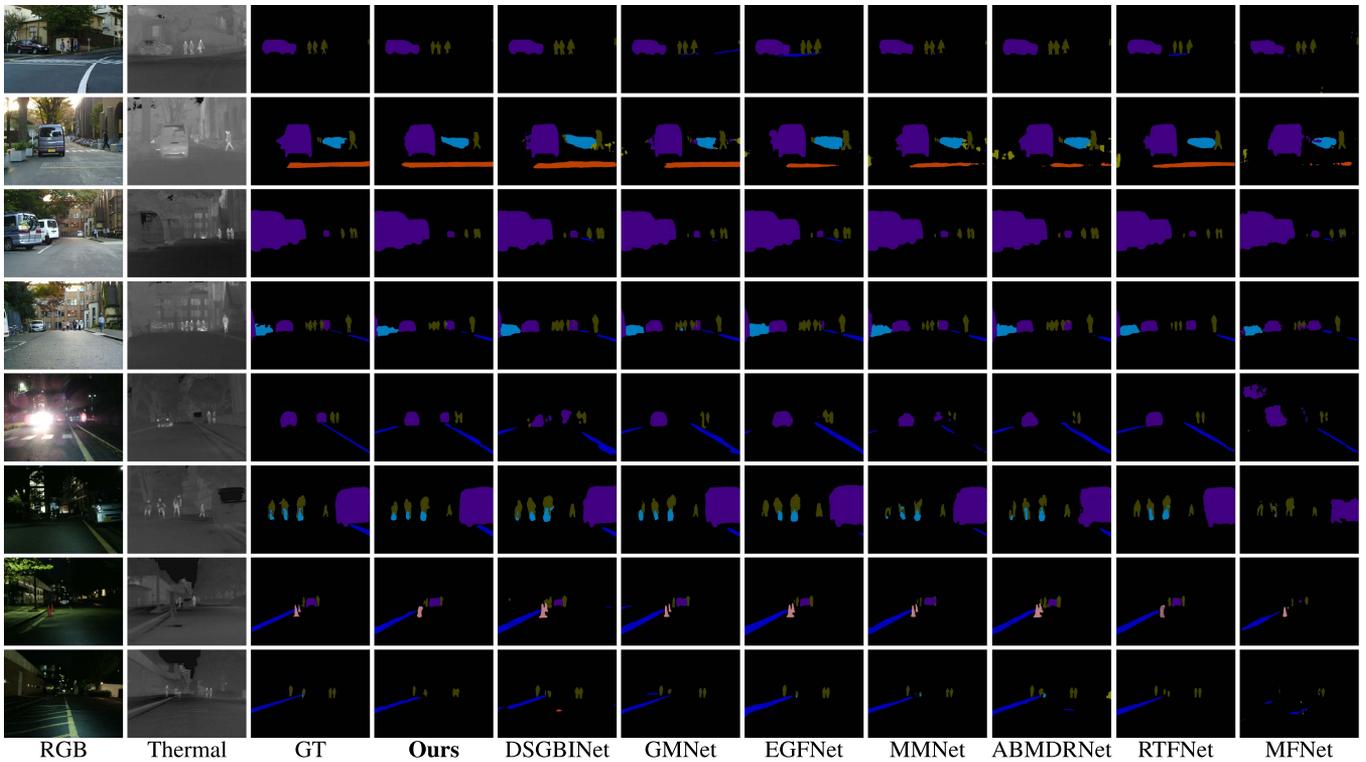
Fig. 5. Visual comparisons of the proposed method and seven state-of-the-art methods in typical daytime and nighttime images of MFNet. The proposed CAINet provides suitable results under a variety of lighting conditions.

methods on the PST900 dataset is reported in Table III. Our method demonstrates competitive performance on the PST900 dataset, securing six first-place and two second-place rankings across all 12 metrics. Of particular significance, our method outperforms the second-place LASNet by 0.34% in mIoU, ranking first in this metric. Additionally, our method achieves first place in mAcc, surpassing EGFNet by 0.25%. It is worth noting that our method exhibits significant performance improvements in the 'Backpack' and 'Survivor' categories. Our analysis underscores the effectiveness of CAINet on the PST900 dataset and its potential for generalization to different datasets. Furthermore, compared to previous methods, our approach provides accurate and reliable segmentation results, as visualized in Fig. 6.

