Region-Adaptive and Context-Complementary Cross Modulation for RGB-T Semantic Segmentation

Fengguang Peng[a], Zihan Ding[b], Ziming Chen[a], Gang Wang[c,*], Tianrui Hui[d], Si Liu[b], Hang Shi[e]

[a]*School of Computer Science and Engineering, Beihang University, China*
[b]*Institute of Artificial Intelligence,Beihang University, China*
[c]*Beijing Institute of Basic Medical Sciences, China*
[d]*Institute of Information Engineering, Chinese Academy of Sciences, China*
[e]*Beijing Research Institute of ZheJiang Laboratory, Beijing, China*

Abstract

RGB-Thermal (RGB-T) semantic segmentation is an emerging task aiming to improve the robustness of segmentation methods under extreme imaging conditions with the aid of thermal infrared modality. Foreground-background distinguishment and complementary information mining are two key challenges of this task. Recent methods use naive channel attention and cross-attention to tackle these challenges, but they still struggle with a sub-optimal solution where salient foreground features and noisy background ones might be equally modulated without distinction. The quadratic computational overhead of cross-attention also blocks its application on high-resolution features. Moreover, lacking complementary information mining in the encoding phase hinders the comprehensive scene encoding as well. To alleviate these limitations, we propose a cross modulation process with two collaborative components. The first Region-Adaptive Channel Modulation (RACM) module conducts channel attention at a fine-grained region level where foreground and background regions can be modulated differently in each channel. The second Context-Complementary Spatial Modulation (CCSM) module mines and transfers complementary information between the two modalities early in the encoding phase. Experiments show that our method achieves state-of-the-art performances on current RGB-T segmentation benchmarks.

Keywords: RGB-Thermal, Semantic Segmentation, Region-adaptive Channel Modulation, Context-complementary Spatial Modulation

1. Introduction

Semantic segmentation is a fundamental problem that aims to partition the image into regions belonging to different semantic categories. It enjoys wide applications in many visual systems, such as medical diagnosis [38], scene understanding [21], autonomous driving [54] and embodied artificial intelligence [26]. Recently, with the emergence of advanced network architectures [27, 30, 41] and context modeling techniques [4, 48, 55], impressive progress has been achieved in this field. Despite the fruitful achievements, these methods still struggle to generate accurate segmentation results in extreme imaging conditions (*e.g.*, low illumination, and adverse weather like fog and heavy rain)

*Corresponding author

Email addresses: pengfg@buaa.edu.cn (Fengguang Peng), dzh19990407@buaa.edu.cn (Zihan Ding), chenzm@buaa.edu.cn (Ziming Chen), g_wang@foxmail.com (Gang Wang), huitianrui@gmail.com (Tianrui Hui), liusi@buaa.edu.cn (Si Liu), sh@zhejianglab.com (Hang Shi)

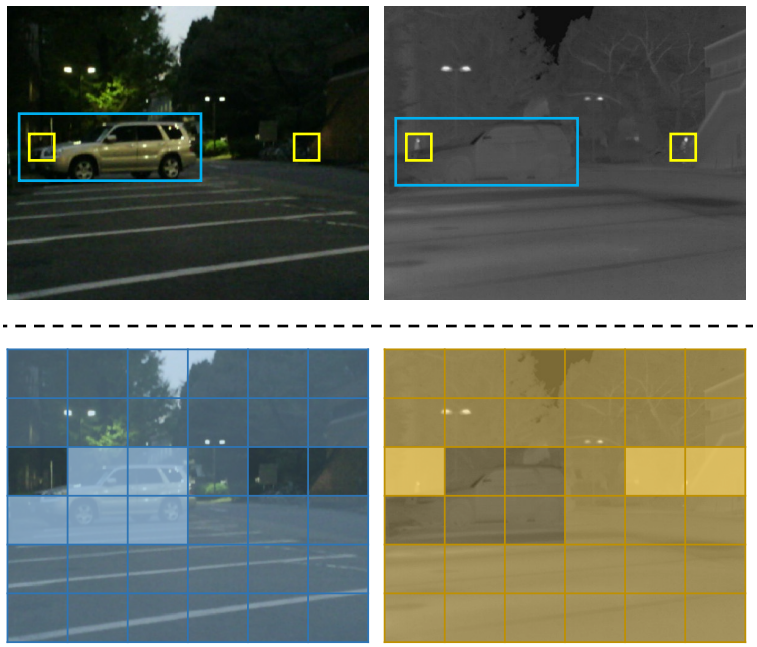

Figure 1: (Top) Due to different imaging mechanisms, RGB and TIR images are complementary to each other in low-light conditions. The white car is invisible in the TIR image because of its similar temperature to the road (highlighted by blue boxes). And the two people are not captured in the RGB image due to the poor illumination condition (highlighted by yellow boxes). (Bottom) With our proposed RACM module, regions around salient foreground objects could be enhanced, while uninformative background ones are suppressed.

based solely on the visible light data (*i.e.*, RGB images). To solve this, researchers [8, 29] propose to leverage an additional visual modality, *i.e.*, thermal infrared (TIR) images. As the thermal infrared radiation captured from the object's surface depends only on the temperature, it is insensitive to illumination variation, which can compensate for the inadequacy of regular visible light camera imaging.

The main challenges of RGB-T semantic segmentation are *foreground-background distinguishment* and *complementary information mining*. Previous methods [8, 32, 33, 56] use addition or concatenation followed by convolution layers to integrate these two modalities' information. However, these simple operators cannot effectively reduce background noise in each modality and are insufficient for cross-modal context modeling. Recently, several representative attention-based approaches are proposed to handle these issues. FEANet [6] designs a FEAM module inspired by CBAM [39], and exploits channel and spatial relations for the detail information and multi-level features in a progressive refinement way to suppress distractors. CCFFNet [40] proposes a Complementarity-Aware Encoder including a cross-modal fusion module (CMF) in each stage of ResNet backbone, a gate which is made up of a 3×3 convolutional layer followed by a Sigmoid activation function is adopted to generate the gate maps for fusing different modalities. AFNet [43] and CMX [49] utilizes the cross-attention mechanism [35] to improve the contextual correlation between two modalities. Despite the methods above achieving good results, there are still some limitations: 1) Multiplying all spatial locations with the identical weight in each channel may cause an isotropic change in both foreground and background regions, resulting in a sub-optimal foreground-background distinguishment process. 2) The quadratic computational overhead brought by the pixel-level cross-attention hinders its application on high-resolution image features. 3) Previous methods mostly fuse different modal features after they are output from the whole backbone, hardly considering complementary

information mining in the early encoding phase, which is essential for comprehensive scene context encoding of two modalities.

In light of the above discussions, we propose a cross modulation process (CM) in the encoding phase including the Region-Adaptive Channel Modulation (RACM) to distinguish foreground features from the background noises in a fine-grained region-adaptive manner, and the Context-Complementary Spatial Modulation (CCSM) to suppress spatial distractors and transfer complementary information between RGB and TIR modalities. As shown in Figure 1 (bottom), the regions containing foreground objects (*i.e.,* the white car in the RGB image and people in the TIR image) are expected to be enhanced, while background ones should be suppressed in each channel. In our RACM module, we first spatially split feature maps into different local windows, from which we extract region representations. Then, we conduct cross-attention between region features of the RGB and TIR data to explore region-wise cross-modal alignment. Finally, we generate channel weights based on the refined region features to adaptively modulate each region. Overall, the RACM module confers benefits in two aspects: 1) Unlike traditional channel attention that constructs a global channel weight across the entire feature without distinguishing between foreground and background regions, adaptive channel attention on different regions allows a finer cross-modal modulation. Namely, foreground features and background noises of the same channel can be enhanced and suppressed flexibly, avoiding a broad-brush foreground-background distinguishment process. 2) By region partitioning, the length of the region sequence is much smaller than that of the pixel sequence, it is possible to apply this approach to the high-resolution features to mine richer cross-modal correlation.

In addition, some spatial details and semantics may inherently not exist in RGB and TIR images due to the specific imaging mechanism of each modality. For example, at the top of Figure 1, the appearance information of the two people is lost in the RGB image due to the poor illumination but is captured in the TIR image. The white car is invisible in the TIR image because its temperature is similar to the road, while it can be obviously seen in the RGB image. Therefore, we argue that the complementary information mining between two modalities should occur early in the encoding phase to collaborate with the RACM module for comprehensive scene encoding. To this end, our proposed CCSM module utilizes channel grouping to generate finer and richer spatial activation weights, effectively suppressing uninformative features. Furthermore, it facilitates the exchange of complementary visual contexts between two modalities through spatial cross-activation and residual connections. Thus, spatial distractors could be replaced with salient details and semantics of the other modality. Moreover, we also devise a lightweight Gated Feature Fusion (GFF) module to effectively integrate multimodal features. Ultimately, based on the Decoder, we can output the final segmentation results.

The main contributions of our paper are three-fold:

- We propose a Region-Adaptive Channel Modulation (RACM) module to adaptively modulate each region channel-wisely for more precise foreground-background distinguishment and more efficient computation.

- We propose a Context-Complementary Spatial Modulation (CCSM) module to suppress spatial distractors and transfer complementary information between two modalities early in the encoding phase.

- To fuse multi-modal features, we propose a lightweight Gated Feature Fusion (GFF) module. Experiments show that our method achieves state-of-the-art performances on MFNet [8] and PST900 [29] datasets.

2. Related Works

2.1. Semantic Segmentation

With an increasing number of scenarios requiring accurate and efficient segmentation technology, semantic segmentation has achieved a pivotal role in dense prediction tasks and made great progress with the development of deep learning methods [36]. Long *et al.* [20] first perform image segmentation at pixel level with a fully convolutional neural network. Since then, researchers put much effort into further improving the performance in aspects including adjusting the receptive field [24, 2], fusing multi-scale features [51, 13, 43], and so on. Recently, attempts of replacing CNN backbones with ViT variants have proved to take effect. SETR [53] and Segformer [41] exploit Transformer [7] as the encoder with a simple decoder to recover pixel-level prediction. MaskFormer [5] reformulates semantic segmentation as a mask classification problem. SAM [14] has indeed gained significant attention recently as a foundation model for image segmentation. It addresses various downstream segmentation problems on new data distributions by utilizing prompt engineering techniques. SAM provides valuable information about object localization and delineation, it may not directly classify the specific objects within the segmented regions. Numerous works [19, 34, 22] have attempted to investigate the performance and versatility of SAM in different scenarios by combining it with other models. However, RGB-based semantic segmentation methods heavily rely on visible light data and may encounter challenges or algorithm failures under extreme imaging conditions, such as low light, heavy fog, or adverse weather conditions, which motivates the integration of other visual modalities such as thermal infrared data.

