Information Fusion 120 (2025) 103030

Contents lists available at ScienceDirect

Information Fusion

journal homepage: www.elsevier.com/locate/inffus

Full length article

PEAFusion: Parameter-efficient Adaptation for RGB-Thermal fusion-based semantic segmentation

Yan Wang [a], Henry K. Chu [a], Yuxiang Sun [b,*]

[a] Department of Mechanical Engineering, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China
[b] Department of Mechanical Engineering, City University of Hong Kong, Kowloon, Hong Kong, China

ARTICLE INFO

Keywords:
RGB-Thermal semantic segmentation
Multi-view adapter-pair
Cross-modal self-attention

ABSTRACT

RGB-Thermal (RGB-T) semantic segmentation has attracted great attention in the research community of autonomous driving. Full fine-tuning pre-trained networks is a common strategy in RGB-T semantic segmentation. However, as model size grows, updating all parameters becomes expensive and impractical, which hinders the wide applications of pre-trained networks despite their effectiveness. To efficiently adapt pre-trained single-modality networks to the multi-modal RGB-T task, we design a module named multi-view adapter-pair. The multi-view adapter-pair bridges the gap between pre-trained features and the features required for RGB-T semantic segmentation. It achieves this by approximating high-dimensional updates to the hidden state during full fine-tuning within low-dimensional spaces. Moreover, we propose cross-modal self-attention, constructed using the self-attention operations in pre-trained transformer models. The cross-modal self-attention is designed to fuse RGB and thermal data by expanding the self-attention mechanism in the pre-trained model from a single modality to multiple modalities. Due to the permutation invariance of the attention mechanism and the differences between the two modalities, we introduce modality bias to guide the attention mechanism in learning dependencies inter- and intra-the two modalities. Leveraging these innovations, our network outperforms state-of-the-art methods on the MFNet dataset, as well as the FMB dataset and PST900 dataset, while maintaining parameter efficiency.

1. Introduction

Semantic image segmentation is an essential capability for autonomous vehicles. Deep learning has significantly improved this field [1]. However, conventional deep learning networks, designed for 3-channel RGB images, often degrade under unsatisfactory lighting conditions, such as total darkness. To overcome this problem, researchers have turned to fusing thermal imaging data with RGB images, which has been demonstrated to enhance the overall segmentation performance across varying lighting conditions [2–5].

Training deep neural networks by fully fine-tuning pre-trained models through updating all the model's parameters has been prevalent in computer vision. This method has been demonstrated to be successful due to the effective transferability from well-pre-trained models to specific tasks. However, this method updates all the parameters of a pre-trained model for downstream tasks, which becomes costly and inefficient as the size of pre-trained models grows.

To mitigate this issue, several efforts have been paid to update only a small amount of extra parameters while keeping most pre-trained parameters frozen in training. Adapter Tuning [6] inserts lightweight adapters into each transformer layer, with only these adapters being updated. Prefix Tuning [7] prepends additional l prefix tokens to the input or hidden layers and trains only these tokens for downstream tasks. Another approach, LoRA [8], approximates parameter updates of fully-connected layers with low-rank matrices. These methods make it possible to transfer pre-trained models to downstream Natural Language Processing tasks in a parameter-efficient way, without or with minimal performance sacrifice. Given the success of the parameter-efficient fine-tuning paradigm in Natural Language Processing, researchers have introduced these ideas into computer vision [9–13].

Although the parameter-efficient fine-tuning methods have proven effective, strategies for designing them are less frequently discussed. As the mainstream parameter-efficient fine-tuning methods [6–8] can be formulated as indirect updates to the hidden states [14], we propose designing modules to indirectly mimic the direct updates to the hidden states during full fine-tuning in low-dimensional spaces. Based on this

* Corresponding author.
E-mail address: yx.sun@cityu.edu.hk (Y. Sun).

https://doi.org/10.1016/j.inffus.2025.103030
Received 13 August 2024; Received in revised form 9 January 2025; Accepted 11 February 2025
Available online 28 February 2025

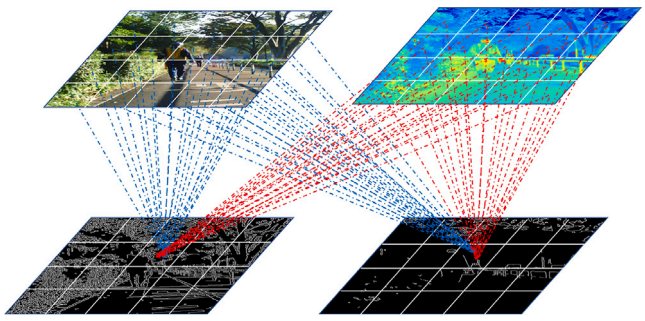


Fig. 1. The illustration of learning inter- and intra-modal dependencies by our cross-modal self-attention. RGB input (top left) and thermal input (top right) correspond to RGB output (bottom left) and thermal output (bottom right). The blue lines denote data flows from the RGB modality, while the red lines denote data flows from the thermal modality.

strategy, we introduce an additive module, named multi-view adapter-pair, to efficiently transfer the pre-trained model to semantic segmentation tasks. The experimental results demonstrate the effectiveness of our strategy.

Another critical aspect of RGB-T semantic segmentation is the fusion method. The current methods for RGB-T fusion can be generally categorized into non-learnable and learnable methods. The non-learnable fusion methods, such as element-wise addition [5], element-wise weighted addition [15], element-wise maximum [16], and feature concatenation [17–19], are popular due to their simplicity. However, these methods usually suffer from inferior performance because they could not adaptively fuse information from different modalities. Therefore, the learnable fusion methods have been introduced to address this issue by incorporating manually designed fusion modules into the network [4,20,21]. These fusion modules can be classified into convolution-based modules and cross-attention-based modules. The convolution-based modules, being insensitive to the input, can achieve multi-scale fusion. However, these modules offer limited explainability, which complicates the analysis of the fusion process. On the other hand, the cross-attention-based modules enhance explainability, but are limited to fusion at the semantic-level due to the substantial computational cost of external cross-attention modules, which also hinders their ability to fuse low-level features [21].

To address these limitations, we propose a novel and interpretable fusion method called cross-modal self-attention (see Fig. 1). This approach fuses RGB and thermal data by constructively learning the dependencies between modalities based on query-key similarities. Using attention maps, this method allows for the analysis of relationships between modalities. The scalability of the attention mechanism also eliminates the obstacle to scale the cross-modal self-attention to more modalities. Most importantly, we implement the cross-modal self-attention in a parameter-efficient way by reusing the self-attention module in the pre-trained backbone, allowing for multi-scale fusion. Our code is open-sourced[1]. The contributions of this work are summarized as follows:

1. We introduce the multi-view adapter-pair as an efficient method to transfer pre-trained single-modality image models to RGB-T semantic segmentation.
2. We design the cross-modal self-attention to fuse RGB and thermal information in a simple, explainable, scalable, and effective way.
3. We achieve superior performance compared to state-of-the-art methods on the MFNet dataset, as well as the FMB and PST900 datasets, while requiring significantly fewer parameter updates.
4. We demonstrate that the transfer from an upstream single modality model to a downstream multi-modal task, such as RGB-T semantic segmentation, can be implemented in a parameter-efficient way.

The remainder of this paper is structured as follows. Section 2 reviews the related work. Section 3 outlines some preliminaries. Section 4 presents the details of our proposed network. Section 5 discusses the experimental results. Section 6 discusses the limitations. Conclusions are drawn in the last section.

2. Related work

2.1. Pre-trained backbone

Since the introduction of AlexNet [22], which marked a shift in the computer vision community from feature engineering and shallow models to deep neural networks, Convolutional Neural Networks (CNNs) [23] have dominated the field for a long time. However, the emergence of the Vision Transformer (ViT) [24] has challenged CNNs' longstanding dominance by demonstrating exceptional performance in image classification tasks. Subsequently, various adaptations of ViT [25–27] have extended the success of transformer to numerous other computer vision tasks.

The success of deep learning stems not only from ingenious structure design but also from meticulously collected datasets, such as ImageNet [28]. Thanks to the vast size and diversity of the ImageNet [28], models trained on ImageNet are able to learn a comprehensive set of features that can be applied to a wide range of downstream computer vision tasks. In our work, we build our deep model based on the Swin Transformer [25,26] pre-trained on ImageNet [28], due to its proven excellent performance in semantic segmentation.