3) Computational Complexity: We present a comparison of model parameters and floating-point operations (FLOPs) in Table II. Compared to other models, our model achieves the highest mIoU and comparable mAcc while utilizing only 12.16 M parameters and 123.62 G FLOPs, showcasing the highest computational efficiency.

C. Ablation Studies

To explore the effectiveness of our proposed modules on RGB-T semantic segmentation, we perform ablation studies on the MFNet dataset using the same hyperparameters as in Section IV-A. Specifically, we evaluate the effectiveness of the proposed four modules of ARLM, DA, CACR and GCM with

individual and joint contributions, as well as multiple auxiliary supervision and target supervision.

1) The Individual and Joint Contributions of Four Modules: We delete and replace the above mentioned modules, and provide five variants to evaluate the individual contributions of the four modules. 1) Baseline, 2) Baseline+ARLM, 3) Baseline+ARLM+DA, 4) Baseline+ARLM+CACR, and 5) Baseline+ARLM+GCM. For "Baseline", we retain three supervised decoders where RGB and thermal image features were fused at the end using the ASPP module [65]. Among them, ARLM is special in the framework, and the other modules must be combined with ARLM to function, so ARLM cannot be removed. For convenience, the deleted modules are replaced with ASPP modules. Quantitative results are shown in Table IV, where "√" indicates that the corresponding module is retained, and no mark represents that the corresponding module is removed. "Baseline" only achieves 51.9% in mIoU, which is 6.7% lower than our vanilla CAINet, indicating that these three modules can also improve the accuracy of segmentation when acting together. With the help of DA, CACR or GCM in conjunction with ARLM, the No.2, No.3, No.4 and No.5 variants improve the performance considerably compared to "Baseline".

In addition, we provide three variants to evaluate the joint contribution of the three modules, as shown in Table IV: 6) Baseline+ARLM+CACR+GCM, 7) Baseline+ARLM+DA+GCM, and 8) Baseline+ARLM+DA+CACR. With the cooperation of the modules, the performance of the above three variants is further improved compared to the single module. We find that with

TABLE III
QUANTITATIVE COMPARISONS (%) ON THE TEST SET OF PST900 DATASET

Methods	Methods	Background		Hand-Drill		Backpack		Fire-Extinguisher		Survivor		mAcc	mIoU
		Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU		
MFNet$_{17}$ [18]	RGB-T	-	98.63	-	41.13	-	64.27	-	60.35	-	20.70	-	57.02
PSTNet$_{20}$ [20]	RGB-T	-	98.85	-	53.60	-	69.20	-	70.12	-	50.03	-	68.36
EGFNet$_{22}$ [24]	RGB-T	99.48	99.26	**97.99**	64.67	**94.17**	83.05	**95.17**	71.29	83.30	74.30	**94.02**	78.51
MTANet$_{22}$ [26]	RGB-T	-	99.33	-	62.05	-	**87.50**	-	64.95	-	**79.14**	-	78.60
MFFENet$_{21}$ [30]	RGB-T	-	99.40	-	72.50	-	81.02	-	66.38	-	75.60	-	78.98
GMNet$_{21}$ [27]	RGB-T	**99.81**	99.44	90.29	**85.17**	89.01	83.82	88.28	73.79	80.86	78.36	89.61	84.12
CCFFNet50$_{22}^*$ [32]	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
DSGBINet$_{22}$ [50]	RGB-T	99.73	99.39	94.53	74.99	88.65	85.11	**94.78**	**79.31**	81.37	75.56	91.81	82.87
FDCNet$_{22}$ [51]	RGB-T	99.72	99.15	82.52	70.36	77.45	72.17	91.77	71.52	78.36	72.36	85.96	77.11
LASNet$_{23}^\#$ [28]	RGB-T	99.77	**99.46**	_91.81_	_82.80_	90.80	86.48	_92.36_	_77.75_	**83.43**	75.49	91.63	**84.40**
CAINet (Ours)	RGB-T	99.66	**99.50**	**95.87**	80.30	**96.09**	**88.02**	88.38	77.21	**91.35**	**78.69**	**94.27**	**84.74**

$*$: The performance of CCFFNet50 is excerpted from the original paper, which originally keep one decimal place.