2.2. RGB-T Semantic Segmentation

With the popularity of thermal infrared imaging cameras, TIR images have been found as a good complement to RGB images in changing light conditions and adverse weather. To date, several studies have investigated fusing the two modalities with primitive methods. For example, MFNet [8] employs two encoders to extract features from RGB and TIR images respectively, and a single decoder to process concatenated cross-modal features.

Recently, various attention mechanisms [35, 46] and advanced methods are introduced to excavate cross-modal correlations. FEANet [6] utilizes the feature-enhanced attention module to enhance multi-layer features after summation between two modalities from spatial and channel dimensions. EGFNet [56] generates a prior edge map with both modalities, which is further embedded into their feature maps to capture detailed information. In addition, ABMDRNet [50] attempts to deploy translation networks between RGB and TIR modalities to reduce bi-directional modality differences. FDCNet [52] first employs a two-stream structure to extract unimodal low-level features and a siamese structure to extract unimodal high-level features from an RGB and Thermal image pair, several repeated Cross-modal Spatial Activation (CSA) modules and a Cross-modal Channel Activation (CCA) module are presented. LASNet [15] commits multi-level features fusion, it designs Collaborative Location Module (CLM), Complementary Activation Module (CAM), and Edge Sharpening Module (ESM) for the high-, middle-, and low-level features fusion. CMX [49] introduces a MiT [41] backbone and proposes a unified framework that exploits complementary features from the supplementary modality (X modality) for RGB-X semantic segmentation. After the calibration of bi-modal features, CMX employs a cross-attention operation between RGB and X modalities for a long-range context exchanging. In this paper, we propose a more fine-grained RACM module to modulate different regions with different channel weights for more precise foreground-background distinguishment. And we also propose a CCSM module to suppress spatial distractors and transfer complementary features between the two modalities early

in the encoder stage, the channel grouping strategy in CCSM not only reduces module complexity but also achieves a greater diversity and flexibility in channel weights modulation.

2.3. Efficient Model Learning

Efficient model learning has been an area of active research, aiming to improve the training process of deep learning models by reducing computational requirements and enhancing resource utilization, and enabling faster and more efficient model training. Dataset Distillation [17, 18] is particularly relevant in situations where the original dataset is large, high-dimensional, or computationally expensive to handle. By condensing the dataset, it becomes more manageable, reduces computational costs, and potentially alleviates issues of overfitting. DeRy [45] and KF [44] decompose pre-trained models into building blocks or factor networks, and then reassemble them to customize models based on users' specifications, giving rise to impressive results on transfer learning. Recently, vision transformer and its variants have shown great promise on various computer vision tasks while capturing both short- and long-range visual dependencies. However, they often suffer from quadratic computational overhead, especially for high-resolution vision tasks, many recent works have focused on reducing computational and memory costs while improving performance.

RegionViT [1] adopts a pyramid structure and utilizes a novel regional-to-local attention instead of global self-attention in vision transformer. Ouyang *et al.* [23] reshape the partly channels into the batch dimensions and group channel dimensions into multiple sub-features, achieving well-distributed spatial semantic features within each feature group. They propose an efficient multi-scale attention module (EMA) that focuses on retaining the information on per channel while reducing computational overhead.

In this paper, we utilize the MiT [41] pre-trained on large-scale datasets for feature extraction, which serves as a strong baseline for RGB-T semantic segmentation task with its powerful transformer-based feature representation and capability to capture short- and long-range visual dependencies. In the design of RACM and CCSM modules, unlike CBAM [39], we employ region splitting and channel grouping strategies in RACM and CCSM modules. Region splitting enables us to obtain different channel weights for different regions rather than a global weight in the entire feature, reducing computational overhead in subsequent cross-attention operations and improving cross-modal interaction. In CCSM, we employ channel grouping to obtain a greater variety of spatial activation maps, which not only significantly reduces computational requirements but also modulates more flexible spatial weights, suppressing spatial distractors and enhancing the overall efficiency and performance of our network.

3. Methodology

The overall architecture of our method is depicted in Figure 2. We feed the paired RGB image $\boldsymbol{I}_{\text{rgb}}$ and TIR image $\boldsymbol{I}_{\text{tir}}$ into two weight-unshared visual backbones to extract hierarchical feature maps, denoted as $\{\boldsymbol{F}^i_{\text{rgb}}\}^4_{i=1}$ and $\{\boldsymbol{F}^i_{\text{tir}}\}^4_{i=1}$ respectively. The CM at each encoding stage includes two parallel modules, where the RACM module aims to independently modulate each region with adaptive channel weights, while the CCSM module is designed to suppress spatial distractors using a channel grouping mechanism. Additionally, the CCSM facilitates the transfer of complementary contexts between two modalities through cross-modal spatial activation and residual connection from the other modality. Subsequently, we use a lightweight GFF module to effectively integrate two modalities' features and send them to the decoder for generating the final prediction $\boldsymbol{P}$.

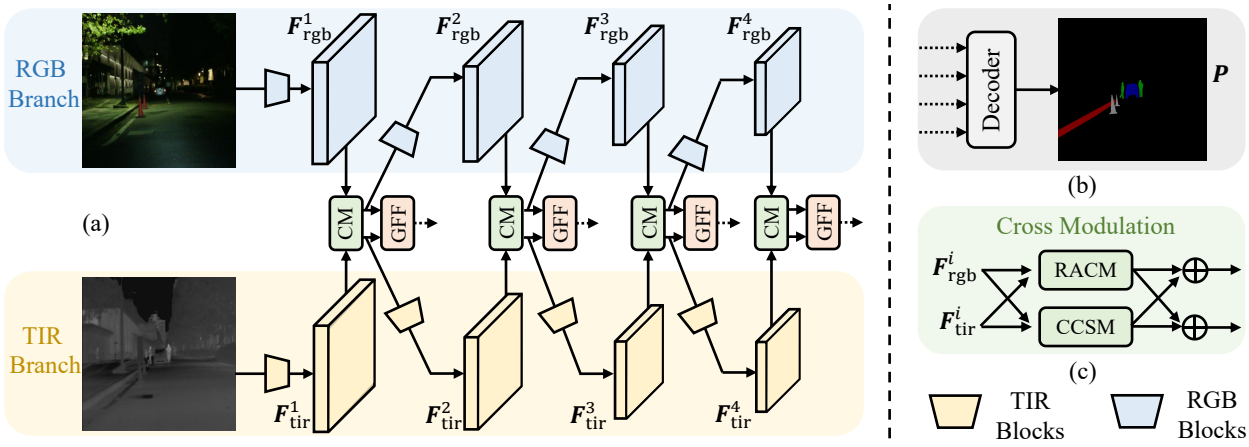

Figure 2: The overall architecture of our method. We feed RGB and TIR images into two weight-unshared backbones to extract hierarchical features $\{\boldsymbol{F}^i_{\text{rgb}}\}^4_{i=1}$ and $\{\boldsymbol{F}^i_{\text{tir}}\}^4_{i=1}$. In the i-th encoding stage, $\boldsymbol{F}^i_{\text{rgb}}$ and $\boldsymbol{F}^i_{\text{tir}}$ are sent into the Cross Modulation (CM) process including two parallel modules (*i.e.*, RACM and CCSM). Then, we use the devised GFF module to integrate two modalities' features and send them to the decoder for final predictions P.

3.1. Feature Extraction

Given the input RGB image $\boldsymbol{I}_{\text{rgb}} \in \mathbb{R}^{3\times H_0\times W_0}$ and TIR image $\boldsymbol{I}_{\text{tir}} \in \mathbb{R}^{1\times H_0\times W_0}$, we first repeat $\boldsymbol{I}_{\text{tir}}$ to $\mathbb{R}^{3\times H_0\times W_0}$ and then utilize the simple and efficient Mix Transformer (MiT) [41] as our backbone to extract multi-scale visual features $\{\boldsymbol{F}^i_{\text{rgb}}\}^4_{i=1}, \boldsymbol{F}^i_{\text{rgb}} \in \mathbb{R}^{C_i\times H_i\times W_i}$ and $\{\boldsymbol{F}^i_{\text{tir}}\}^4_{i=1}, \boldsymbol{F}^i_{\text{tir}} \in \mathbb{R}^{C_i\times H_i\times W_i}$, where C_i, $H_i = \frac{H_0}{2^{i+1}}$, and $W_i = \frac{H_0}{2^{i+1}}$ are the channel number, height, and width of features from the i-th encoding stage. H_0 and W_0 are the height and width of the original input images. To elaborate the whole modulation and fusion processes, we take the i-th encoding stage as an example and omit the superscript i for presentation clarity.

3.2. Cross Modulation (CM)

As shown in Figure 2(c), CM includes two parallel modules, the Region-Adaptive Channel Modulation module (RACM) and the Context-Complementary Spatial Modulation module (CCSM). We insert the CM module into every stage of the backbone to maximize the complementary information mining between different modalities in the encoding phase. By splitting into different regions and channel groups, we achieve a more flexible channel and spatial attention modeling, suppress background noise, and enhance the features of foreground objects.

3.2.1. RACM

RACM module aims to adaptively modulate each spatial region via independent channel weights (Figure 3). Firstly, we split both $\boldsymbol{F}_{\text{rgb}}$ and $\boldsymbol{F}_{\text{tir}}$ into $N = \frac{HW}{hw}$ non-overlapped local windows $\boldsymbol{L}_{\text{rgb}} \in \mathbb{R}^{N\times C\times h\times w}$ and $\boldsymbol{L}_{\text{tir}} \in \mathbb{R}^{N\times C\times h\times w}$ respectively, where h and w are height and width of each local window. Then we reshape $\boldsymbol{L}_{\text{rgb}}$ and $\boldsymbol{L}_{\text{tir}}$ to $\mathbb{R}^{NC\times hw}$ and extract region representations from them with linear layers:

$$\boldsymbol{R}_{\text{rgb}} = \boldsymbol{L}_{\text{rgb}}\boldsymbol{W}_1, \boldsymbol{R}_{\text{tir}} = \boldsymbol{L}_{\text{tir}}\boldsymbol{W}_2, \tag{1}$$

where $\boldsymbol{W}_1 \in \mathbb{R}^{hw\times 1}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{hw\times 1}$ are projection parameters, $\boldsymbol{R}_{\text{rgb}} \in \mathbb{R}^{N\times C}$ and $\boldsymbol{R}_{\text{tir}} \in \mathbb{R}^{N\times C}$ are region features after reshaping.