2.2. RGB-only semantic segmentation

The goal of semantic segmentation is to label each pixel in an image into different semantic classes. Early approaches, such as Fully Convolutional Networks (FCNs) [29], laid the foundation by replacing fully-connected layers with convolution layers for pixel-wise predictions. Further advancements were made with U-Net [30], which incorporated skip connections to combine low- and high-level features, enhancing segmentation performance, particularly in medical imaging. To capture multi-scale context, Dilated Convolutions [31] expanded receptive fields without losing resolution, a technique utilized by models like DeepLab [32]. Subsequent models, such as SegFormer [27], focused on integrating global context through pyramid structures and transformers, respectively, to capture multi-scale information and achieve state-of-the-art results. In more recent developments, models like MaskFormer [33] and Mask2Former [34] were introduced. MaskFormer [33] employed mask classification to effectively address both semantic and instance segmentation tasks, demonstrating superior performance over traditional per-pixel classification methods, especially in datasets with a large number of classes. Similarly, Mask2Former [34] utilized masked attention within a Transformer framework to handle various image segmentation tasks, including panoptic, instance, and semantic segmentation. Building on this foundation, our method leveraged the pixel decoder and mask decoder of Mask2Former [34] to address RGB-Thermal segmentation tasks.

2.3. RGB-depth semantic segmentation

RGB cameras provide rich color information but lack spatial depth data, while depth cameras offer abundant spatial depth information. The combination of these two modalities enables the model to achieve a more comprehensive 3D understanding of the scene, which has led to the emergence of the RGB-Depth semantic segmentation task. A key

[1] https://github.com/lab-sun/PEAFusion.

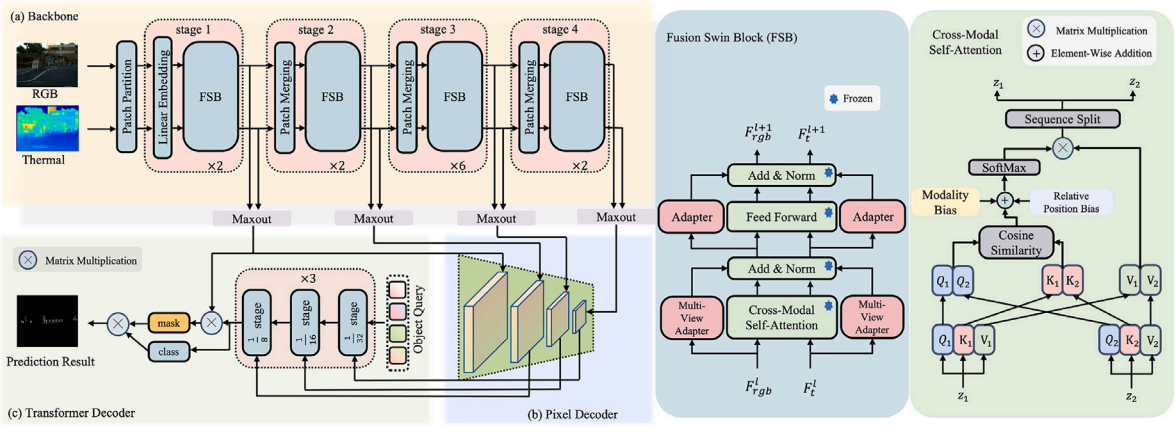


Fig. 2. The structure of our PEAFusion (Tiny). It mainly consists of three components: (a) Backbone, (b) Pixel Decoder, and (c) Transformer Decoder. The Backbone contains a sequence of Fusion Swin Blocks (FSB). Within these Fusion Swin Blocks, cross-modal interaction between the RGB and thermal modalities is achieved via Cross-Modal Self-Attention.

challenge in this field is how to effectively integrate RGB and depth information.

Methods in this area can be broadly categorized into three classes [35]. The first class [36,37] treats depth information as a separate modality, using distinct backbones to extract features from both RGB and depth data, followed by their fusion. The second class [38] incorporates depth information as a bias within the RGB network, thereby aiding in the learning of depth-related features. The third class [39,40] uses depth information as labels, enabling the model to learn both semantic and depth information from a single RGB image.

2.4. RGB-thermal semantic segmentation

Thermal cameras have the unique ability to detect infrared radiation emitted by all objects with a temperature above absolute zero [41]. This capability allows vehicle systems to address the limitations of standard grayscale and RGB cameras, catalyzing the creation of RGB-T semantic segmentation networks. Most RGB-T semantic segmentation networks focus on developing multi-modal fusion modules [4,5,20]. While these modules have greatly improved segmentation accuracy, the mechanisms driving them remain largely unclear, which obstructs efforts to analyze and enhance these fusion modules further.

In our research, we conceptualize the fusion process as the learning of interdependencies between RGB and thermal data and introduce cross-modal self-attention to establish relationships between different modalities. This fusion approach not only offers better performance than previous methods but also saves parameters by being implemented based on the self-attention module in a transformer-based backbone.

2.5. Parameter-efficient fine-tuning

Parameter-efficient fine-tuning technologies [6–8,14] are initially introduced in Natural Language Processing tasks to reduce training costs while maintaining or even surpassing the performance of full fine-tuning. With the recent escalation in model sizes, such as the Swin V2 Giant [26] which possesses 3 billion parameters, the computer vision community has begun to adopt these techniques [9–13,42].

Among the various innovations, AdaptFormer [11] and AIM [9] have implemented the initial adapter concept from Adapter Tuning [6] directly into the ViT layers, facilitating more efficient transfer of pre-trained ViT models. The Convolutional Bypass Adapter [12] integrates the inductive bias of the convolutional layer into the adapter, making it better suited for computer vision tasks. Additionally, LoRand [10] minimizes the parameters of the adapter further using a low-rank approximation approach, employing multiple low-dimensional tensors to represent the relatively high-dimensional fully-connected layer within the adapter. Meanwhile, ViT-Adapter [13] introduces image-related inductive biases to standard ViTs, boosting their performance on dense prediction tasks.

In our study, we investigate parameter-efficient fine-tuning from a different perspective by developing additive parameters that simulate the modifications of the hidden state typically achieved through full fine-tuning. These additive parameters are based on the adapter in Adapter Tuning [6] due to its simplicity.

3. Preliminaries

3.1. A brief review of swin transformer

Models based on the Swin Transformer (Swin) [26] have gained significant popularity in computer vision, particularly in the tasks that require dense prediction, such as RGB-T semantic segmentation [16]. Similar to ViT [24], Swin processes an image as a sequence of small patches. For an image of size $H \times W$, Swin divides the input into N non-overlapping patches and projects them into a specified dimension (denoted as C). Several Swin blocks are applied on these patches. As these Swin blocks maintain the number of patches, a hierarchical representation is produced by reducing the number of patches through a patch merging layer as the network gets deeper. The Swin architecture consists of four stages, each containing of $2n$ successive Swin blocks.

A Swin block consists of a shifted window-based multi-head self-attention (MHSA) module and a feed forward network (FFN) module. It computes self-attention within localized windows and employs a shift window operation to facilitate connections across windows.

To enhance the initial Swin's capacity and window resolution, Swin V2 [26] was developed. The computation of two successive Swin V2 blocks can be written as

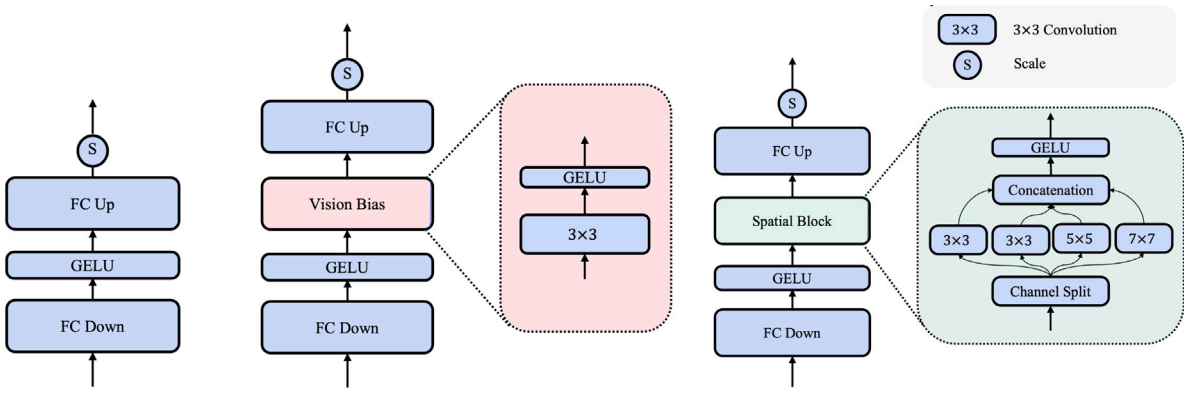
$$\hat{z}^{l} = \text{LayerNorm}\{W\text{-MHSA}(z^{l-1})\} + z^{l-1} \tag{1}$$

$$z^{l} = \text{LayerNorm}\{\text{FFN}(\hat{z}^{l})\} + \hat{z}^{l} \tag{2}$$

$$\hat{z}^{l+1} = \text{LayerNorm}\{\text{SW-MHSA}(z^{l})\} + z^{l} \tag{3}$$

$$z^{l+1} = \text{LayerNorm}\{\text{FFN}(\hat{z}^{l+1})\} + \hat{z}^{l+1} \tag{4}$$

where $\hat{z}^{l}$ and z^{l} represent the output features of the (S)W-MHSA module and the FFN module for block l, respectively. W-MHSA and SW-MHSA refer to window-based multi-head self-attention using regular and shifted window partitioning configurations, respectively. LayerNorm denotes the LayerNorm [43] layer.