$\#$: We correct the values of "Hand-Drill" and "Fire-Extinguish" in LASNet, as indicated by underlining.

'-' means that the authors do not provide the corresponding results. The top two results in each column are highlighted in red and blue.
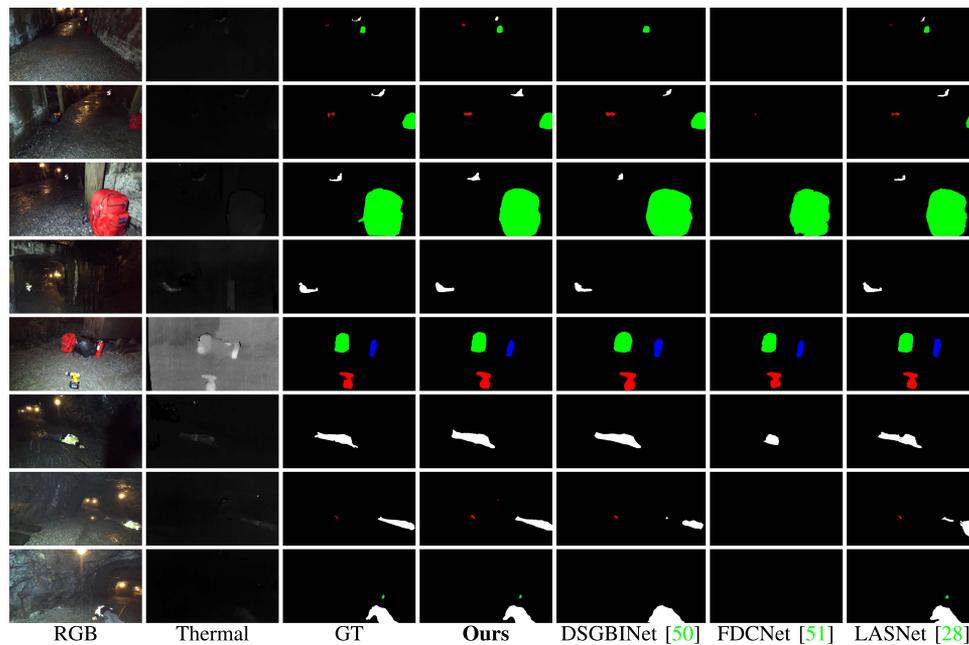

Fig. 6. Qualitative comparison for semantic segmentation of RGB-T images on PST900 [20] dataset.

the fusion of the CACR and GCM modules, the results are higher than the results of the other two joint modules, which confirms the boost of the RGB-T semantic segmentation performance by explicit auxiliary supervision with attention maps. The perfect cooperation of the three modules results in an excellent full CAINet, reaching 58.6%.

2) The Effectiveness of Auxiliary Supervision and Target Supervision: The multi-task supervision model is widely used in industry because it can improve the model performance and accelerate the convergence of model training without increasing

the reasoning time and computational complexity of the model. To evaluate the effectiveness of the ARLM module in conjunction with DA, CACR, and GCM for explicit auxiliary supervision and target supervision, we provide eight combinations of supervision in Table V, including 1) three supervisions of the decoder at the end of the model, 2) four target supervisions of residual learning branch for four semantic segmentation loss L_{seg}, 3) retaining supervision of two attention maps, 4) keeping supervision of only the binary map, 5) retaining supervision of only the boundary map, 6) joint supervision of the binary and