To achieve global cross-modal alignment, a bidirectional cross-attention operation is adopted. We first project $\boldsymbol{R}_{\text{rgb}}$ and $\boldsymbol{R}_{\text{tir}}$ into the same embedding subspace and multiply them to obtain the affinity matrix $\boldsymbol{A} \in \mathbb{R}^{N\times N}$:

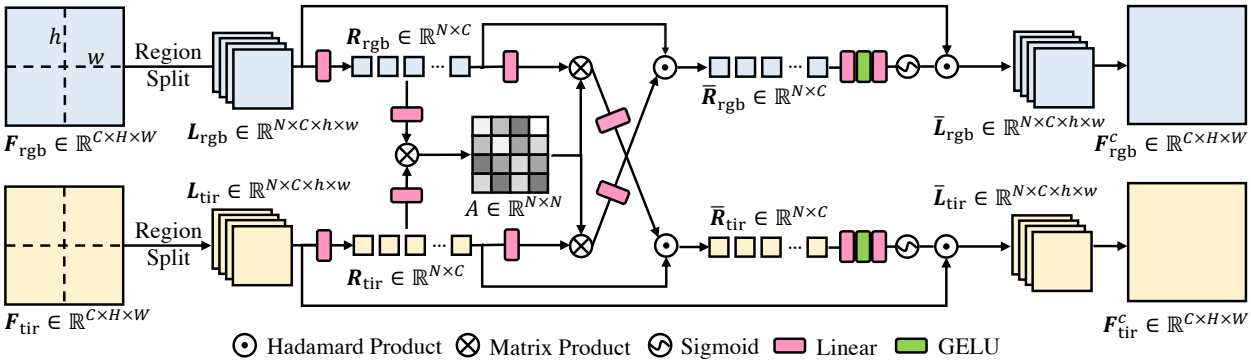

Figure 3: Illustration of our RACM module. We first split $\boldsymbol{F}_{\text{rgb}}$ and $\boldsymbol{F}_{\text{tir}}$ into N local windows, from which we extract region representations $\boldsymbol{R}_{\text{rgb}}$ and $\boldsymbol{R}_{\text{tir}}$. Then, we explore cross-modal correlations via an efficient cross-attention operation and utilize the refined region features $\overline{\boldsymbol{R}}_{\text{rgb}}$ and $\overline{\boldsymbol{R}}_{\text{tir}}$ to generate adaptive channel weights.

$$\boldsymbol{A} = \frac{1}{\sqrt{C}}(\boldsymbol{R}_{\text{rgb}}\boldsymbol{W}_3)(\boldsymbol{R}_{\text{tir}}\boldsymbol{W}_4)^T, \tag{2}$$

where $\boldsymbol{W}_3 \in \mathbb{R}^{C\times C}$ and $\boldsymbol{W}_4 \in \mathbb{R}^{C\times C}$ are projected parameters, $(\boldsymbol{R}_{\text{tir}}\boldsymbol{W}_4)^T$ is the transpose of $\boldsymbol{R}_{\text{tir}}\boldsymbol{W}_4$. Then $\boldsymbol{A}$ is utilized to aggregate the corresponding context as follows:

$$\boldsymbol{R}_{\text{rgb}\to\text{tir}} = \text{Softmax}(\boldsymbol{A})(\boldsymbol{R}_{\text{rgb}}\boldsymbol{W}_5), \tag{3}$$

$$\boldsymbol{R}_{\text{tir}\to\text{rgb}} = \text{Softmax}(\boldsymbol{A}^T)(\boldsymbol{R}_{\text{tir}}\boldsymbol{W}_6), \tag{4}$$

where $\boldsymbol{W}_5 \in \mathbb{R}^{C\times C}$ and $\boldsymbol{W}_6 \in \mathbb{R}^{C\times C}$ are projection parameters, $\boldsymbol{R}_{\text{rgb}\to\text{tir}}$ denotes the context flow from RGB modality to TIR modality, and vice versa. Next, we fuse $\boldsymbol{R}_{\text{rgb}\to\text{tir}}$ and $\boldsymbol{R}_{\text{tir}\to\text{rgb}}$ with original region features via linear transform and Hadamard product:

$$\overline{\boldsymbol{R}}_{\text{rgb}} = \boldsymbol{R}_{\text{rgb}} \odot (\boldsymbol{R}_{\text{tir}\to\text{rgb}}\boldsymbol{W}_7), \tag{5}$$

$$\overline{\boldsymbol{R}}_{\text{tir}} = \boldsymbol{R}_{\text{ir}} \odot (\boldsymbol{R}_{\text{rgb}\to\text{tir}}\boldsymbol{W}_8), \tag{6}$$

where $\boldsymbol{W}_7 \in \mathbb{R}^{C\times C}$ and $\boldsymbol{W}_8 \in \mathbb{R}^{C\times C}$ are projected parameters, $\overline{\boldsymbol{R}}_{\text{rgb}} \in \mathbb{R}^{N\times C}$ and $\overline{\boldsymbol{R}}_{\text{ir}} \in \mathbb{R}^{N\times C}$ are refined region features. Finally, we generate channel-wise modulation weights via the single-layer MLP and sigmoid activation. The modulated local region features $\overline{\boldsymbol{L}}_{\text{rgb}}$ and $\overline{\boldsymbol{L}}_{\text{tir}}$ are obtained as follows:

$$\overline{\boldsymbol{L}}_{\text{rgb}} = \boldsymbol{L}_{\text{rgb}} \odot \boldsymbol{\Psi}_{\text{r}}(\overline{\boldsymbol{R}}_{\text{rgb}}), \tag{7}$$

$$\overline{\boldsymbol{L}}_{\text{tir}} = \boldsymbol{L}_{\text{tir}} \odot \boldsymbol{\Psi}_{\text{t}}(\overline{\boldsymbol{R}}_{\text{tir}}), \tag{8}$$

where $\boldsymbol{\Psi}_{\text{r}}$ and $\boldsymbol{\Psi}_{\text{t}}$ denote a sequential of Linear, GELU, Linear, and Sigmoid layers in RGB and TIR branches, respectively. Finally, we obtain $\boldsymbol{F}^c_{\text{rgb}} \in \mathbb{R}^{C\times H\times W}$ and $\boldsymbol{F}^c_{\text{tir}} \in \mathbb{R}^{C\times H\times W}$ by reshaping $\overline{\boldsymbol{L}}_{\text{rgb}}$ and $\overline{\boldsymbol{L}}_{\text{tir}}$ to the same shape as $\boldsymbol{F}_{\text{rgb}}$ and $\boldsymbol{F}_{\text{tir}}$.

3.2.2. CCSM

The motivation behind the CCSM module is to mine the complementary visual context and transfer it between the two modalities early in the encoding phase. In addition, inspired by group convolution [42], we introduce a channel grouping mechanism to enable finer spatial activation and reduce module size (Figure 4). Concretely, we reshape RGB features $\boldsymbol{F}_{\text{rgb}} \in \mathbb{R}^{C\times H\times W}$ and TIR

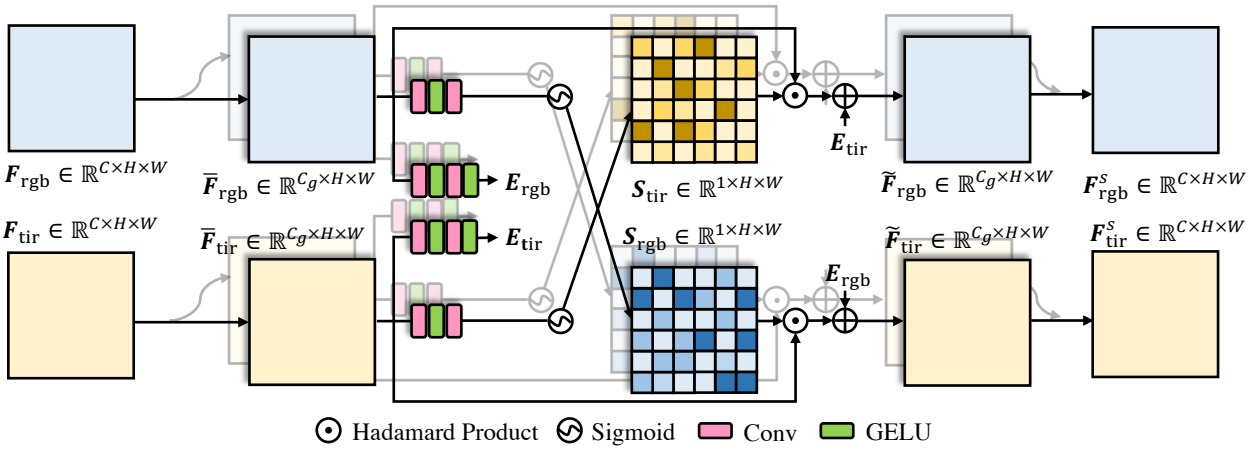

Figure 4: Illustration of our CCSM module. We first suppress spatial distractors in each modality with spatial activation maps $\boldsymbol{S}_{\text{rgb}}$ and $\boldsymbol{S}_{\text{tir}}$, then we transfer complementary context between the two modalities with residual connections.