(a) Adapter. (b) Conv Bypass Adapter. (c) Multi-view adapter.

Fig. 3. The structures of multi-view adapter-pair (composed of the multi-view adapter and the adapter) and the conv bypass adapter. The outputs from these adapters are scaled before being added to the main branch.

3.2. An overview of adapter tuning

Adapter Tuning [6] offers a parameter-efficient and effective strategy for fine-tuning pre-trained models on downstream tasks by integrating several adapters into the existing model framework. During training, only the newly added adapters and LayerNorm [43] modules are updated, preserving the integrity of the rest of the model. Adapter Tuning [6] is later reframed as modification to specific hidden state within the pre-trained model [14].

Each adapter employs a bottleneck architecture, consisting of two fully-connected (FC) layers with an activation layer situated between them, see Fig. 3(a). The first FC layer reduces the dimensionality of the input, while the second FC layer restores it to its original dimension. In Adapter Tuning [6], a standard transformer block includes two adapters: one following the MHSA module and the other following the FFN module.

4. The proposed method

This section begins with a concise overview of our proposed method (Section 4.1). Subsequently, we provide a detailed explanation of the multi-view adapter-pair (Section 4.2) and the cross-modal self-attention (Section 4.3).

4.1. Method overview

In this work, we propose a method to adapt pre-trained single-modality Swin Transformer models [25,26] for RGB-T semantic segmentation, treating thermal information as an additional modality to enhance scene understanding. Our approach addresses two core components: the efficient transfer of knowledge from single-modality models to multi-modal tasks and the effective fusion of multi-modal information.

4.1.1. Parameter-efficient transfer of pre-trained models

Full fine-tuning could be regarded as direct update of the hidden states in a pre-trained model, achieved by optimizing all parameters to adapt the model to downstream tasks. In our work, we have designed an additive module referred to as multi-view adapter-pair and inserted it into each transformer block. These modules approximate the updates to hidden states achieved through full fine-tuning while operating within a low-dimensional space to reduce computational complexity and improve parameter efficiency. Within the multi-view adapter-pair, the multi-view adapter emulates the role of the attention module, while the adapter functions as part of the feed-forward network [44]. By employing these modules, we achieve parameter-efficient adaptation of the pre-trained single-modality Swin Transformer [25,26] to downstream tasks, enabling effective adaptation from upstream RGB modality to both downstream RGB and thermal modalities in the context of RGB-Thermal semantic segmentation.

4.1.2. Fusion of multimodal information

For modality integration, we leverage the flexibility of the self-attention mechanism [44] in the pre-trained Swin Transformer [25,26] to adaptively fuse modalities without extra fusion modules. Self-attention is well-suited for this task as it captures dependencies among elements, which is crucial for effective modality fusion, enabling an effective transition from single-modality to multi-modality by integrating diverse input sources dynamically. However, given the permutation invariance of self-attention and the distinct characteristics of modalities, we introduce modality-bias, a learnable term specific to each modality. This mechanism effectively guides the learning of intra-modality dependencies and inter-modality interactions, addressing the limitations of basic token concatenation methods, which model dependencies by combining input sources along the token dimension.

4.2. Multi-view adapter-pair

4.2.1. Structure of multi-view adapter-pair

In full fine-tuning, the MHSA constructs both spatial and channel relationships, while the FFN focuses specifically on channel connections. The MHSA module processes a sequence of feature tokens, represented as $X \in \mathrm{R}^{N\times C}$, by first projecting these tokens into a number of lower-dimensional spaces, defined as $X \in \mathrm{R}^{heads\times N\times \frac{C}{heads}}$. Within these spaces, the MHSA learns varying dependencies between tokens through attention operations. Then it projects the tokens back to their original dimensions, returning to $X \in \mathrm{R}^{N\times C}$. Meanwhile, the FFN first projects the input tokens to a certain channel dimension, usually higher than the input dimension, using a fully-connected layer. This is followed by an activation function to introduce non-linearity. Finally, the hidden state is restored to its initial dimension through another fully-connected layer.

Here we design distinct adapters for the MHSA and the FFN modules within a transformer block, individually, to mimic the update of hidden state in full fine-tuning. For the adapter attached parallel to the MHSA module, which we refer to as the multi-view adapter (as shown in Fig. 3(c)), the process begins with a down-sampling fully-connected layer that projects the input into a lower-dimensional space. The features are then divided evenly along the channel dimension into four segments, each processed by convolution layers of different kernel sizes: 3×3 conv, 3×3 conv, 5×5 conv, and 7×7 conv respectively, drawing inspiration from the multi-head mechanism in the MHSA. Subsequently, the features are recombined along the channel dimension and the

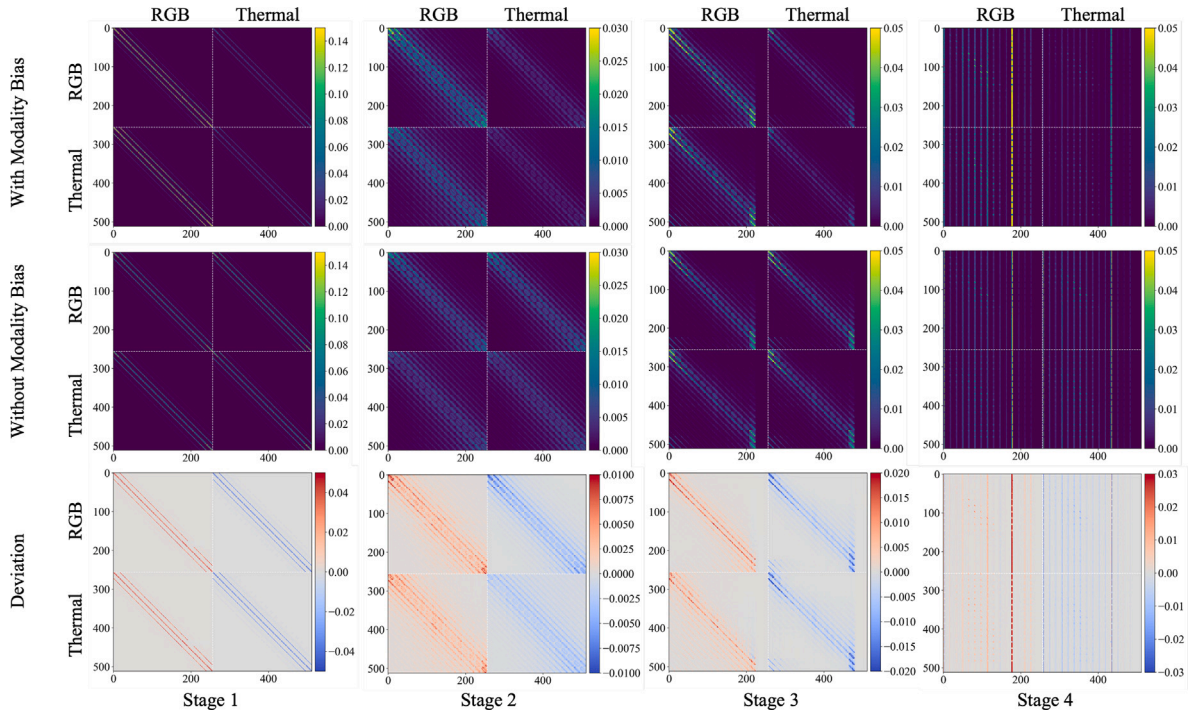


Fig. 4. Visualization of central window attention maps focuses on the first head in the first block of each stage, with the average computed over the MFNet test set to capture the overall distribution. The rows represent: (1) with modality bias; (2) without modality bias; and (3) their differences. Color bar ranges are not standardized to better show attention distribution. PEAFusion-tiny is used with a window size of 16 and an input resolution of 640 × 480.

Table 1
The per-class and average results (%) on the MFNet Dataset. IoU (Intersection over Union), mIoU (mean Intersection over Union), Acc (Accuracy), and mAcc (mean Accuracy) are the evaluation metrics used in this comparison. The best results are highlighted in **bold**. The symbol '–' denotes missing data. The main comparative data come from the article [16].