TABLE IV
QUANTITATIVE RESULTS (%) OF ASSESSING THE INDIVIDUAL AND JOINT CONTRIBUTIONS OF THE FOUR MODULES IN CAINET ON MFNET [18] DATASET

No.	Baseline	ARLM	DA	CACR	GCM	mAcc	mIoU
1	✓					67.2	51.9
2	✓	✓				72.0	53.0
3	✓	✓	✓			64.8	54.9
4	✓	✓		✓		70.7	55.3
5	✓	✓			✓	69.3	55.7
6	✓	✓		✓	✓	**73.8**	56.4
7	✓	✓	✓		✓	69.2	56.6
8	✓	✓	✓	✓		63.3	55.7
9	✓	✓	✓	✓	✓	73.2	**58.6**

The best one is red.

TABLE V
QUANTITATIVE RESULTS (%) OF ASSESSING THE INDIVIDUAL AND JOINT CONTRIBUTIONS OF THE FIVE KINDS OF SUPERVISION IN CAINET ON MFNET [18] DATASET

No.	$L_{decoder}$	L_{target}	L_{Att}	L_{binary}	$L_{boundary}$	mAcc	mIoU
1	✓					70.5	54.3
2	✓	✓				60.4	54.9
3	✓	✓	✓			70.6	55.7
4	✓	✓		✓		69.0	55.1
5	✓	✓			✓	70.5	54.7
6	✓	✓		✓	✓	71.2	55.9
7	✓	✓	✓		✓	**74.8**	56.7
8	✓	✓	✓	✓		70.4	56.9
9	✓	✓	✓	✓	✓	73.2	**58.6**

The best one is red.

TABLE VI
QUANTITATIVE COMPARISON RESULT(%) ON THE TEST SET OF THE NYU-DEPTH V2 DATASET

Methods	mAcc	mIoU
D-CNN [70]	56.3	43.9
3D Graph [71]	55.7	43.1
LSTM [72]	60.0	45.9
RefineNet [73]	58.9	46.5
RDFNet [74]	62.8	50.1
RTFNet [19]	64.8	49.1
SA-Gate [75]	-	52.4
ECGFNet [29]	65.2	51.5
Ours	**65.9**	**52.6**

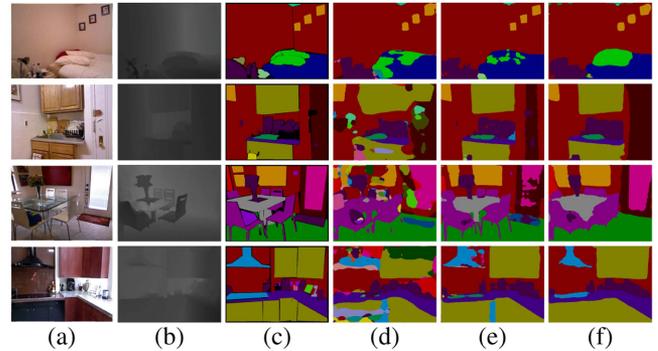The best results are highlighted in boldface.

Fig. 7. Segmentation results for RGB-D images. The proposed CAINet delivers superior segmentation across various scenes compared to the other methods on the test set of the NYU-Depth V2 dataset. (a) RGB image, (b) Depth image, (c) GT, (d) RDFNet [74], (e) SA-Gate [75], (f) CAINet.

Table VI. CAINet demonstrates superior performance, indicating its applicability to RGB-D data. Many RGB-D segmentation methods struggle with issues stemming from the limitations of depth image capture devices, such as sensitivity to lighting conditions, which often results in suboptimal segmentation. The proposed CAINet, with its CACR and GCM modules, effectively addresses these challenges. Some segmentation results are shown in Fig. 7 for a visual comparison.

