CPAL: Cross-Prompting Adapter of Large Vision Foundation Model for Multi-Modal Semantic Segmentation

Ye Liu, *Member, IEEE*, Pengfei Wu, Miaohui Wang, *Senior Member, IEEE* and Jun Liu, *Senior Member, IEEE*

***Abstract*—With the development of sensor technology, various sensors are being applied to visual perception tasks, expanding the perception capabilities of traditional RGB cameras and providing richer information from heterogeneous modalities. However, how to fully exploit the information from different modalities in important tasks such as semantic segmentation remains an open and challenging issue. Existing methods typically have smaller parameter scales, which restricts the representation and generalization capabilities of the models. Training large-scale models requires massive amounts of multi-modal data, which is often difficult to obtain. However, in the RGB modality, people can access relatively vast amounts of data to train large vision foundation models (LVFM). Therefore, fine-tuning LVFMs in the RGB modality to enable them to solve multi-modal segmentation problems may be a more viable option. In this paper, instead of focusing on the ability of LVFM solely on RGB modality, we are dedicated to developing the potential of LVFM in both RGB and non-RGB modalities simultaneously, which is non-trivial due to the semantic gap between modalities. Specifically, we present a novel bi-directional cross-prompting adapter to simultaneously fully exploit the complementarity and bridging the semantic gap between modalities. We also introduce modality specific LoRA to fine-tune the foundation models of each modal. With the support of these elements, we have successfully unleashed the potential of LVFM in both RGB and non-RGB modalities simultaneously. Our method achieves state-of-the-art (SOTA) performance on five multi-modal benchmarks, including RGB+Depth, RGB+Thermal, RGB+Event, and a multi-modal video object segmentation benchmark, as well as four multi-modal salient object detection benchmarks. This demonstrates its generalization ability and robustness across diverse tasks. The code and results are available at: https://github.com/abelny56/CPAL.**

***Index Terms*—Multi-modal Semantic Segmentation, large vision foundation model, prompt-tuning, cross-prompting adapter, LoRA.**

I. Introduction

IMAGE semantic segmentation aims at categorizing each pixel into specific class, which has been applied in many domains such as robot vision and autonomous driving, and has seen significant progress [1]–[5]. However, semantic segmentation relying solely on RGB images remains susceptible to imaging conditions such as illumination variations, rainy or snowy weather, motion blur etc. Incorporating more sensors can solve these problems to a certain extent. For example, depth sensor [6]–[12] is capable of obtaining the three-dimensional geometric information of the scene, thermal camera [13]–[15] can capture the temperature information of a scene, and event camera is able to detect and output only local pixel-level brightness changes with low latency [16]–[18]. However, current multi-modal semantic segmentation methods face two major challenges:

1) At the data level, there is a shortage of large-scale multi-modal data, due to equipment limitations and the high costs associated with modal alignment and annotation.
2) At the methodological level, the challenge lies in how to maximize the extraction of information from different modalities.

Most existing methods are designed to address the second challenge by devising models with better strategies to mine useful information from different modalities [19]–[21] as

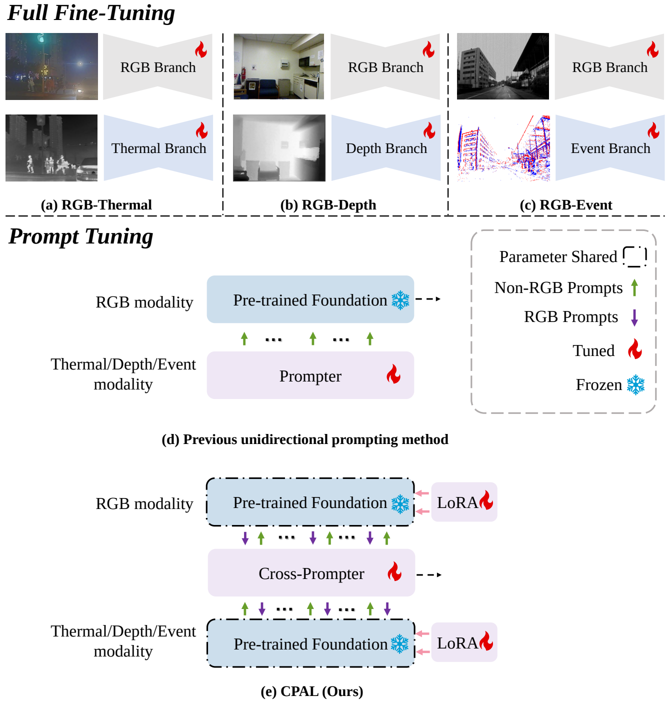


Fig. 1. Differences between our CPAL and previous multi-modal semantic segmentation paradigm. (a-c) Previous methods which extend pre-trained RGB models with fusion strategies and conduct full fine-tuning on multi-modal semantic segmentation tasks. (d) Previous unidirectional prompting method. (e) The proposed bi-directional cross-prompting tuning framework.