features $\boldsymbol{F}_{\text{tir}} \in \mathbb{R}^{C\times H\times W}$ to $\mathbb{R}^{g\times C_g\times H\times W}$, where $C_g = C/g$ and g is the number of groups. Then we generate spatial activation maps $\boldsymbol{S}_{\text{rgb}} \in \mathbb{R}^{g\times 1\times H\times W}$ and $\boldsymbol{S}_{\text{tir}} \in \mathbb{R}^{g\times 1\times H\times W}$ as follows:

$$\boldsymbol{S}_{\text{rgb}} = \sigma(\text{Conv}(\phi(\text{Conv}(\boldsymbol{F}_{\text{rgb}})))), \tag{9}$$

$$\boldsymbol{S}_{\text{tir}} = \sigma(\text{Conv}(\phi(\text{Conv}(\boldsymbol{F}_{\text{tir}})))), \tag{10}$$

where $\text{Conv}(\cdot)$ denotes 1×1 convolution layer, σ is sigmoid function, ϕ is GELU activation. And then the complementary spatial details and semantic features are obtained as follows:

$$\boldsymbol{E}_{\text{rgb}} = \phi(\text{Conv}(\phi(\text{Conv}(\boldsymbol{F}_{\text{rgb}})))), \tag{11}$$

$$\boldsymbol{E}_{\text{tir}} = \phi(\text{Conv}(\phi(\text{Conv}(\boldsymbol{F}_{\text{tir}})))), \tag{12}$$

where $\boldsymbol{E}_{\text{rgb}} \in \mathbb{R}^{g\times C_g\times H\times W}$ and $\boldsymbol{E}_{\text{tir}} \in \mathbb{R}^{g\times C_g\times H\times W}$. Overall, the spatially modulated features $\boldsymbol{F}^s_{\text{rgb}}$ and $\boldsymbol{F}^s_{\text{tir}}$ which provide complementary visual context can be obtained via Hadamard Product to suppress spatial distractors and cross-modal residual connections shown in Figure 4.

$$\boldsymbol{F}^s_{\text{rgb}} = \boldsymbol{S}_{\text{tir}} \odot \boldsymbol{F}_{\text{rgb}} + \boldsymbol{E}_{\text{tir}}, \tag{13}$$

$$\boldsymbol{F}^s_{\text{tir}} = \boldsymbol{S}_{\text{rgb}} \odot \boldsymbol{F}_{\text{tir}} + \boldsymbol{E}_{\text{rgb}}. \tag{14}$$

After all the steps above, $\boldsymbol{F}^s_{\text{rgb}} \in \mathbb{R}^{g\times C_g\times H\times W}$ and $\boldsymbol{F}^s_{\text{tir}} \in \mathbb{R}^{g\times C_g\times H\times W}$ are reshaped to $\mathbb{R}^{C\times H\times W}$ finally.

3.3. GFF

The modulated RGB and TIR features are denoted as $\boldsymbol{F}^m_{\text{rgb}}$ and $\boldsymbol{F}^m_{\text{tir}}$, and we use the devised GFF module to integrate them. As it is shown in Figure 5, We first combine $\boldsymbol{F}^m_{\text{rgb}}$ and $\boldsymbol{F}^m_{\text{tir}}$ via element-wise addition. Then we generate selection gates $\boldsymbol{G} \in \mathbb{R}^{C\times H\times W}$ via 3×3 depth-wise convolutions for a lightweight and efficient module. The final fused features $\boldsymbol{D} \in \mathbb{R}^{C\times H\times W}$ are obtained as follows:

$$\boldsymbol{D} = \boldsymbol{G} \odot \boldsymbol{F}^m_{\text{rgb}} + (1 - \boldsymbol{G}) \odot \boldsymbol{F}^m_{\text{tir}}. \tag{15}$$

For multi-stage GFF modules, multi-modal fused features $\boldsymbol{D}^i$ (*i.e.*, $i = 1$ to 4) are obtained, in the decoding phase, we feed hierarchical multi-modal features $\{\boldsymbol{D}^i\}_{i=1}^4$ to the lightweight MLPDecoder [41] to generate the final prediction results $\boldsymbol{P} \in \mathbb{R}^{C_{\text{cls}}\times H_0\times W_0}$, where C_{cls} is the class number of the dataset.

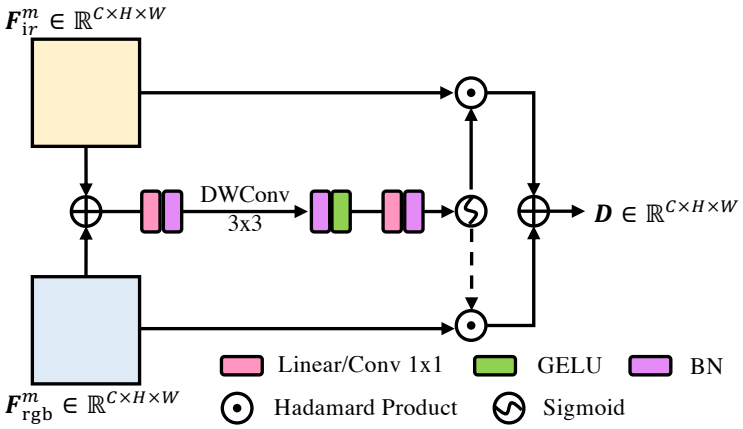

Figure 5: Illustration of our proposed GFF Module.

4. Experiments

4.1. Dataset and Evaluation Criteria

MFNet. MFNet dataset [8] is the first RGB-Thermal urban scene image dataset with pixel-level annotation. It contains 1569 image pairs belonging to 8 classes, including car, person, bike, curve, car stop, guardrail, color cone, and bump. 820 pairs are taken daytime and 749 are taken nighttime, and the resolution of all RGB and TIR images is 480×640. The whole dataset is divided into three parts, specifically, 784 pairs (410 daytime and 374 nighttime) are used for training, 392 pairs (205 daytime and 187 nighttime) for validation, and 393 pairs (205 daytime and 188 nighttime) for testing.

PST900. PST900 dataset [29] includes 894 aligned RGB-Thermal image pairs annotated in 4 foreground classes, including hand-drill, backpack, fire-extinguisher, and survivor. The resolution of all RGB and TIR images is 1280×720. All the image pairs are divided into two parts, of which 597 RGB-T pairs for training and 288 pairs for testing.

We adopt two widely used quantitative evaluation metrics to evaluate the segmentation performance of our method and other compared methods, including **m**ean **I**ntersection **o**ver **U**nion (mIoU) and **m**ean **Acc**uracy (mAcc). They are calculated in the formulas:

$$\text{mIoU} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i + \text{FN}_i}, \tag{16}$$

$$\text{mAcc} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i}. \tag{17}$$

4.2. Implementation Details

We use MiT-B2 [41] pretrained on ImageNet-1K dataset as our encoder. The input images are resized to 480×640 on MFNet and 640×1280 on PST900. AdamW [12] with weight decay 1e-2 is utilized as the optimizer. The width and height of local regions in RACM are empirically set to {10, 10, 5, 5} for different encoding stages. Our whole network is trained with a batch size 8 and a learning rate 6e-5 for 500 epochs with a poly learning rate schedule on a single 32G NVIDIA V100 GPU. In terms of efficiency, we report the FPS values of our method and previous ones on the same machine with a single NVIDIA A800 GPU.

Table 1: Comparison with state-of-the-art methods on the testing set of MFNet dataset. *3c* and *4c* represent that the networks are tested with the three-channel RGB data and four-channel RGB-Thermal data, respectively. Others without *3c* or *4c* indicate that the input is a pair of RGB and Thermal images. "*" denotes the result is replicated based on the author's publicly available checkpoint. The best two performances are respectively highlighted with red and blue colors.

Method	Backbone	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		mIoU	mAcc
		IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc		
MFNet [8]	CNN	65.9	77.2	58.9	67.0	42.9	53.9	29.9	36.2	9.9	12.5	0.0	0.1	25.2	30.3	27.7	30.0	39.7	45.1
FRRN(3c) [25]	CNN	71.2	80.0	46.1	53.0	53.0	65.1	27.1	34.0	19.1	21.6	0.0	0.0	32.5	34.7	30.5	36.2	41.8	47.1
FRRN(4c) [25]	CNN	74.4	81.9	60.8	66.2	50.3	62.8	35.0	41.2	11.5	12.5	0.0	0.0	34.0	37.2	34.6	35.2	44.2	48.5
FuseNet [9]	VGG	75.6	81.0	66.3	75.2	51.9	64.5	37.8	51.0	15.0	17.4	0.0	0.0	21.4	31.1	45.0	51.9	45.6	52.4
DepthAwareCNN [37]	VGG	77.0	85.2	53.4	61.7	56.5	76.0	30.9	40.2	29.3	41.3	8.5	22.8	30.1	32.9	32.3	36.5	46.1	55.1
BiSeNet(3c) [46]	CNN	84.5	90.0	54.3	65.0	61.4	75.0	25.7	32.1	26.2	32.3	0.9	3.2	43.3	49.6	40.5	48.1	48.2	54.9
BiSeNet(4c) [46]	CNN	84.1	89.7	63.2	72.0	60.1	74.1	36.7	45.1	25.3	34.2	5.0	18.2	42.2	47.4	35.9	39.8	50.0	48.2
DFN(3c) [47]	ResNet	81.4	90.7	52.8	67.7	57.5	71.5	34.9	49.2	23.8	35.1	0.9	4.1	31.0	44.2	47.5	54.6	47.5	57.3
DFN(4c) [47]	ResNet	54.5	90.0	65.0	73.2	60.9	75.5	40.4	54.0	25.7	38.9	4.0	10.2	42.5	48.3	47.4	55.8	52.0	60.5
SegHRNet(3c) [31]	HRNet	86.6	92.2	59.8	73.1	61.3	74.9	33.2	47.0	28.7	38.3	1.4	7.3	47.2	54.6	46.2	61.5	51.3	60.9
SegHRNet(4c) [31]	HRNet	87.6	92.8	71.0	79.3	63.4	78.3	42.5	59.8	19.1	25.7	2.7	18.8	49.8	56.5	44.5	63.5	53.2	63.7
RTFNet [32]	ResNet	87.4	93.0	70.3	79.3	62.7	76.8	45.3	60.7	29.8	38.5	0.0	29.8	29.1	45.5	55.7	74.7	53.2	63.1
FuseSeg [33]	DenseNet	87.9	93.1	71.7	81.4	64.6	78.5	44.8	68.4	22.7	29.1	6.4	63.7	46.9	55.8	47.9	66.4	54.5	70.6
AFNet [43]	ResNet	86.0	91.2	67.4	76.3	62.0	72.8	43.0	49.8	28.9	35.3	4.6	24.5	44.9	50.1	56.6	61.0	54.6	62.2
ABMDRNet [50]	ResNet	84.8	94.3	69.6	90.0	60.3	75.7	45.1	64.0	33.1	44.1	5.1	31.0	47.4	61.7	50.0	66.2	54.8	69.5
EGFNet [56]	ResNet	87.6	95.8	69.8	89.0	58.8	80.6	42.8	71.5	33.8	48.7	7.0	33.6	48.3	65.3	47.1	71.1	54.8	72.7
FEANet [6]	ResNet	87.8	93.3	71.1	82.7	61.1	76.7	46.5	65.5	22.1	26.6	6.6	70.8	55.3	66.6	48.9	77.3	55.3	73.2
MFFENet [57]	DenseNet	87.1	91.4	74.4	82.6	61.3	76.1	45.6	58.7	30.6	44.9	5.2	60.0	57.0	64.4	47.4	72.7	55.5	72.3
LASNet [15]	ResNet	84.2	94.9	67.1	81.7	56.9	82.1	41.1	70.7	39.6	56.8	18.9	59.5	48.8	58.1	40.1	77.2	54.9	75.4
FDCNet [52]	ResNet	87.5	94.1	72.4	91.4	61.7	78.1	43.8	70.1	27.2	34.4	7.3	61.5	52.0	64.0	56.6	74.5	56.3	74.1
CCFFNet [40]	ResNet	89.6	94.5	74.2	83.6	63.1	73.2	50.5	67.2	31.9	38.7	4.8	30.6	49.7	55.2	56.3	72.9	57.6	68.3
CMX [49]	MiT-B2	89.4	-	74.8	-	64.7	-	47.3	-	30.1	-	8.1	-	52.4	-	59.4	-	58.2	71.3*
CMX [49]	MiT-B4	90.1	-	75.2	-	64.5	-	50.2	-	35.3	-	8.5	-	54.2	-	60.6	-	59.7	-
Ours	MiT-B2	89.8	95.1	75.1	88.9	65.4	78.2	45.7	69.5	37.4	46.6	10.2	60.0	56.1	65.8	61.6	76.7	60.0	75.5