V. CONCLUSION

In this paper, we review existing state-of-the-art multimodal fusion methods for RGB-T semantic segmentation, including feature fusion and feature interaction paradigms. By considering the advantages of both, we propose a new fusion paradigm, the context-aware interaction network (CAINet), in which the interaction space is constructed to exploit auxiliary tasks and global contexts for explicitly guided learning. Compared to previous work, it is more effective in tapping the complementary relationship between multimodal regions in terms of long-term dependencies in spatial and channel dimensions. Moreover, the CACR module views the inputs as a set of features and performs long-term dependencies in the interaction space. This

the boundary map, 7) joint supervision of the attention and the boundary map, and 8) joint supervision of the attention map and the binary map. We observe that discarding one or all of the supervision is detrimental to the final segmentation performance. With the effective and explicit guidance of multi-auxiliary supervision, the role of each functional module can be fully exploited. The GCM interacts as a global context-guided feature of the CACR in an explicit paradigm and the DA can enhance the representation of target edges and contour surfaces, which helps to refine the CAINet final segmentation results. The joint supervision achieves our CAINet final segmentation results in up to 58.6% outperforming other results.

D. Generalization to RGB-D Data

To assess the generalization ability of CAINet, we conduct training and testing using the NYU-Depth V2 dataset [76]. In addition, we compare our results with those of other multimodal semantic segmentation methods, taking depth images instead of thermal images as input. The comparative results are presented in

mitigates cross-modal divergence, greatly improving the semantic segmentation performance. Furthermore, with our proposed GCM and DA modules, global context information and boundary details are well-explored. Extensive experiments and analysis demonstrate our method outperforms state-of-the-art methods on MFNet and PST900 datasets. Ablation experimental results verify the efficacy of our method. While we achieve better results with CAINet, there is room for improvement in terms of the number of parameters and computational complexity. Furthermore, if we intend to deploy CAINet on mobile devices, a more lightweight model is required. Therefore, in our future work, we will focus on investigating lightweight algorithms to better adapt to embedded platforms.

REFERENCES

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

[2] 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. Representations*, 2015, pp. 1–14.

[3] 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.

[4] E. Xie et al., "SegFormer: Simple and efficient design for semantic segmentation with transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 12077–12090.

[5] G. Gao et al., "FBSNet: A fast bilateral symmetrical network for real-time semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 3273–3283, 2022.

[6] T. Chen et al., "Semantically meaningful class prototype learning for one-shot image segmentation," *IEEE Trans. Multimedia*, vol. 24, pp. 968–980, 2022.

[7] M. Zhang et al., "Weakly supervised few-shot semantic segmentation via pseudo mask enhancement and meta learning," *IEEE Trans. Multimedia*, vol. 25, pp. 7980–7991, 2023.

[8] L. Ma, H. Xie, C. Liu, and Y. Zhang, "Learning cross-channel representations for semantic segmentation," *IEEE Trans. Multimedia*, vol. 25, pp. 2774–2787, 2023.

[9] L. Zhao et al., "LIF-Seg: LiDAR and camera image fusion for 3D LiDAR semantic segmentation," *IEEE Trans. Multimedia*, early access, May 17, 2023, doi: 10.1109/TMM.2023.3277281.

[10] 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.

[11] 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.

[12] S. Ainetter and F. Fraundorfer, "End-to-end trainable deep neural network for robotic grasp detection and semantic segmentation from RGB," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2021, pp. 13452–13458.

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

[14] W. Li, W. Jia, Y. Fan, and S. Ma, "Virtual reality image dataset vCAT helps research on semantic segmentation algorithms," in *Proc. IEEE Int. Conf. Ind. Technol.*, 2022, pp. 1–6.

[15] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, "Encoder-decoder with atrous separable convolution for semantic image segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 833–851.

[16] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2021, pp. 17864–17875.

[17] G. Li, Z. Bai, Z. Liu, X. Zhang, and H. Ling, "Salient object detection in optical remote sensing images driven by transformer," *IEEE Trans. Image Process.*, vol. 32, pp. 5257–5269, 2023.

[18] 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.*, 2017, pp. 5108–5115.

[19] 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.