Method	Venue	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		
RTFNet [5]	RAL 2019	87.4	93.0	70.3	79.3	62.7	76.8	45.3	60.7	29.8	38.5	0	0	29.1	45.5	55.7	74.7	53.2	63.1
ABMDRNet [45]	CVPR 2021	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
FEANet [46]	IROS 2021	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
EAEFNet [47]	RAL 2023	87.6	**95.4**	72.6	85.2	63.8	79.9	48.6	70.6	35.0	47.9	14.2	62.8	52.4	62.7	58.3	71.9	58.9	75.1
IGFNet(B2) [20]	ROBIO 2023	88.0	93.2	74.0	83.4	62.7	71.8	48.2	67.6	36.0	45.4	14.2	68.5	52.4	58.8	57.5	68.3	59.0	72.9
CMX(B2) [4]	TITS 2023	89.4	–	74.8	–	64.7	–	47.3	–	30.1	–	8.1	–	52.4	–	59.4	–	58.2	–
CMX(B4) [4]	TITS 2023	**90.1**	–	**75.2**	–	64.5	–	**50.2**	–	35.3	–	8.5	–	54.2	–	60.6	–	59.7	–
CRM-T [16]	ICRA 2024	90.0	94.8	73.1	85.1	63.7	80.6	47.9	**73.0**	40.7	51.3	9.9	64.4	54.4	60.0	54.2	68.1	59.1	71.8
CRM-B [16]	ICRA 2024	90.0	95.2	75.1	85.6	67.0	81.8	45.2	54.2	49.7	71.2	18.4	12.9	54.2	82.9	54.4	72.9	61.4	72.9
PEAFusion-tiny (ours)		87.6	93.4	72.6	86.3	62.7	78.4	42.9	62.5	45.5	69.6	15.0	18.0	51.2	**89.1**	56.8	**82.6**	59.1	75.4
PEAFusion-base (ours)		89.2	94.6	72.3	87.2	63.5	76.6	46.6	62.8	**49.7**	**77.6**	**20.3**	31.6	52.0	84.5	57.2	78.8	61.0	**77.0**
PEAFusion-large (ours)		88.5	94.4	72.3	84.0	**67.5**	**85.2**	49.2	56.7	49.0	76.2	8.5	8.8	**57.8**	74.5	**69.6**	74.9	**62.3**	72.6

dimensions are restored to their original size through an up-sampling fully-connected layer. Since the FFN module in a transformer block preserves the spatial relationships within a sequence while primarily learning channel relationships, we utilize the initial adapter attached parallel to the FFN module [6].

We consider these two adapters within a block to constitute a multi-view adapter-pair. Table 5 demonstrates how our proposed adapter-pair effectively facilitates the transfer of the pre-trained image model to the RGB-T semantic segmentation task.

4.2.2. Training initialization of multi-view adapter-pair

In parameter-efficient fine-tuning methods, such as those involving additive modules like adapter modules [6] and LoRA modules [8], it is common to initialize these modules to zero or near-zero to minimize their initial perturbation to the pre-trained model. Typically, this involves zero or near-zero initialization of components within the additive module, such as a linear projection layer [6,8,9,11,12]. However, initializing all weights of a linear layer to the same value can lead to issues like symmetry, where neurons learn identical features, hindering the layer's learning effectiveness [48]. Inspired by Highway Networks [49], we introduce a gating mechanism to mitigate the additive module's impact on the pre-trained model at the onset of training without initializing certain components as 0. Specifically, we employ a gate scale—initialized to zero—to regulate the information flow through the additive modules. Meanwhile, the components of the additive module are initialized using standard methods; for example, we initialize the linear layer in the multi-view adapter-pair with a normal distribution [25,26] and the convolution operator with Xavier Initialization [50].

4.2.3. Parameter comparison: Transformer block vs. Multi-view adapter-pair

In this section, we compare the parameter counts of a transformer block and a multi-view adapter-pair. For simplicity, we consider only the linear projection layer of the transformer block. The parameter count of the transformer block [25,26,44] P_T consists of two parts: the

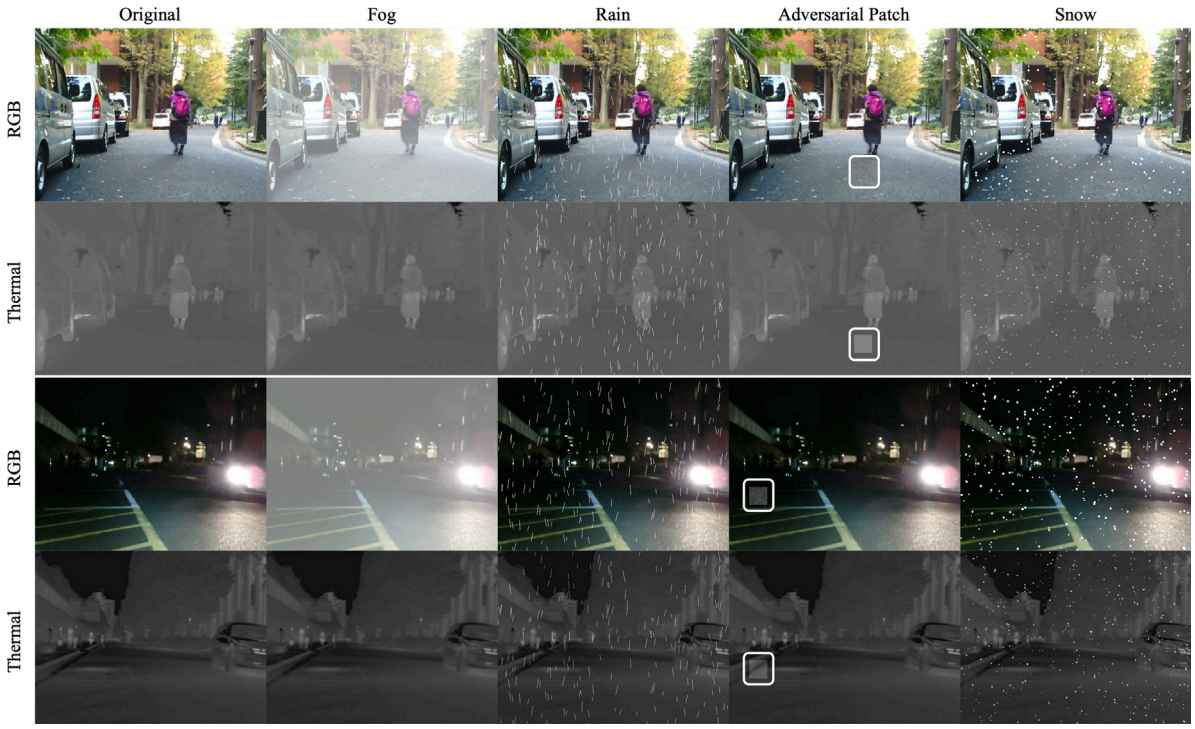


Fig. 5. Illustration of simulated adversarial attack scenarios: fog, rain (depicted by white streaks symbolizing raindrops), snow (depicted by white dots symbolizing snowflakes), and an adversarial patch scene (with the adversarial patch delineated within a white frame).

parameters of the attention module, denoted as P_{attn}, and the parameters of the feed-forward network (FFN), denoted as P_{ffn}. Assuming the input dimension is C, we have:

$$P_T = P_{\text{attn}} + P_{\text{ffn}} = 4 \cdot C \cdot C + C \cdot 4C \cdot 2 = 12C^2 \tag{5}$$

For the multi-view adapter-pair, the parameter count $P_{\text{mv-ap}}$ includes the parameters of the multi-view adapter, denoted as $P_{\text{mv-ad}}$, and the parameters of the adapter, denoted as P_{ad}. Assuming the downsampling ratio in the multi-view adapter is γ_1, and the downsampling ratio in the adapter is γ_2, we have:

$$P_{\text{mv-ad}} = C^2 \cdot \gamma_1 \cdot 2 + \left(\frac{C \cdot \gamma_1}{4}\right)^2 \cdot \left(2 \cdot 3^2 + 5^2 + 7^2\right) \tag{6}$$

$$P_{\text{ad}} = C^2 \cdot \gamma_2 \cdot 2 \tag{7}$$

$$P_{\text{mv-ap}} = P_{\text{mv-ad}} + P_{\text{ad}} \tag{8}$$

Based on our work's settings, where $\gamma_1 = 0.125$ and $\gamma_2 = 0.5$, the proportion of the parameter count between the multi-view adapter-pair and the transformer block is:

$$\frac{P_{\text{mv-ap}}}{P_T} \approx 0.11 \tag{9}$$

4.3. Cross-modal self-attention

The Self-Attention excels at building spatial relationship by learning dependency between the tokens of the input sequence [24–27]. The Self-Attention in Swin V2 [26] could be written as

$$\text{Attention}(Q, K, V) = \text{SoftMax}(\text{cosine}(Q, K)/\tau + B)V \tag{10}$$

where $Q, K, V \in \mathbb{R}^{M^2 \times d}$ are the query, key and value; d is the dimension of $query/key/value$, and M^2 is the number of patches in a window; $B \in \mathbb{R}^{M^2 \times M^2}$ is the relative position bias term for each head.