4.3. Comparison with State-of-the-Art Methods

We compare our method with 18 state-of-the-art methods, including MFNet [8],FRRN [25], FuseNet [9], DepthAwareCNN [37], BiSeNet(4c) [46], DFN(3c) [47], SegHRNet(3c) [31], RTFNet [32], FuseSeg [33], AFNet [43], ABMDRNet [50], EGFNet [56], FEANet [6], MFFENet [57], LASNet [15], FDCNet [52], CCFFNet [40], and CMX [49]. In the tables of experiments, "3c" represents an input of RGB three-channel image, while "4c" refers to an input of RGBT four-channel image, those methods without "3c" or "4c" indicate that images from the two modalities are input separately.

4.3.1. The Overall Results on MFNet Dataset

Table 1 shows the quantitative results of the testing split on the MFNet dataset. Our proposed method achieves a significant improvement over previous SOTA methods. Specifically, our method has an improvement of 2.4 mIoU (4.2%) and 7.2 mAcc (10.5%) than CCFFNet, and a minor increment of 0.1 mAcc, but a significant improvement of 5.1 mIoU (9.3%) than LASNet. Based on the published results and replication, our model outperforms CMX (MiT-B2) by 1.8 mIoU (3.1%) and 4.2 mAcc (5.9%). Though achieving a modest 0.3 mIoU improvement over CMX (MiT-B4), our model significantly reduces computational requirements and achieves double the FPS as shown in Table 4. Meanwhile, our method achieves outstanding results across all categories in the MFNet dataset. In the "Guardrail" category, although our mIoU is relatively low due to limited training samples and few pixels, our result still ranks second only to LASNet.

Table 2: Comparison of daytime and nighttime results on the testing set of MFNet dataset. The best two performances are respectively highlighted with red and blue colors.

Method	Daytime		Nighttime	
	mIoU	mAcc	mIoU	mAcc
FRRN(3c) [25]	40.0	45.1	37.3	41.6
FRRN(4c) [25]	38.0	42.4	42.3	46.2
DFN(3c) [47]	42.2	53.7	44.6	52.4
DFN(4c) [47]	43.9	53.4	51.8	57.4
BiSeNet(3c) [46]	44.5	52.1	45.0	50.3
BiSeNet(4c) [46]	44.8	52.9	47.7	53.1
SegHRNet(3c) [31]	47.2	59.7	49.1	55.7
SegHRNet(4c) [31]	41.4	50.0	44.9	50.2
DepthAwareCNN [37]	42.4	50.6	43.2	50.7
MFNet [8]	36.1	42.6	36.8	41.4
FuseNet [9]	41.0	49.5	43.9	48.9
RTFNet [32]	45.8	60.0	54.8	60.7
FuseSeg [33]	47.8	62.1	54.6	67.3
AFNet [43]	48.1	54.5	53.8	60.2
EGFNet [56]	47.3	66.2	55.0	68.0
MFFENet [57]	47.9	70.5	56.7	70.0
FDCNet [52]	47.8	62.0	56.8	72.2
LASNet [15]	45.2	73.3	58.7	72.8
CCFFNet [40]	50.6	69.7	57.6	66.0
CMX [49]	52.5	-	59.4	-
Ours	51.1	71.2	60.4	74.0

4.3.2. Daytime and Nighttime Results

We also test our method and previous methods in both daytime and nighttime scenarios. As shown in Table 2, our approach has also shown promising results in various scenarios, especially in the nighttime, which indicates that our method has a better performance in low light environments, as well as adaptability to various scenarios and strong generalization capabilities.

4.3.3. Evaluation on PST900 Dataset

As shown in Table 3, our method is superior to other methods on the PST900 dataset. The network we design not only achieves better results in three of the four categories, but also outperforms the other methods in overall testing results.

Table 3: Comparison with state-of-the-art methods on the testing sets of PST900 datasets. The best two performances are respectively highlighted with red and blue colors.

Method	Hand-Drill	Backpack	Fire-Extinguisher	Survivor	mIoU
UNet(3c) [28]	40.26	63.64	49.28	23.37	54.99
UNet(4c) [28]	38.27	52.89	42.96	31.64	52.74
FCN(3c) [16]	30.12	58.15	39.96	28.00	50.98
FCN(4c) [16]	38.58	67.59	46.28	35.06	57.27
MFNet [8]	41.13	64.27	60.35	20.70	57.02
RTFNet [32]	7.07	74.17	51.93	70.11	60.46
CCNet(3c) [11]	32.27	66.42	51.84	57.50	61.42
CCNet(4c) [11]	51.01	72.95	73.80	33.52	66.00
PSTNet [29]	53.60	69.20	70.12	50.03	68.36
ACNet [10]	51.46	83.19	59.95	65.19	71.81
MFFENet [57]	66.79	76.61	79.76	63.01	77.10
EGFNet [56]	64.67	83.05	71.29	74.30	78.51
SA-Gate [3]	81.01	79.77	72.97	62.22	79.05
FDCNet [52]	70.36	72.17	71.52	72.36	77.11
CCFFNet [40]	82.80	75.80	79.90	72.70	82.10
LASNet [15]	77.75	86.48	82.80	75.49	84.80
Ours	83.66	89.24	79.82	77.11	86.16

4.3.4. Experiments of Different Backbones

As shown in Table 4, to further evaluate the effectiveness of our proposed modules (*i.e.*, CM and GFF), we conduct a comprehensive comparison of the computational requirements with other methods using different backbones including VGG, ResNet, and MiT. Under the same training strategy, our proposed modules along with VGG or ResNet backbones still achieve promising results on two metrics, as well as real-time inference speeds. When using the VGG backbone, our approach incorporates a more sophisticated design compared to earlier works, yet it significantly outperforms them in terms of mIoU and mAcc. For the ResNet backbone, our designed CM and GFF modules exhibit high efficiency and effectiveness, achieving performance on par with SOTA models, while significantly surpassing them in terms of computational cost and inference speed. Compared to CMX (Mit-B2), our method has improved the mIoU by 3.1% and the mAcc by 5.9% while maintaining comparable computational cost and inference speed. Additionally, our inference speed is twice as fast as CMX (Mit-B4) while surpassing it with 0.3 mIoU.

4.4. Comparison with Naive Channel-Spatial Attention Method

In order to demonstrate the superity of our proposed CM module, we replace the corresponding modules with traditional channel-spatial attention modules designed in CBAM [39] and compare their performance on MFNet dataset. Specifically, we replace RACM with channel attention and CCSM with spatial attention, and explore both parallel and sequential structures in different orders of traditional attention modules. As shown in the Table 5, with approximately the same inference speed, our proposed RACM-CCSM parallel structure design achieves better performance than the naive one.

In Figure 6, we also visualize the RGB and TIR features output from the first stage of the backbone network. Specifically, we extract RGB and TIR images after they pass through the CM

Table 4: Comparison results with previous state-of-the-art methods using different backbones on the MFNet dataset.

Backbone	Method	#FLOPs (G)	#Param (M)	FPS	mIoU	mAcc
VGG	FuseNet [9]	283.5	44.2	68.0	45.6	52.4
	DepthAwareCNN [37]	-	-	-	46.1	55.1
	VGG+CM+GFF	382.6	57.4	35.0	**55.3**	**68.1**
ResNet	DFN [47]	469.3	47.7	44.0	52.0	60.5
	RTFNet [32]	337.5	254.5	28.7	53.2	63.1
	AFNet [43]	-	-	-	54.6	62.2
	ABMDRNet [50]	566.7	210.9	20.5	54.8	69.5
	EGFNet [56]	201.5	62.8	14.6	54.8	72.7
	FEANet [6]	337.5	255.2	27.1	55.3	73.2
	LASNet [15]	234.4	93.6	27.0	54.9	**75.4**
	FDCNet [52]	159.1	52.9	-	56.3	74.1
	CCFFNet [40]	527.5	109.1	-	57.6	68.3
	ResNet+CM+GFF	104.1	78.4	52.5	**57.9**	73.4
MiT	CMX(B2) [49]	66.9	66.6	34.5	58.2	71.3
	CMX(B4) [49]	134.2	139.9	16.8	59.7	-
	Ours	62.1	77.1	34.3	**60.0**	**75.5**

Table 5: comparasion with naive channel-spatial attention network.