[20] S. S. Shivakumar et al., "PST900: RGB-thermal calibration, dataset and segmentation network," in *Proc. IEEE Int. Conf. Robot. Automat.*, 2020, pp. 9441–9447.

[21] 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.

[22] 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.

[23] Q. Zhang et al., "ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2633–2642.

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

[25] 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.*, 2021, pp. 4467–4473.

[26] 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. Veh.*, vol. 8, no. 1, pp. 48–58, Jan. 2023, doi: 10.1109/TIV.2022.3164899.

[27] W. Zhou, J. Liu, J. Lei, L. Yu, and J.-N. 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] 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.

[29] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Embedded control gate fusion and attention residual learning for RGB–thermal urban scene parsing," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 5, pp. 4794–4803, May 2023.

[30] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "MFFENet: Multi-scale feature fusion and enhancement network for RGB-thermal urban road scene parsing," *IEEE Trans. Multimedia*, vol. 24, pp. 2526–2538, 2022.

[31] H. Zhou et al., "Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7507105.

[32] W. Wu, T. Chu, and Q. Liu, "Complementarity-aware cross-modal feature fusion network for RGB-T semantic segmentation," *Pattern Recognit.*, vol. 131, 2022, Art. no. 108881.

[33] 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 Trans. Intell. Transp. Syst.*, vol. 24, pp. 14679–14694, Aug. 2023.

[34] A. M. Saxe et al., "On the information bottleneck theory of deep learning," in *Proc. Int. Conf. Learn. Representations*, 2018. [Online]. Available: https://openreview.net/forum?id=ry_WPG-A-

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

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

[37] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[38] 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.

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

[40] Y. Cao, J. Xu, S. Lin, F. Wei, and H. Hu, "GCNet: Non-local networks meet squeeze-excitation networks and beyond," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, 2019, pp. 1971–1980.

[41] 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. Eur. Conf. Comput. Vis.*, 2020, pp. 646–662.

[42] S. Zheng et al., "Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 6881–6890.

[43] B. Zhang et al., "SEGVit: Semantic segmentation with plain vision transformers," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2022, pp. 4971–4982.

[44] D. Kim, J. Kim, S. Cho, C. Luo, and S. Hong, "Universal few-shot learning of dense prediction tasks with visual token matching," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/forum?id=88nT0j5jAn

[45] X. Lan, X. Gu, and X. Gu, "MMNet: Multi-modal multi-stage network for RGB-T image semantic segmentation," *Appl. Intell.*, vol. 52, no. 5, pp. 5817–5829, Mar. 2022.

[46] Y. Wang, T. Lu, Y. Yao, Y. Zhang, and Z. Xiong, "Learning to hallucinate face in the dark," *IEEE Trans. Multimedia*, early access, Jul. 12, 2023, doi: 10.1109/TMM.2023.3294808.

[47] J. Liu, W. Zhou, Y. Cui, L. Yu, and T. Luo, "GCNet: Grid-like context-aware network for RGB-thermal semantic segmentation," *Neurocomputing*, vol. 508, pp. 60–67, Sep. 2022.

[48] T. Gong, W. Zhou, X. Qian, J. Lei, and L. Yu, "Global contextually guided lightweight network for RGB-thermal urban scene understanding," *Eng. Appl. Artif. Intell.*, vol. 117, Jan. 2023, Art. no. 105510.

[49] S. Yi, J. Li, X. Liu, and X. Yuan, "CCAFFMNet: Dual-spectral semantic segmentation network with channel-coordinate attention feature fusion module," *Neurocomputing*, vol. 482, pp. 236–251, Apr. 2022.

[50] C. Xu, Q. Li, X. Jiang, D. Yu, and Y. Zhou, "Dual-space graph-based interaction network for RGB-thermal semantic segmentation in electric power scene," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1577–1592, Apr. 2023.

[51] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for RGB-T semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2892–2905, Jun. 2023, doi: 10.1109/TCSVT.2022.3229359.