As shown in Fig. 2, to establish relationships between modalities, we first divide the window to prepare the tokens for attention calculation. Since the RGB and thermal images are highly aligned, we adopt a shared window partitioning approach for both modalities. Next, we concatenate the query, key, and value vectors from each modality along the sequence dimension. Following this, the attention operation is applied to the concatenated vectors. Similarly, because the RGB and thermal images are highly aligned, we retain the relative position bias from the pre-trained model for each modality to distinguish spatial positions within the same modality. To further differentiate tokens across modalities, we use the modality bias alongside the relative position bias for each attention head, much like the relative position bias distinguishes tokens based on their positions. As indicated in Table 7 and Fig. 4, the importance of modality bias is clearly demonstrated. Initially, elements in the RGB bias are set to 2, reflecting the typically richer information content in RGB modality, while elements in the thermal bias start at 0. Both biases are updated during training via gradient backpropagation.

The Cross-Modal Self-Attention (CM-SA) can be described as:

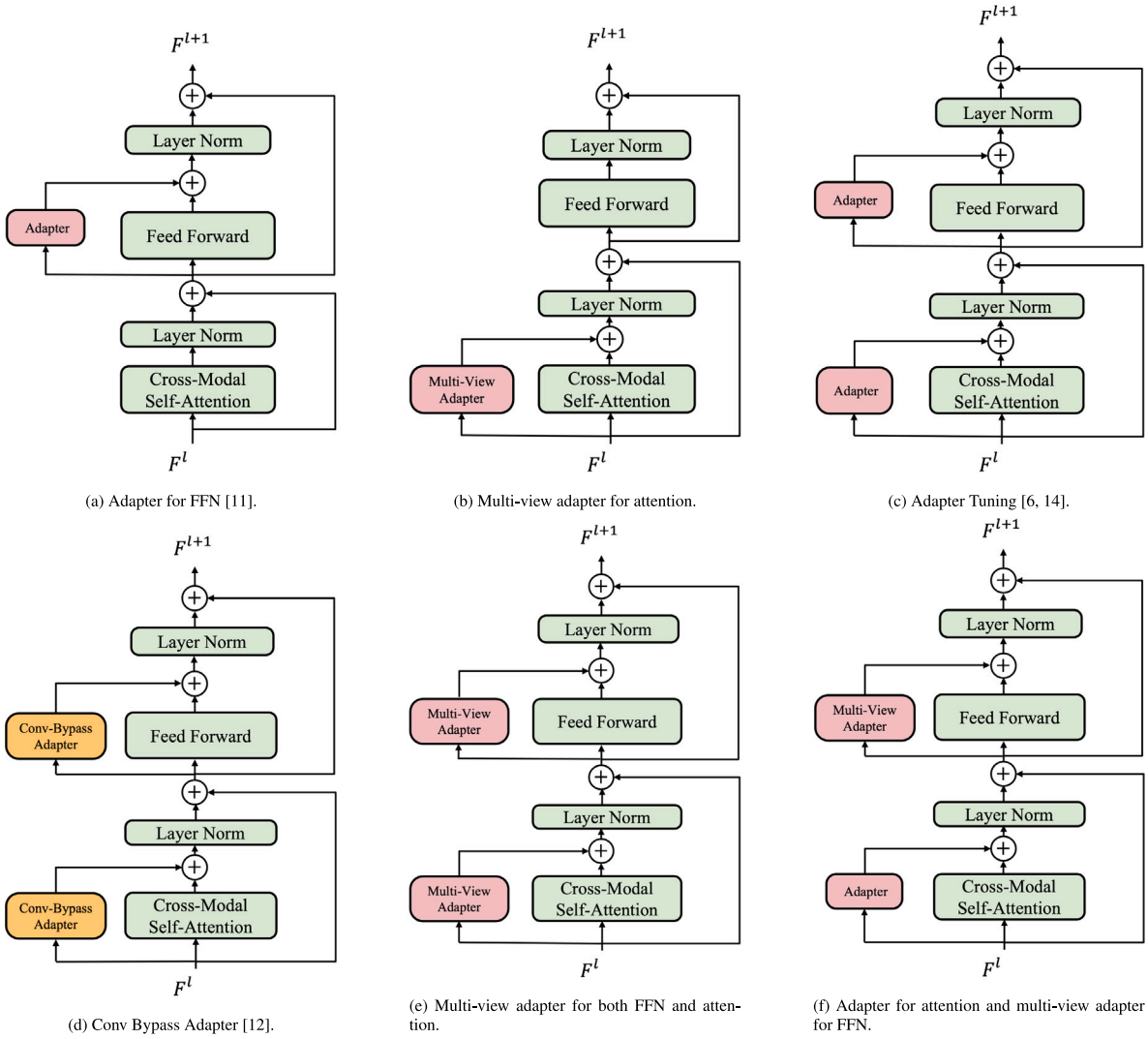
$$\begin{gathered} Q = \binom{Q_1}{Q_2}; K = \binom{K_1}{K_2}; V = \binom{V_1}{V_2}; Bias = \begin{bmatrix} B+\beta_1 & B+\beta_2 \\ B+\beta_1 & B+\beta_2 \end{bmatrix}; \\ \text{CM-SA} = \text{SoftMax}(\text{cosine}(Q, K)/\tau + Bias)V \end{gathered} \tag{11}$$

where $\beta_1, \beta_2 \in \mathbb{R}^{n \times M^2 \times M^2}$ are the modality biases for different modalities; n is the number of heads, and M^2 is the number of patches in a window, and SoftMax normalizes the weighted similarity scores across all patches to ensure that the attention coefficients sum to 1 for each query.

The Cross-Modal Self-Attention (CM-SA) functions like traditional self-attention, serving as a sampling mechanism that is based on the similarity between query and key pairs. This offers enhanced clarity over previous fusion methods. With appropriate position and modality biases, along with effective similarity calculations, CM-SA can seamlessly fuse an arbitrary number of modalities through a single attention operation.

5. Experimental results and discussions

In this section, we introduce the datasets used in our RGB-T semantic segmentation networks, followed by a detailed description of the



(a) Adapter for FFN [11].

(b) Multi-view adapter for attention.

(c) Adapter Tuning [6, 14].

(d) Conv Bypass Adapter [12].

(e) Multi-view adapter for both FFN and attention.

(f) Adapter for attention and multi-view adapter for FFN.

Fig. 6. Illustration of ablation study settings for multi-view adapter-pair.

Table 2
The per-class and average results (%) on the PST900 Dataset. The best results are highlighted in **bold**. The symbol '–' denotes missing data in the original publication. The main comparative data come from the article [16].

Method	Venue	Background		Fire-Extinguisher		Backpack		Hand-Drill		Survivor		mIoU	mAcc
		IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc		
RTFNet [5]	RAL 2019	98.9	–	52.0	–	75.3	–	25.2	–	36.4	–	57.6	–
PSTNet [18]	ICRA 2020	98.9	–	70.1	–	69.2	–	53.6	–	50.0	–	68.4	–
ABMDRNet [45]	CVPR 2021	99.0	–	66.2	–	67.9	–	61.5	–	62.0	–	71.3	–
CRM-T [16]	ICRA 2024	99.5	–	79.1	–	86.0	–	86.2	–	78.7	–	85.9	–
CRM-B [16]	ICRA 2024	99.6	–	79.5	–	89.6	–	89.0	–	82.2	–	88.0	–
MMSFormer [51]	OJSP 2024	99.6	–	81.45	–	89.86	–	89.65	–	76.68	–	87.4	–
PEAFusion-tiny (ours)		99.6	99.7	80.7	89.5	88.8	96.7	89.5	94.4	**85.0**	94.5	88.7	95.0
PEAFusion-base (ours)		**99.7**	99.8	**83.3**	89.2	**90.7**	97.0	**90.2**	93.4	84.3	94.6	**89.6**	94.8

network architecture and training configurations. Next, we present the results of ablation studies to evaluate the effectiveness of our proposed methods and analyze the robustness of our approach against adversarial attacks. We also conduct a comprehensive comparison with CRM [16], a full fine-tuning method that employs a similar model architecture. Finally, we provide a quantitative comparison of our approach with previous methods. We employ the evaluation metrics, Intersection over Union (IoU) and Accuracy (Acc), in our experiments [5].

5.1. Dataset

In this study, we utilize three publicly available RGB-T datasets for the training and evaluation of the proposed methodology.

5.1.1. MFNet dataset [17]

The MFNet dataset comprises a collection of 820 daytime and 749 nighttime RGB-thermal images captured in urban driving scenarios,

Table 3
The per-class (%) results and average results on the FMB Dataset. The best results are highlighted in **bold**. The symbol '–' denotes missing data in the original publication. As class 'Bicycle' is absent from the test set, we report average results excluding the 'Bicycle' class.

Method	Venue	Building		T-Lamp		T-Sign		Vegetation		Person		Car		Truck		Pole		mIoU	mAcc
		IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc		
SegMiF [52]	ICCV 2023	82.0	–	43.1	–	74.8	–	85.0	–	65.4	–	78.3	–	47.3	–	49.8	–	57.6	–
MMSFormer [51]	OJSP 2024	83.0	–	**45.2**	–	**79.7**	–	87.3	–	69.8	–	82.6	–	44.6	–	51.4	–	61.7	–
PEAFusion-tiny (ours)		**84.2**	91.8	38.6	82.4	79.4	93.6	**88.0**	93.0	**72.3**	88.8	**84.7**	91.2	**49.1**	70.3	**54.1**	77.1	**69.8**	85.3

Table 4
The settings for the multi-view adapter-pair in the Backbone.