Method	Car		Person		Bike		Curve		Car Stop		Guardrail		Color Cone		Bump		#FLOPs	#Param	FPS	mIoU	mAcc
	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc					
Naive	89.5	94.3	74.7	87.0	65.6	77.2	46.4	67.7	26.9	29.3	6.4	55.5	53.5	59.2	56.6	72.0	60.7	50.9	38.9	57.5	71.9
Ours	89.8	95.1	75.1	88.9	65.4	78.2	45.7	69.5	37.4	46.6	10.2	60.0	56.1	65.8	61.6	76.7	62.1	77.1	34.3	**60.0**	**75.5**

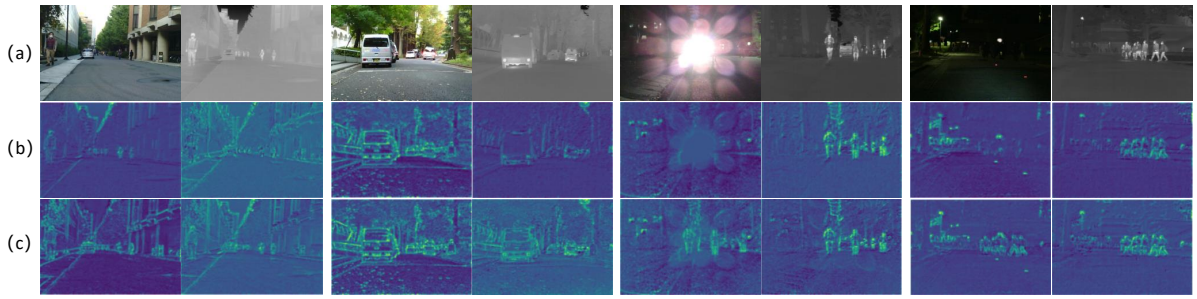

Figure 6: Qualitative comparison between our method and naive channel-spatial attention method on MFNet dataset. We chose 4 representative RGB-T image pairs taken at daytime and nighttime. In each pair, (a) represents RGB-T images, (b) represents the RGB and TIR features output from naive channel-spatial method, and (c) represents features from our method.

module, before being input into GFF. We select four representative examples, and in each example, (a) represents RGB and TIR image pair, (b) represents features output from naive channel-spatial attention CM module, with the left being the RGB feature and the right being the TIR feature, and (c) presents the results of our approach, which utilizes a CM module composed of RACM and CCSM. As it shows, the feature map information obtained by our method is abundant compared with the traditional attention method, for the network has a more precise foreground-background distinguishment, and makes more effective information interaction between different modalities.

For instance, in the 2-nd example, benefitting from our proposed RACM and CCSM modules, the car information in the TIR feature map of our method is much richer than the naive one, and in the 3-rd example, RGB modality succeeds in obtaining foreground target information by exchanging with TIR modality under abrupt illumination.

The thermal infrared camera can capture the temperature difference between objects and provide additional discriminative cues to the visible modality under extreme imaging conditions. In the task setting of RGB-T segmentation, only some common obstacles and visible objects are labeled as targets which require thermal data to recognize them, resulting in serious foreground-background imbalance and introducing a lot of noises. Given this special characteristic of RGB-T segmentation, both our proposed RACM and CCSM contain carefully designed sub-modules to reduce the influence of background noises.

4.5. Ablation Studies

In this section, we carry out a series of ablation studies to validate the effectiveness of different components in our model. All ablation studies are under the same hyper-parameters.

4.5.1. The Effect of Different Components in Our Network

Table 6: The effect of different components in our network. We adopt the same training strategies as before.

Methods	#FLOPs	#Param	FPS	mIoU	mAcc
Baseline	60.0	50.0	44.7	57.0	68.6
Baseline+CM	61.3	76.3	34.6	58.1	72.1
Baseline+GFF	60.7	50.7	43.4	57.9	70.0
Full Model	62.1	77.1	34.3	**60.0**	**75.5**

In this section, we verify the effect of different components of our network, the quantitative results are shown in Table 6. The "Baseline" in the first row represents our base network consisting of the MiT backbone and MLPDecoder. We observe that our transformer-based baseline has a comparable performance with previous non-transformer SOTA methods on two metrics, so introducing transformer backbone into the RGB-T semantic segmentation task may also qualify as a contribution in our work. When we add the CM module, there is a very noticeable improvement of 1.1 mIoU and 3.5 mAcc, demonstrating the effectiveness of our proposed CM module. When we add the lightweight module GFF alone, there is a 0.9 mIoU and 1.4 mAcc improvement. Results show that the well-designed CM module extracts finer features and incorporates more effective modal information than the GFF module, while the lightweight GFF module achieves a quite good result despite its simplicity. When we add both the CM and GFF modules, our full model achieves a substantial increase of 3.0 mIoU and 6.9 mAcc. So it has been proved that our two modules are designed to work harmoniously, effectively utilizing complementary information from different modalities, filtering out background noise, and successfully extracting valuable information for foreground objects. When it comes to computational complexity, although our proposed CM module incorporates operations such as region partitioning and cross-modal attention that result in a decrease in FPS, the full model still exhibits a real-time inference speed.

Table 7: Ablation studies on different components of our proposed RACM and CCSM modules. We adopt the same training strategies as before.

(a) Modulation order.

Order	mIoU	mAcc
RACM → CCSM	58.3	72.1
CCSM → RACM	58.7	73.4
Parallel	**60.0**	**75.5**

(b) Modulation process.

RACM	CCSM	mIoU	mAcc
		57.9	70.0
✓		59.2	72.3
	✓	58.2	70.6
✓	✓	**60.0**	**75.5**

(c) RACM components.

CA	RA	#FLOPs	#Param	FPS	mIoU	mAcc
		61.0	52.6	38.0	58.5	70.5
	✓	61.8	74.9	36.0	58.7	71.0
✓	✓	62.1	77.1	34.3	**60.0**	**75.5**

(d) Group number in CCSM.

g	#Param/K	mIoU	mAcc
1	2324.4	59.3	73.8
2	631.5	58.7	72.5
4	146.4	**60.0**	**75.5**
8	37.2	59.0	72.4

4.5.2. Ablation Studies of RACM and CCSM Modules

Modulation Order. Table 7(a) shows the influence of the modulation order in each encoding stage. Compared with "RACM → CCSM", transferring complementary information between the two modalities first (*i.e.*, "CCSM → RACM") achieves 0.4 mIoU and 1.3 mAcc improvements, indicating that a more complete environmental context is conducive to foreground-background distinguishment. In contrast to the sequential modulation, organizing these two modulation processes in parallel (*i.e.*, "Parallel") is better as it may ease the optimization of each module.

Modulation Process. The RACM module can enhance foreground regions and suppress background regions adaptively via channel modulation. The CCSM module can filter and transfer complementary information between the two modalities. We try to remove them and find that the model performance drops to various degrees in Table 7(b), which indicates that both foreground-background distinguishment and complementary information mining are crucial for RGB-T semantic segmentation.

RACM Components. Table 7(c) shows ablation experiments on components of RACM, where "CA" and "RA" represent cross-attention and region-adaptive channel modulation respectively. When "RA" and "CA" are both removed from RACM, and all spatial locations in each channel share a global weight, the result is shown in the 1-st row. The 2-nd row shows that our region-adaptive channel modulation mechanism leads to a decrease of 2 FPS but brings 0.2 mIoU and 0.5 mAcc gains. As shown in the 3-rd row, applying "CA" in our RACM module can further boost the performance since it can enhance the cross-modal correlation between different modalities. It is worth noting that we do not report the performance where only the "CA" is included in our RACM module, which is because the GPU memory limit is exceeded when applying the cross-attention to the high-resolution low-level features. This problem can be avoided with our efficient "RA" design, which can reduce the computational overhead by switching the scope of cross-attention from pixel-level to region-level, thus facilitating the implementation of "CA".

Group Number in CCSM. We compare different group numbers of CCSM in Table 7(d).

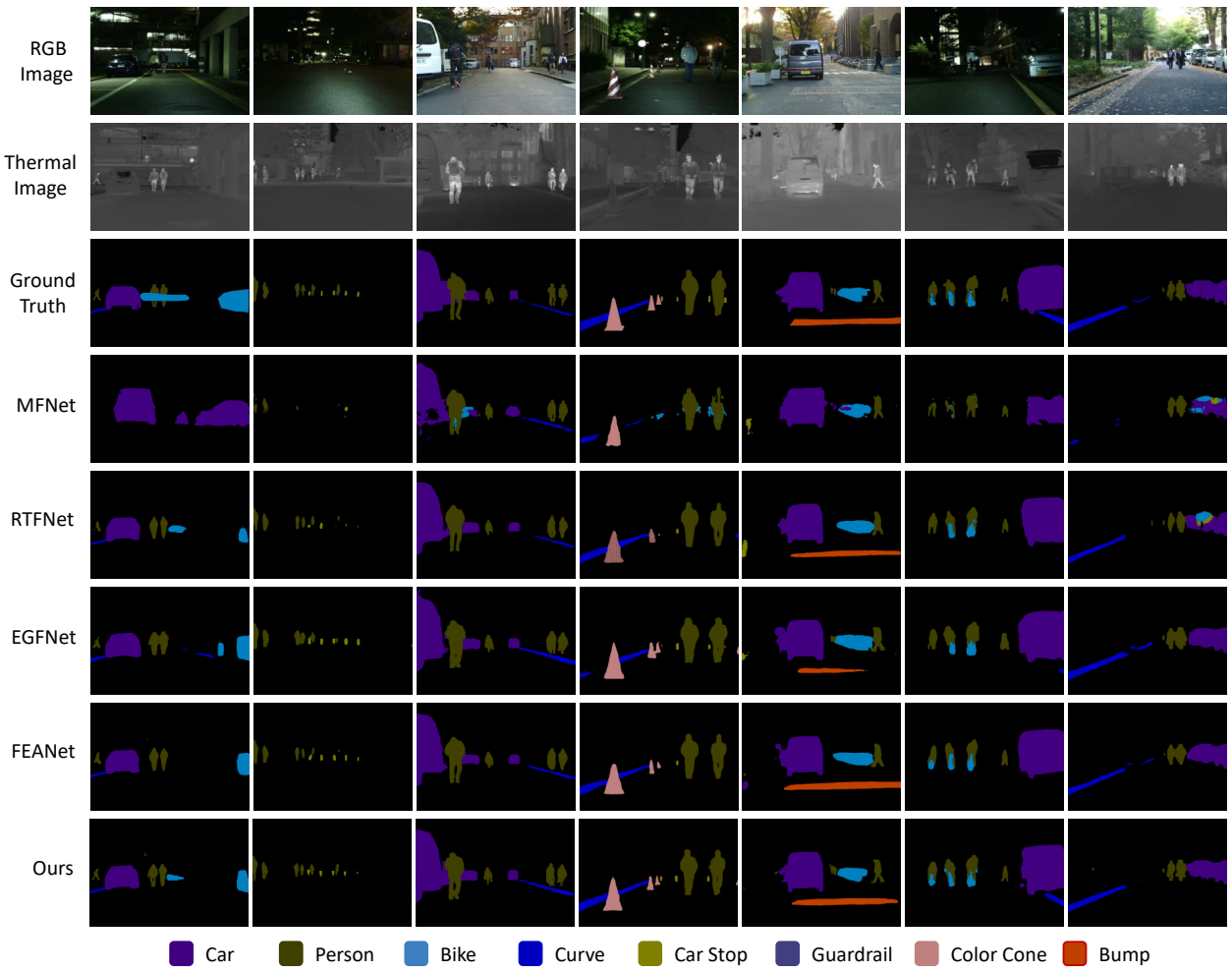

Figure 7: Qualitative comparison between our method and other SOTA methods on the MFNet dataset, which demonstrates the superiority of our method in a variety of different scenarios.