[52] O. Frigo, L. Martin-Gaffe, and C. Wacongne, "DooDLeNet: Double deeplab enhanced feature fusion for thermal-color semantic segmentation," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2022, pp. 3020–3028.

[53] Y. Wang, Z. Cui, and Y. Li, "Distribution-consistent modal recovering for incomplete multimodal learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2023, pp. 22025–22034.

[54] Y. Wang, Y. Li, and Z. Cui, "Incomplete multimodality-diffused emotion recognition," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2023. [Online]. Available: https://openreview.net/forum?id=BuGFwUS9B3

[55] Y. Fu, Q. Chen, and H. Zhao, "CGFNet: Cross-guided fusion network for RGB-thermal semantic segmentation," *Vis. Comput.*, vol. 38, no. 9/10, pp. 3243–3252, Jul. 2022.

[56] Z. Feng, Y. Guo, and Y. Sun, "CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images," *IEEE Robot. Autom. Lett.*, vol. 8, no. 4, pp. 2205–2212, Apr. 2023.

[57] Y. Cai, W. Zhou, L. Zhang, L. Yu, and T. Luo, "DHFNet: Dual-decoding hierarchical fusion network for RGB-thermal semantic segmentation," *Vis. Comput.*, to be published, doi: 10.1007/s00371-023-02773-6.

[58] H. Zhou et al., "Multispectral fusion transformer network for RGB-thermal urban scene semantic segmentation," *IEEE Geosci. Remote Sens. Lett.*, vol. 19, 2022, Art. no. 7507105.

[59] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[60] J. Deng et al., "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[61] X. Ma et al., "Image as set of points," in *Proc. Int. Conf. Learn. Representations*, 2023. [Online]. Available: https://openreview.net/pdf?id=awnvqZja69

[62] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[63] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6000–6010.

[64] Y. Chen et al., "Graph-based global reasoning networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 433–442.

[65] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "DenseASPP for semantic segmentation in street scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 3684–3692.

[66] W. Zhou, Y. Lv, J. Lei, and L. Yu, "Global and local-contrast guides content-aware fusion for RGB-D saliency prediction," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 51, no. 6, pp. 3641–3649, Jun. 2021.

[67] M. Berman, A. Rannen Triki, and M. B. Blaschko, "The Lovász-softmax loss: A tractable surrogate for the optimization of the intersection-over-union measure in neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4413–4421.

[68] A. Paszke, A. Chaurasia, S. Kim, and E. Culurciello, "ENet: A deep neural network architecture for real-time semantic segmentation," 2016, *arXiv:1606.02147*.

[69] A. Paszke et al., "PyTorch: An imperative style, high-performance deep learning library," in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, 2019, pp. 8024–8035.

[70] W. Wang and U. Neumann, "Depth-aware CNN for RGB-D segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 144–161.

[71] X. Qi, R. Liao, J. Jia, S. Fidler, and R. Urtasun, "3D graph neural networks for RGBD semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 5199–5208.

[72] Z. Li et al., "LSTM-CF: Unifying context modeling and fusion with LSTMs for RGB-D scene labeling," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 541–557.

[73] G. Lin, A. Milan, C. Shen, and I. Reid, "RefineNet: Multi-path refinement networks for high-resolution semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 5168–5177.

[74] S. Lee, S.-J. Park, and K.-S. Hong, "RDFNet: RGB-D multi-level residual feature fusion for indoor semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4990–4999.

[75] 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.*, 2020, pp. 561–577.

[76] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from RGBD images," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 746–760.

Ying Lv is currently working toward the Ph.D. degree with the School of Communication and Information Engineering, Shanghai University, Shanghai, China. His research interests include multimodal semantic segmentation and audio-visual segmentation.

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, China. From 2012 to 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 authored or coauthored 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 the Area Editor of *Signal Processing: Image Communication* and was the Guest Editor of the special issue on *Recent Advances in Saliency Models, Applications and Evaluations* in *Signal Processing: Image Communication*.

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