Setting		PEAFusion-tiny	PEAFusion-base	PEAFusion-large
Multi-View adapter	FC down ratio	0.125	0.125	0.125
Adapter	FC down ratio	0.5	0.5	0.5
Scale		4.0	4.0	4.0

with a resolution of 640 × 480 pixels. It includes semantic annotations for nine categories, which consist of one unlabeled class and eight classes corresponding to common urban objects.

5.1.2. PST900 dataset [18]

The PST900 dataset consists of 894 synchronized RGB-thermal image pairs with a resolution of 1280 × 720, captured in cave and subterranean environments for the DARPA Subterranean Challenge. It includes per-pixel human annotations across four object classes, with one background class (unlabeled).

5.1.3. FMB dataset [52]

The FMB dataset contains 1500 well-registered infrared and visible image pairs, each annotated with 14 pixel-level categories. It covers a diverse range of environments, including dense fog, heavy rain, and low-light conditions, providing rich scenes under varying illumination. With images of 800 × 600 resolution, the dataset is designed to enhance the generalization ability of fusion and segmentation models.

5.2. Implementation details

In this subsection, we present the details of the networks and training settings. The network architecture is shown in Fig. 2.

5.2.1. Backbone

We employ Swin V2 (tiny, base, and large) [26] pre-trained on ImageNet [28] as our base backbones and augment them by incorporating the multi-view adapter-pair and cross-modal self-attention into each Swin V2 block. We refer to this enhanced block as the Fusion Swin Block. Depending on the size of the pre-trained Swin V2 model, we label our networks as PEAFusion-tiny, PEAFusion-base, and PEAFusion-large. PEAFusion refers to parameter-efficient adaptation for multi-modal fusion-based semantic segmentation.

As illustrated in Fig. 2, traditional self-attention is substituted with the cross-modal self-attention by reconstructing the forward pass process and incorporating modality bias. Additionally, two multi-view adapters are aligned parallel to the cross-modal self-attention module. Furthermore, the FFN module is equipped with two parallel adapters [6], each serving a different modality. Detailed settings for the multi-view adapter-pair configurations can be found in Table 4. In addition, the feature maps of the RGB modality and the thermal modality, output from the backbone, undergo a maxout [16,53]operation to reduce dimensionality.

5.2.2. Head network

In our design, we incorporate Mask2former [34], mainly including Pixel Decoder and Transformer Decoder, as the head network of our architecture.

Table 5
The results (%) of the ablation study on the components in the multi-view adapter-pair in terms of mIoU. The term 'parameters' refers to the number of parameters introduced by this configuration, measured in millions.

Method	Parameters	mIoU
Adapter (FFN)	4.3 M	57.6
Multi-View Adapter (Attn)	1.5 M	57.0
Adapter (Attn) + Adapter (FFN)	5.4 M	58.4
Adapter (Attn) + Multi-View Adapter (FFN)	11.6 M	57.1
Multi-View Adapter (Attn) + Multi-View Adapter (FFN)	12.0 M	56.6
Conv Bypass Adapter [12]	15.7 M	56.6
AdapterFormer [11]	4.3 M	57.0
Multi-View Adapter (Attn) + Adapter (FFN) (ours)	5.7 M	59.1

5.2.3. Training settings

Our networks are built with the PyTorch [54] and Detectron2 [55] libraries, and our experiments are accelerated by RTX 3090 GPU. For data augmentation, we employ techniques such as random color jittering [56], random horizontal flipping, and random cropping on both RGB and thermal images. AdamW [57] optimizer is adopted for all experiments.

5.3. Ablation study for multi-view adapter-pair

In this subsection, we analyze the efficacy of the multi-view adapter-pair. Here we illustrate the additive module setting for only one modality for simplicity, as different modalities share the same setting. Ablation studies are carried out using the PEAFusion-tiny configuration. Unless otherwise specified, the ablation study experiments are conducted on the MFNet dataset.

5.3.1. Ablation on the components in the multi-view adapter-pair

Here, we conduct two experiments to verify the effectiveness of each component in the multi-view adapter-pair. In the first experiment, we attach only the multi-view adapter to the attention module, while in the second, we attach only the adapter to the FFN module. The experimental results show that, without the adapter (i.e., using only the multi-view adapter for the attention module, as shown in Fig. 6(b)), there is a loss of 2.1 mIoU. Similarly, when only the adapter for the FFN module is used (as depicted in Fig. 6(a)), a loss of 1.5 mIoU is observed.

5.3.2. Ablation on the strategy for designing the additive adapter

In Section 4.2, we propose a low-dimensional approximation of the high-dimensional updates to the hidden states in the transformer-based block during full fine-tuning to achieve parameter-efficient transfer. Specifically, we simulate the role of the attention module in learning spatial-channel relationships and the function of the FFN module in establishing channel-wise relationships. To validate the effectiveness of our approach, we conduct three experiments.

Table 6
The results (%) of the ablation study on the initialization methods for the multi-view adapter-pair in terms of mIoU.

Setting	Zero-initialization	Gate-initialization
AdapterFormer	57.0	57.6
PEAFusion-tiny (ours)	57.1	59.1

Table 7
The results (%) of the ablation study on the effectiveness of cross-modal self-attention in terms of mIoU.

Method	mIoU
Baseline	57.8
Baseline+ CM-SA (without modality bias)	56.9
Baseline+ CM-SA (with modality bias) (ours)	59.1

First, we swap the positions of the multi-view adapter and the adapter (illustrated in Fig. 6(f)) and observe a loss of 2.0 mIoU, despite adding 5.9M additional parameters. In the second experiment, we attach the multi-view adapter to both the attention and FFN modules, see Fig. 6(e). After introducing an additional 6.3M parameters, a performance drop of 2.5 mIoU is observed. Finally, when we attach the adapter to both the attention and FFN modules, see Fig. 6(c), a milder performance loss of 0.7 mIoU is recorded. This behavior can be interpreted as follows: attaching the adapter to the attention module essentially introduces an identity projection. However, this identity projection is highly ineffective in capturing the complexities of spatial relationships, which explains the performance degradation observed in comparison to the multi-view adapter-pair method. The results of these experiments strongly support the design strategy we propose for the additive adapter.

5.3.3. Ablation on the initialization method of the multi-view adapter pair

In this section, we validate the gate initialization method proposed in Section 4.2.2. We conduct comparative experiments using the common zero-initialization method as a baseline, comparing both our method and the AdapterFormer [11] method. For our approach, the gate initialization shows an improvement of 2.0 mIoU compared to zero-initialization. Additionally, we observe a performance increase of 0.6 mIoU for the AdapterFormer [11] method (see Table 6).

5.3.4. Comparison to other parameter-efficient fine-tuning methods

Compared to other adapter-based parameter-efficient fine-tuning methods, such as AdapterFormer [11], shown in Fig. 6(a), and Conv Bypass Adapter [12], illustrated in Fig. 6(d), our multi-view adapter-pair achieves superior performance, with improvements of +2.1 mIoU and +2.5 mIoU, respectively (see Table 5).

5.4. Ablation on cross-modal self-attention

In this subsection, we perform a series of experiments to investigate the impact of the cross-modal self-attention. The results of these experiments are listed in Table 7. We start with two frozen Swin V2 [26] backbones equipped with two parallel multi-view adapter-pairs as our baseline. Then, we construct cross-modal self-attention without modality bias. Since the RGB and thermal images are well-aligned, we apply the same window partitioning strategy for the both modalities and share the relative position bias. As shown in Fig. 4, under the condition of shared position relationships, both intra-modality and inter-modality attention maps exhibit high spatial consistency. This suggests that even without modality bias, cross-modal self-attention can establish spatial relationships between modalities. However, this cross-modal self-attention setup without modality bias leads to a performance drop of 0.9 mIoU. When modality bias is introduced, we observe a performance improvement of 1.3 mIoU compared to the baseline.

To further investigate the role of modality bias, we visualize the differences between the attention maps with and without modality bias. Our analysis reveals that, with the introduction of modality bias, the importance of the RGB modality relative to the thermal modality increases in the attention maps. We explain this effect by noting that the RGB modality contains richer information compared to the thermal modality. The presence of modality bias enables the model to focus more on the RGB modality rather than treating both modalities equally, which enhances the model's ability to better understand the scene. This also indicates that ignoring the differences between input modalities and constructing cross-modal attention through naive token dimension concatenation can potentially lead to performance degradation, even when spatial relationships between modalities are considered.

5.5. Robustness against adversarial attacks

We simulate several adversarial attack scenarios to evaluate the robustness of our method, which are illustrated in Fig. 5.