With the increase of the group number, the amount of parameters drops by several times. Compared with the naive non-grouping solution ($g = 1$), our CCSM with 4 groups achieves a better trade-off between the number of parameters and channel weight diversity, resulting in 0.7 mIoU and 1.7 mAcc gains with only 6.3% amount of parameters, which validates the efficiency and effectiveness of our channel grouping mechanism.

4.6. Qualitative Analysis

We select typical samples in the MFNet dataset for qualitative visualization, as shown in Figure 7. Compared with other SOTA methods, it can be found that our method can accurately segment objects in the distance (*e.g.*, the distant person and color cone in the 4-th column). At the same time, We select image pairs in different scenarios, daytime, and nighttime, and our method can also obtain more precise segmentation results under challenging illumination conditions. In general, compared with the ground truth, our method can predict more complete results with solid robustness, zoom in and there will be more details to prove the superiority.

We also visualize the segmentation results of our method on the PST900 dataset, as shown in Figure 8. It can be seen that our method can obtain accurate segmentation results very close to ground truth even under low-illumination underground environments.

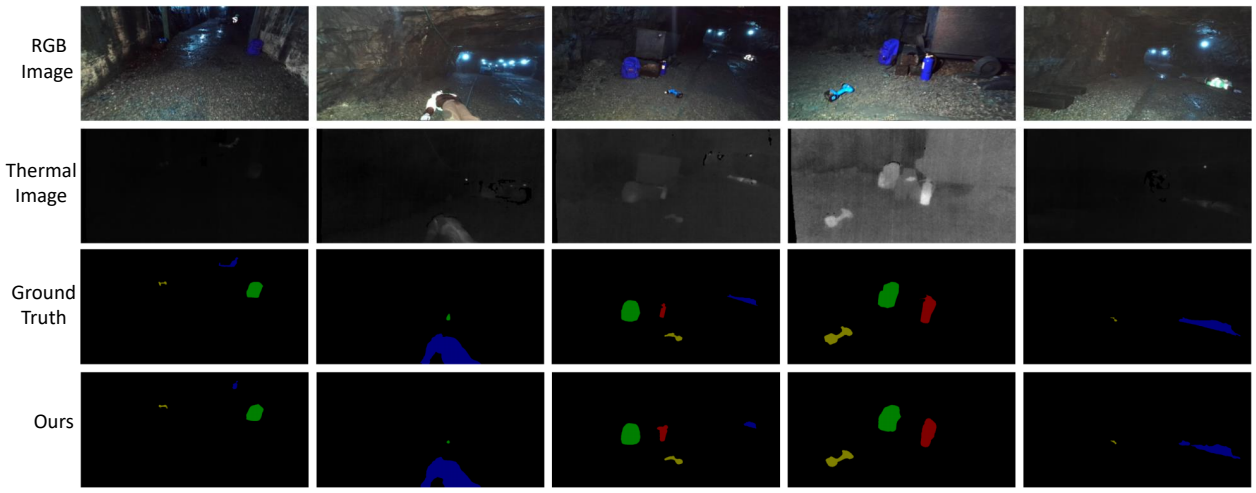

Figure 8: Visualization of our method on PST900 dataset. We select five RGB-T image pairs, and our method achieves results comparable to the ground truth.

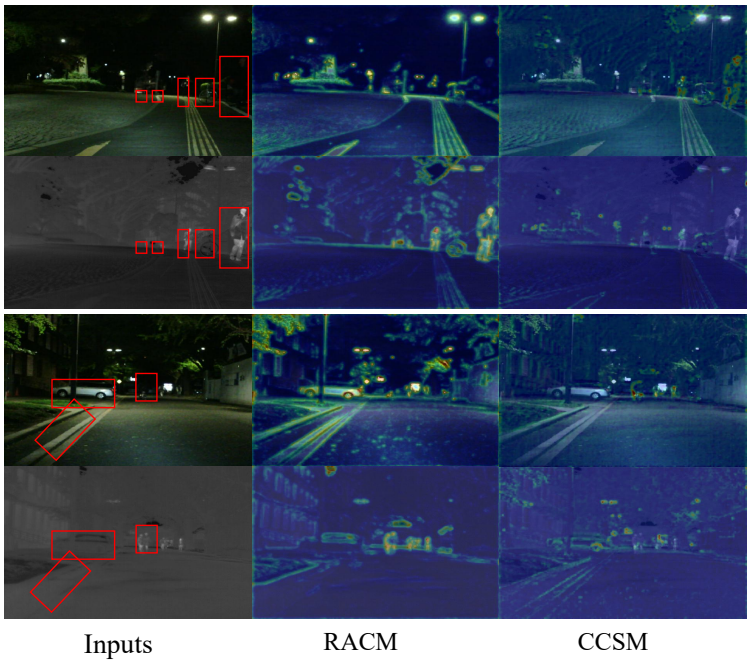

Figure 9: Visualization of features from RACM and CCSM modules. Redder color means a higher response. Objects to be segmented are highlighted with red boxes in the left column.

In Figure 9, we further analyze the RACM and CCSM modules separately via visualization. The 1-st column is the paired RGB and TIR images. The 2-nd and 3-rd columns are features of the two modalities from the 1-st encoding stage, which are modulated by RACM and CCSM respectively. It reveals that RACM can achieve foreground-background distinguishment by activating the foreground regions while suppressing the background regions of the two modalities. Besides, complementary information can be transferred between the two modalities via CCSM. For example, the appearance information of the two people captured in the TIR modality is transferred to the RGB modality in the 1-st row. And the full silhouette of the white car can be completed in the TIR modality as shown in the last row.

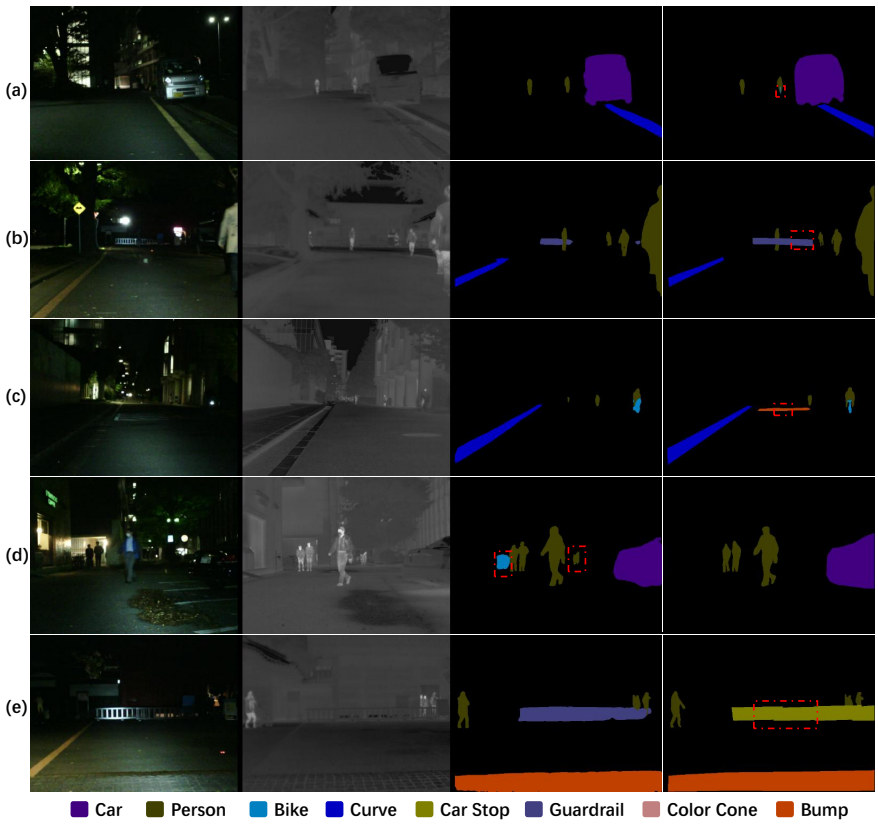

Figure 10: Failure cases on MFNet of our method. Each row represents a case, and from left to right are the RGB image, TIR image, our result, and GT. The red boxes indicate missing or incorrect targets compared to GT.

5. Limitations

Exchanging complementary information between multiple modalities contributes to addressing challenges faced in segmentation task, while in extremely low light conditions, when the object is almost indistinguishable through the RGB image and does not emit enough infrared radiation itself, in other words, there is a paucity of information in either modality, it's difficult to achieve a desired result with the current method in this specific scenario, as the car and guardrail in Figure 10(a)(b). Secondly, factors such as distance and perspective of imaging will make objects visually challenging to identify based solely on their appearances. As the dump in (c), due to the unique visual perspective of the target, it is difficult to determine from a distance whether it is a dump or a flat road surface. Furthermore, because of the high cost of annotation, there are some cases where the dataset has missing or incorrect labels, which can introduce noise into the data and impede the network's ability to achieve optimal performance, as in (d)(e), the bicycles, pedestrians, and guardrail are not properly marked.

6. Conclusion

In this paper, we have proposed a cross modulation process including the Region-Adaptive Channel Modulation (RACM) to achieve fine-grained and efficient foreground-background distinguishment, and the Context-Complementary Spatial Modulation (CCSM) to suppress spatial distractors and transfer complementary information between different modalities. Besides, we integrate information from each modality with the devised Gated Feature Fusion (GFF) module to

obtain the multi-modal features. Experimental results show that our method outperforms previous methods on two popular RGB-T segmentation benchmarks. Furthermore, a comprehensive comparison of the computational requirements and time complexity demonstrates the efficiency and effectiveness of our framework.

Our method provides a general Transformer-based framework including plug-and-play modules (*e.g.*, RACM, CCSM, and GFF) for vision tasks with RGB-T input. In future works, we plan to exploit various RGB-X modality data such as RGB-Depth and RGB-Event apart from RGB-T input and apply our proposed modules to various tasks such as visual object tracking, object detecting, and instance segmentation. The further reduction of computational cost and acceleration of inference speed of our model is also a direction for future exploration. Additionally, with the powerful semantic segmentation capabilities of foundational models like SAM, we plan to enhance segmentation accuracy and improve visual perception in RGB-T semantic segmentation task by integrating multi-modal information while specifying specific targets, leading to a better understanding and analysis of objects and scenes.

References

[1] Chen, C.F., Panda, R., Fan, Q., 2021. Regionvit: Regional-to-local attention for vision transformers. arXiv preprint arXiv:2106.02689 .

[2] Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H., 2018. Encoder-decoder with atrous separable convolution for semantic image segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 801–818.

[3] Chen, X., Lin, K.Y., Wang, J., Wu, W., Qian, C., Li, H., Zeng, G., 2020. Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation, in: European Conference on Computer Vision, Springer. pp. 561–577.

[4] Cheng, B., Misra, I., Schwing, A.G., Kirillov, A., Girdhar, R., 2022. Masked-attention mask transformer for universal image segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1290–1299.

[5] Cheng, B., Schwing, A., Kirillov, A., 2021. Per-pixel classification is not all you need for semantic segmentation. Advances in Neural Information Processing Systems 34, 17864–17875.

[6] Deng, F., Feng, H., Liang, M., Wang, H., Yang, Y., Gao, Y., Chen, J., Hu, J., Guo, X., Lam, T.L., 2021. Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation, in: 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 4467–4473.

[7] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al., 2020. An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 .

[8] Ha, Q., Watanabe, K., Karasawa, T., Ushiku, Y., Harada, T., 2017. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes, in: 2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS), IEEE. pp. 5108–5115.

[9] Hazirbas, C., Ma, L., Domokos, C., Cremers, D., 2017. Fusenet: Incorporating depth into semantic segmentation via fusion-based cnn architecture, in: Asian conference on computer vision, Springer. pp. 213–228.

[10] Hu, X., Yang, K., Fei, L., Wang, K., 2019. Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation, in: 2019 IEEE International Conference on Image Processing (ICIP), IEEE. pp. 1440–1444.

[11] Huang, Z., Wang, X., Huang, L., Huang, C., Wei, Y., Liu, W., 2019. Ccnet: Criss-cross attention for semantic segmentation, in: Proceedings of the IEEE/CVF international conference on computer vision, pp. 603–612.

[12] Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

[13] Kirillov, A., Girshick, R., He, K., Dollár, P., 2019. Panoptic feature pyramid networks, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6399–6408.

[14] Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al., 2023. Segment anything. arXiv preprint arXiv:2304.02643 .

[15] Li, G., Wang, Y., Liu, Z., Zhang, X., Zeng, D., 2022. Rgb-t semantic segmentation with location, activation, and sharpening. IEEE Transactions on Circuits and Systems for Video Technology .

[16] Liu, J., He, J., Zhang, J., Ren, J.S., Li, H., 2020. Efficientfcn: Holistically-guided decoding for semantic segmentation, in: European Conference on Computer Vision, Springer. pp. 1–17.

[17] Liu, S., Wang, K., Yang, X., Ye, J., Wang, X., 2022. Dataset distillation via factorization. Advances in Neural Information Processing Systems 35, 1100–1113.

[18] Liu, S., Ye, J., Yu, R., Wang, X., 2023a. Slimmable dataset condensation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3759–3768.

[19] Liu, S., Zeng, Z., Ren, T., Li, F., Zhang, H., Yang, J., Li, C., Yang, J., Su, H., Zhu, J., et al., 2023b. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. arXiv preprint arXiv:2303.05499 .

[20] Long, J., Shelhamer, E., Darrell, T., 2015. Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 3431–3440.

[21] López-Cifuentes, A., Escudero-Vinolo, M., Bescós, J., García-Martín, Á., 2020. Semantic-aware scene recognition. Pattern Recognition 102, 107256.

[22] Ma, J., Wang, B., 2023. Segment anything in medical images. arXiv preprint arXiv:2304.12306 .

[23] Ouyang, D., He, S., Zhang, G., Luo, M., Guo, H., Zhan, J., Huang, Z., 2023. Efficient multi-scale attention module with cross-spatial learning, in: ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE. pp. 1–5.

[24] Peng, C., Zhang, X., Yu, G., Luo, G., Sun, J., 2017. Large kernel matters–improve semantic segmentation by global convolutional network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4353–4361.

[25] Pohlen, T., Hermans, A., Mathias, M., Leibe, B., 2017. Full-resolution residual networks for semantic segmentation in street scenes, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 4151–4160.

[26] Qi, Y., Wu, Q., Anderson, P., Wang, X., Wang, W.Y., Shen, C., Hengel, A.v.d., 2020. Reverie: Remote embodied visual referring expression in real indoor environments, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9982–9991.

[27] Qiu, Y., Shen, Y., Sun, Z., Zheng, Y., Chang, X., Zheng, W., Wang, R., 2023. Sats: Self-attention transfer for continual semantic segmentation. Pattern Recognition 138, 109383.

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

[29] Shivakumar, S.S., Rodrigues, N., Zhou, A., Miller, I.D., Kumar, V., Taylor, C.J., 2020. Pst900: Rgb-thermal calibration, dataset and segmentation network, in: 2020 IEEE International Conference on Robotics and Automation (ICRA), IEEE. pp. 9441–9447.

[30] Strudel, R., Garcia, R., Laptev, I., Schmid, C., 2021. Segmenter: Transformer for semantic segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 7262–7272.

[31] Sun, K., Zhao, Y., Jiang, B., Cheng, T., Xiao, B., Liu, D., Mu, Y., Wang, X., Liu, W., Wang, J., 2019a. High-resolution representations for labeling pixels and regions. arXiv preprint arXiv:1904.04514 .

[32] Sun, Y., Zuo, W., Liu, M., 2019b. Rtfnet: Rgb-thermal fusion network for semantic segmentation of urban scenes. IEEE Robotics and Automation Letters 4, 2576–2583.

[33] Sun, Y., Zuo, W., Yun, P., Wang, H., Liu, M., 2020. Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion. IEEE Transactions on Automation Science and Engineering 18, 1000–1011.

[34] Tang, L., Xiao, H., Li, B., 2023. Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709 .

[35] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

[36] Wang, Q., Yuan, C., Liu, Y., 2019. Learning deep conditional neural network for image segmentation. IEEE Transactions on Multimedia 21, 1839–1852.

[37] Wang, W., Neumann, U., 2018. Depth-aware cnn for rgb-d segmentation, in: Proceedings of the European Conference on Computer Vision (ECCV), pp. 135–150.

[38] Wei, J., Wu, Z., Wang, L., Bui, T.D., Qu, L., Yap, P.T., Xia, Y., Li, G., Shen, D., 2022. A cascaded nested network for 3t brain mr image segmentation guided by 7t labeling. Pattern Recognition 124, 108420.

[39] Woo, S., Park, J., Lee, J.Y., Kweon, I.S., 2018. Cbam: Convolutional block attention module, in: Proceedings of the European conference on computer vision (ECCV), pp. 3–19.

[40] Wu, W., Chu, T., Liu, Q., 2022. Complementarity-aware cross-modal feature fusion network for rgb-t semantic segmentation. Pattern Recognition 131, 108881.

[41] Xie, E., Wang, W., Yu, Z., Anandkumar, A., Alvarez, J.M., Luo, P., 2021. Segformer: Simple and efficient design for semantic segmentation with transformers. Advances in Neural Information Processing Systems 34, 12077–12090.

[42] Xie, S., Girshick, R., Dollár, P., Tu, Z., He, K., 2017. Aggregated residual transformations for deep neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1492–1500.

[43] Xu, J., Lu, K., Wang, H., 2021. Attention fusion network for multi-spectral semantic segmentation. Pattern Recognition Letters 146, 179–184.

[44] Yang, X., Ye, J., Wang, X., 2022a. Factorizing knowledge in neural networks, in: European Conference on Computer Vision, Springer. pp. 73–91.

[45] Yang, X., Zhou, D., Liu, S., Ye, J., Wang, X., 2022b. Deep model reassembly. Advances in neural information processing systems 35, 25739–25753.

[46] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018a. Bisenet: Bilateral segmentation network for real-time semantic segmentation, in: Proceedings of the European conference on computer vision (ECCV), pp. 325–341.

[47] Yu, C., Wang, J., Peng, C., Gao, C., Yu, G., Sang, N., 2018b. Learning a discriminative feature network for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1857–1866.

[48] Yu, Q., Wang, H., Kim, D., Qiao, S., Collins, M., Zhu, Y., Adam, H., Yuille, A., Chen, L.C., 2022. Cmt-deeplab: Clustering mask transformers for panoptic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2560–2570.

[49] Zhang, J., Liu, H., Yang, K., Hu, X., Liu, R., Stiefelhagen, R., 2023. Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers. IEEE Transactions on Intelligent Transportation Systems .

[50] Zhang, Q., Zhao, S., Luo, Y., Zhang, D., Huang, N., Han, J., 2021. Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 2633–2642.

[51] Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J., 2017. Pyramid scene parsing network, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 2881–2890.

[52] Zhao, S., Zhang, Q., 2022. A feature divide-and-conquer network for rgb-t semantic segmentation. IEEE Transactions on Circuits and Systems for Video Technology .

[53] Zheng, S., Lu, J., Zhao, H., Zhu, X., Luo, Z., Wang, Y., Fu, Y., Feng, J., Xiang, T., Torr, P.H., et al., 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 6881–6890.

[54] Zhou, B., Zhao, H., Puig, X., Fidler, S., Barriuso, A., Torralba, A., 2017. Scene parsing through ade20k dataset, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 633–641.

[55] Zhou, Q., Wu, X., Zhang, S., Kang, B., Ge, Z., Latecki, L.J., 2022a. Contextual ensemble network for semantic segmentation. Pattern Recognition 122, 108290.

[56] Zhou, W., Dong, S., Xu, C., Qian, Y., 2022b. Edge-aware guidance fusion network for rgb–thermal scene parsing, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 3571–3579.

[57] Zhou, W., Lin, X., Lei, J., Yu, L., Hwang, J.N., 2021. Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing. IEEE Transactions on Multimedia 24, 2526–2538.