As shown in Table 8, both methods experience a certain degree of performance degradation under the simulated extreme scenarios of fog, rain, and snow. In the simulated sensor interference scenarios, both methods demonstrate strong robustness. The consistent performance variations across the two methods also suggest that current models exhibit instability in extreme weather conditions, which may be attributed to the absence of such scenarios in the dataset. Nonetheless, extreme scenarios play a critical role in ensuring the reliable operation of autonomous systems, highlighting the importance of developing more robust systems for these conditions in future research.

In contrast, the performance loss in the simulated sensor interference scenarios is negligible. This may be because the adversarial patch on the thermal images is less deceptive compared to its appearance on RGB images, allowing the thermal modality to assist the model in mitigating the confusion caused by the adversarial patch.

5.6. Comprehensive comparison to state-of-the-art CRM [16]

CRM [16] proposes a complementary random masking strategy and a self-distillation loss to encourage the network to extract complementary and meaningful representations from a single modality or complementary masked modalities. The CRM method has achieved state-of-the-art results on multiple datasets, including MFNet and PST900.

Here, we compare our method with CRM, as both utilize the same Swin [25,26] backbone and Mask2Former [34] head. We compare CRM and our proposed method across three aspects. First, we evaluate their performance on the MFNet and PST900 datasets, using metrics such as mIoU, mAcc, learnable parameters, and total parameters. Second, we compare the training performance of CRM-T and PEAFusion-tiny, focusing on training speed (measured by iterations per second) and memory usage. Finally, we assess the inference performance of CRM-T and PEAFusion-tiny, measured in frames per second (fps).

5.6.1. Performance and parameters comparison

As shown in Table 9, our method demonstrates significantly fewer trainable parameters compared to CRM, with 34.3M vs. 74.9M and 44.9M vs. 193M. On the MFNet dataset, our method achieves comparable performance to CRM in terms of mIoU. However, when evaluated using mAcc, our method exhibits a clear advantage. On the PST900 dataset, PEAFusion-tiny outperforms the corresponding CRM-T by a considerable margin in terms of mIoU and even surpasses CRM-B.

5.6.2. Training performance comparison

As illustrated in Table 10, our method demonstrates significant advantages over CRM during the training process. With the same batch size, our method reduces memory usage by approximately 50% compared to CRM, while achieving nearly a twofold improvement in training speed.

Table 8
Robustness against adversarial attacks.

Method	Fog		Rain		Snow		Patch		Original	
	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
CRM-T	55.4	67.8	54.9	68.6	55.8	72.57	58.9	71.8	59.1	71.8
PEAFusion-tiny (ours)	55.1	73.4	52.4	75.7	52.3	74.1	58.9	75.4	59.1	75.4

Table 9
Performance and parameter comparison. Parameters are measured in millions. The notation '–' indicates missing values in the original paper.

Method	Learnable parameter	Total parameter	Performance on MFNet		Performance on PST900	
			mIoU (%)	mAcc (%)	mIoU (%)	mAcc (%)
CRM-T	74.9 M	74.9 M	59.1	71.8	85.9	–
PEAFusion-tiny (ours)	34.3 M	61.9 M	59.1	75.4	88.7	95.0
CRM-B	193 M	193 M	61.4	72.9	88.0	–
PEAFusion-base (ours)	44.9 M	131 M	61.0	77.0	89.6	94.8

Table 10
Training performance comparison. The term 'it/s' denotes iterations per second, while 'Memory' refers to video memory usage during training.

Method	Batch size = 8		Batch size = 4	
	Memory	Speed	Memory	Speed
CRM-T	22.7 GB	0.7 it/s	12.9 GB	1.1 it/s
PEAFusion-tiny (ours)	11.7 GB	2.2 it/s	6.9 GB	3.4 it/s

Table 11
Inference performance comparison. The numbers in the table represent the frames per second (fps) measured with a batch size of 1.

Method	2080Ti	3090	V100	A100	A40
CRM-T	14.1	19.6	9.5	12.6	17.9
PEAFusion-tiny (ours)	8.5	12.3	7.0	8.4	10.8

5.6.3. Inference performance comparison

Here, we compare the runtime performance of our model, PEAFusion-tiny, with CRM-T across different GPUs, measured in frames per second (fps). The GPUs evaluated include the NVIDIA RTX 2080 Ti, RTX 3090, Tesla V100, Tesla A100, and RTX A40. As observed in Table 11, PEAFusion-tiny shows a disadvantage in inference speed of about 30% compared to CRM-T, which may be attributed to the inference latency introduced by the additive adapters.

5.6.4. Performance comparison summary

Overall, our method achieves comparable or superior performance with relatively fewer total parameters and learnable parameters. During the training process, our approach demonstrates advantages in reducing memory usage and improving training speed. However, during the inference phase, our method exhibits higher inference latency compared to CRM [16]. This highlights the need for future research to explore methods that balance both training efficiency and inference efficiency.

5.7. Per-class results comparison

From Table 1, we observe that our method achieves a 0.9 mIoU improvement on the MFNet dataset, with notable accuracy gains in detecting challenging classes such as bike, color cone, and bump. According to Table 2, on the PST900 dataset [18], our method achieves at least a 1.6 mIoU improvement over the previously best-performing model. Furthermore, as shown in Table 3, our method achieves an improvement on the FMB dataset [52], surpassing previous methods across most classes.

6. Limitations

While we propose a method for parameter-efficient fine-tuning of upstream single-modality vision backbones for downstream multi-modal perception tasks, it is subject to certain limitations: (1) Our approach is trained on datasets with well-aligned RGB-Thermal image pairs. While such alignment ensures optimal performance during experiments, real-time systems often involve RGB and thermal images that are not perfectly aligned, posing challenges for practical applications. However, our proposed cross-modal self-attention mechanism does not strongly depend on precise alignment between RGB-Thermal pairs. This motivates future research to explore its application on misaligned pairs, aiming to enable dynamic real-time perception in practical scenarios; (2) While the multi-view adapter-pair method is parameter-efficient and effective for transfer learning, it introduces additional latency during inference, a common drawback of adapter-based approaches. Future work could aim to mitigate or eliminate this latency to improve deployment efficiency without compromising the benefits during training; (3) Although the modality bias demonstrates strong performance in guiding dependencies, its design is tailored to our experimental settings. Exploring adaptive designs that dynamically adjust the modality bias based on input data could enhance its applicability in real-world scenarios, enabling human-like adaptability. Despite these limitations, our method significantly improves parameter efficiency in multi-modal perception tasks and provides a strong foundation for future enhancements. Addressing the identified constraints will further unlock its potential in broader applications.

7. Conclusions and future work

Our work addresses the challenges of parameter-efficient fine-tuning for RGB-Thermal semantic segmentation. We propose the strategy of external lightweight modules to indirectly approximate the direct updates of hidden states in the full fine-tuning process within low-dimensional spaces. By analyzing the construction of spatial-channel relationships within the pre-trained model, we introduce the multi-view adapter-pair to enable parameter-efficient knowledge transfer from the upstream single-modality model to the downstream RGB-T semantic segmentation task. Meanwhile, considering the scalability of the attention mechanism, we extend the self-attention modules in pre-trained transformer models to accommodate multi-modalities, achieving adaptive fusion between the RGB and thermal modalities. Additionally, we highlight the importance of considering the inherent differences between modalities when fusing features via the attention mechanism, offering valuable insights for other multi-modal research efforts. In training, we freeze the main body of the model, significantly reducing memory usage during training and lowering hardware memory requirements, thereby enhancing the practical utility of our method. Furthermore, we enable easy-to-interpret analysis of modality fusion by visualizing the attention maps during inference, which broadens the model's potential real-world applications. In terms of performance, our approach achieves superior results compared to state-of-the-art methods on the MFNet, FMB, and PST900 datasets, while maintaining parameter efficiency. These results underscore the potential of our method for multi-modal applications.

Future research could address several key challenges to improve the performance and applicability of RGB-Thermal semantic segmentation systems. One major issue is handling misaligned RGB-Thermal data, which often arises due to differences in sensor positions, sensor types, orientations, or temporal misalignment between the modalities. Developing robust models capable of handling these misalignments would significantly improve accuracy and generalization across various environments. Another important area for future work is reducing inference latency. While our approach maintains parameter efficiency, real-time applications in autonomous vehicles or robotics require even faster processing times. Research could explore techniques such as model compression, knowledge distillation, or hardware acceleration to achieve faster inference speeds without compromising result quality. Additionally, exploring more efficient attention mechanisms, such as sparse attention or low-rank approximations, could help reduce the computational cost of multi-modal fusion and improve overall model efficiency. A promising direction also lies in utilizing large language models (LLMs) to guide open-vocabulary perception tasks in downstream applications. With their powerful zero-shot learning capabilities, LLMs can serve as valuable tools to interpret and contextualize sensory inputs, such as RGB and thermal data, in real-time. This would enable systems to recognize and understand novel objects, actions, and environments without explicit prior training.

CRediT authorship contribution statement

Yan Wang: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Henry K. Chu:** Writing – review & editing, Supervision, Project administration, Conceptualization. **Yuxiang Sun:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was supported in part by Hong Kong Innovation and Technology Fund under Grant ITS/145/21, and in part by City University of Hong Kong under Grant 9610675.

References

[1] X. Li, H. Ding, H. Yuan, W. Zhang, J. Pang, G. Cheng, K. Chen, Z. Liu, C.C. Loy, Transformer-based visual segmentation: A survey, IEEE Trans. Pattern Anal. Mach. Intell. (2024).

[2] H. Li, H.K. Chu, Y. Sun, Temporal consistency for RGB-thermal data-based semantic scene understanding, IEEE Robot. Autom. Lett. 9 (11) (2024) 9757–9764.

[3] Z. Feng, Y. Guo, Y. Sun, CEKD: Cross-modal edge-privileged knowledge distillation for semantic scene understanding using only thermal images, IEEE Robot. Autom. Lett. 8 (4) (2023) 2205–2212.

[4] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, R. Stiefelhagen, CMX: Cross-modal fusion for RGB-X semantic segmentation with transformers, IEEE Trans. Intell. Transp. Syst. (2023).

[5] Y. Sun, W. Zuo, M. Liu, RTFNet: RGB-thermal fusion network for semantic segmentation of urban scenes, IEEE Robot. Autom. Lett. 4 (3) (2019) 2576–2583.

[6] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, S. Gelly, Parameter-efficient transfer learning for NLP, in: International Conference on Machine Learning, PMLR, 2019, pp. 2790–2799.

[7] X.L. Li, P. Liang, Prefix-tuning: Optimizing continuous prompts for generation, in: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2021, pp. 4582–4597.

[8] E.J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022.

[9] T. Yang, Y. Zhu, Y. Xie, A. Zhang, C. Chen, M. Li, AIM: Adapting image models for efficient video understanding, in: International Conference on Learning Representations, 2023, URL https://openreview.net/forum?id=CIoSZ_HKHS7.

[10] D. Yin, Y. Yang, Z. Wang, H. Yu, K. Wei, X. Sun, 1% vs 100%: Parameter-efficient low rank adapter for dense predictions, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 20116–20126.

[11] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, P. Luo, Adaptformer: Adapting vision transformers for scalable visual recognition, Adv. Neural Inf. Process. Syst. 35 (2022) 16664–16678.

[12] S. Jie, Z.-H. Deng, Convolutional bypasses are better vision transformer adapters, 2022, arXiv preprint arXiv:2207.07039.

[13] Z. Chen, Y. Duan, W. Wang, J. He, T. Lu, J. Dai, Y. Qiao, Vision transformer adapter for dense predictions, 2022, arXiv preprint arXiv:2205.08534.

[14] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, G. Neubig, Towards a unified view of parameter-efficient transfer learning, in: International Conference on Learning Representations, 2022, URL https://openreview.net/forum?id=0RDcd5Axok.

[15] Y. Mo, X. Kang, P. Duan, B. Sun, S. Li, Attribute filter based infrared and visible image fusion, Inf. Fusion 75 (2021) 41–54.

[16] U. Shin, K. Lee, I.S. Kweon, J. Oh, Complementary random masking for rgb-thermal semantic segmentation, in: 2024 IEEE International Conference on Robotics and Automation, ICRA, IEEE, 2024, pp. 11110–11117.

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

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

[19] S. Gutiérrez, J. Fernández-Novales, T. Garde-Cerdán, S. Marín-San Román, J. Tardaguila, M.P. Diago, Multi-sensor spectral fusion to model grape composition using deep learning, Inf. Fusion 99 (2023) 101865.

[20] H. Li, Y. Sun, IGFNet: Illumination-guided fusion network for semantic scene understanding using RGB-thermal images, in: 2023 IEEE International Conference on Robotics and Biomimetics, ROBIO, IEEE, 2023, pp. 1–6.

[21] H. Li, X.-J. Wu, CrossFuse: A novel cross attention mechanism based infrared and visible image fusion approach, Inf. Fusion 103 (2024) 102147.

[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, Adv. Neural Inf. Process. Syst. 25 (2012).

[23] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.

[24] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, N. Houlsby, An image is worth 16x16 words: Transformers for image recognition at scale, in: ICLR, 2021.

[25] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, B. Guo, Swin transformer: Hierarchical vision transformer using shifted windows, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2021, pp. 10012–10022.

[26] Z. Liu, H. Hu, Y. Lin, Z. Yao, Z. Xie, Y. Wei, J. Ning, Y. Cao, Z. Zhang, L. Dong, et al., Swin transformer v2: Scaling up capacity and resolution, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2022, pp. 12009–12019.

[27] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J.M. Alvarez, P. Luo, SegFormer: Simple and efficient design for semantic segmentation with transformers, Adv. Neural Inf. Process. Syst. 34 (2021) 12077–12090.

[28] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.

[29] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR, 2015.

[30] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18, Springer, 2015, pp. 234–241.

[31] F. Yu, V. Koltun, Multi-scale context aggregation by dilated convolutions, in: ICLR, 2016.

[32] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs, IEEE Trans. Pattern Anal. Mach. Intell. 40 (4) (2017) 834–848.

[33] B. Cheng, A.G. Schwing, A. Kirillov, Per-pixel classification is not all you need for semantic segmentation, 2021,

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

[35] S. Barchid, J. Mennesson, C. Djéraba, Review on indoor RGB-D semantic segmentation with deep convolutional neural networks, in: 2021 International Conference on Content-Based Multimedia Indexing, CBMI, IEEE, 2021, pp. 1–4.
[36] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, H. Fan, Efficient multimodal semantic segmentation via dual-prompt learning, in: 2024 IEEE/RSJ International Conference on Intelligent Robots and Systems, IROS, IEEE, 2024, pp. 14196–14203.
[37] Y. Lu, S. Sirejiding, Y. Ding, C. Wang, H. Lu, Prompt guided transformer for multi-task dense prediction, IEEE Trans. Multimed. (2024).
[38] Y. Xing, J. Wang, G. Zeng, Malleable 2.5 D convolution: Learning receptive fields along the depth-axis for RGB-D scene parsing, in: European Conference on Computer Vision, Springer, 2020, pp. 555–571.
[39] P. Taghavi, R. Langari, G. Pandey, SwinMTL: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images, 2024, arXiv preprint arXiv:2403.10662.
[40] I. Lopes, T.-H. Vu, R. de Charette, Cross-task attention mechanism for dense multi-task learning, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2023, pp. 2329–2338.
[41] R. Gade, T.B. Moeslund, Thermal cameras and applications: a survey, Mach. Vis. Appl. 25 (2014) 245–262.
[42] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, S.-N. Lim, Visual prompt tuning, in: European Conference on Computer Vision, Springer, 2022, pp. 709–727.
[43] J.L. Ba, J.R. Kiros, G.E. Hinton, Layer normalization, 2016, arXiv preprint arXiv:1607.06450.
[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Adv. Neural Inf. Process. Syst. 30 (2017).
[45] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, J. Han, ABMDRNet: Adaptive-weighted bi-directional modality difference reduction network for RGB-T semantic segmentation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021, pp. 2633–2642.
[46] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, T.L. Lam, 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, 2021, pp. 4467–4473.
[47] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, T.L. Lam, Explicit attention-enhanced fusion for RGB-thermal perception tasks, IEEE Robot. Autom. Lett. (2023).
[48] G.B. Orr, K.-R. Müller, Neural Networks: Tricks of the Trade, Springer, 1998.
[49] R.K. Srivastava, K. Greff, J. Schmidhuber, Highway networks, 2015, arXiv preprint arXiv:1505.00387.
[50] X. Glorot, Y. Bengio, Understanding the difficulty of training deep feedforward neural networks, in: Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, JMLR Workshop and Conference Proceedings, 2010, pp. 249–256.
[51] M.K. Reza, A. Prater-Bennette, M.S. Asif, MMSFormer: Multimodal transformer for material and semantic segmentation, IEEE Open J. Signal Process. 5 (2024) 599–610, http://dx.doi.org/10.1109/OJSP.2024.3389812.
[52] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, X. Fan, Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 8115–8124.
[53] I. Goodfellow, D. Warde-Farley, M. Mirza, A. Courville, Y. Bengio, Maxout networks, in: International Conference on Machine Learning, PMLR, 2013, pp. 1319–1327.
[54] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al., Pytorch: An imperative style, high-performance deep learning library, Adv. Neural Inf. Process. Syst. 32 (2019).
[55] Y. Wu, A. Kirillov, F. Massa, W.-Y. Lo, R. Girshick, Detectron2, 2019, https://github.com/facebookresearch/detectron2.
[56] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, A.C. Berg, Ssd: Single shot multibox detector, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, the Netherlands, October 11–14, 2016, Proceedings, Part I 14, Springer, 2016, pp. 21–37.
[57] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, 2017, arXiv preprint arXiv:1711.05101.