Ye Liu and Pengfei Wu are with School of Automation and Artificial Intelligence, Nanjing University of Posts and Telecommunications, P.R. China. (E-mails: yeliu@njupt.edu.cn, 1223055916@njupt.edu.cn)

Miaohui Wang is with Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Institute of Artificial Intelligence and Robotics for Society, and Shenzhen University, Shenzhen, P.R. China. (Email: wang.miaohui@gmail.com)

Jun Liu is with School of Computing and Communications, Lancaster University, UK. (Email: j.liu81@lancaster.ac.uk)

Corresponding author: Ye Liu.

shown in Fig. 1 (a-c). Although these methods have achieved notable progress, their performance is limited by the scale of training data and model parameters, making it difficult to further enhance the performances.

Recent advancements in large language models [22], [23] within the field of natural language processing have also spurred the development of large-scale foundational visual models. Thanks to the relatively abundant RGB data, these models demonstrate excellent generalization capabilities. However, for multi-modal vision tasks, training large-scale multi-modal foundational visual models is currently challenging due to the scarcity of multi-modal data. Therefore, fine-tuning large-scale foundational models trained on RGB modalities using limited multi-modal data is a more feasible option. Recent method utilize a unidirectional scheme: using non-RGB modalities [24] or both RGB and non-RGB modalities [25] to prompt LVFMs trained on RGB modalities as shown in Fig. 1 (d), this kind of scheme neglects the potential of LVFMs in non-RGB modalities. The large number of parameters and the scale of training data enable LVFMs to possess much stronger generalization ability compared to smaller models. Therefore, tapping into the capabilities of LVFMs in both RGB and non-RGB modalities while bridging the semantic gap between modalities are our primary design goals.

To this end, we propose a novel bi-directional multi-modal prompting learning framework named cross-prompt learning, as shown in Fig. 1 (e). In this framework, a large vision foundation model is applied simultaneously to both RGB and non-RGB modality with its parameters frozen. In this framework, a prompting network named cross-prompting adapter is proposed to prompt the frozen models of both modalities simultaneously. Specifically in the cross-prompting adapter, multiple multi-modal cross prompter (MCP) blocks act at different stages of the foundation models, which fully integrates information from both RGB and non-RGB modalities and provides prompts for the frozen models of both modalities. The staged prompt results are collected and purified through the proposed gated perception module (GPM) to generate feature maps for decoding and prediction. Additionally, to further account for modal differences, we introduce LoRA (Low-Rank Adaptation) to fine-tune the frozen models specifically for each modality to learn modality-specific representations. Excessive experiments on various types of multi-modal segmentation datasets have demonstrated the effectiveness of the proposed method.

We summarize the contributions of our work as follows:

- We propose a bi-directional cross-prompt learning framework for multi-modal semantic segmentation, which simultaneously exploits the potential of a pre-trained LVFM in both RGB and non-RGB modalities.
- Within the above framework, we propose a prompting network named cross-prompting adapter with multiple MCP and GPM blocks acting at different stages of the foundation model and producing high-quality features for decoding and prediction.
- We introduce the modality-specific LoRAs to further fine-tune large foundation RGB models for multi-modal segmentation tasks.
- We have conducted systematical experiments on five multi-modal semantic segmentation benchmarks, whose modality ranging from RGB-depth, RGB-thermal, and RGB-event. Furthermore, we extended our approach to a multi-spectral video object segmentation benchmark and four RGB-D salient object detection benchmarks, achieving state-of-the-art performance across all datasets. This validates strong generalization ability and robustness of proposed method.

II. Related Work

A. Multi-modal Semantic Segmentation

General semantic segmentation (RGB inputs) aims to predict the pixel-wise segmentation mask of input RGB images, wherein each pixel is categorized into a specific class. The continuous expansion of large-scale datasets [26]–[28], coupled with the rapid development of deep neural networks [29], [30] has propelled this field forward. Although numerous semantic segmentation models [31]–[33] have made significant breakthroughs in segmentation accuracy and precision, they encounter challenges under real-world conditions where RGB cameras fail to provide adequate information, such as strong or weak illumination, complex weather scenarios, *etc*. Accounting for this, multi-modal semantic segmentation has attracted growing attention owing to the capacity of various modalities, such as depth [21], [34], thermal [19], [35], and event [17], [36], to offer complementary information, thereby improving segmentation performance in challenging scenarios that cannot be effectively tackled solely with single-modal images. Recently, the work of [37], [38] further effectuate the shift from modality-specific fusion towards unified fusion. Mamba network is introduced as encoder and fusion module in Sigma [39] which achieves impressive performances.

Most methods above fully fine-tune models with complex and redundant fusion strategies which fail to fully leverage the advantages gained from pre-training with large amounts of data. Recently, GoPT [24] has advanced the application of Parameter-Efficient Fine-Tuning (PEFT) in multi-modal semantic segmentation by utilizing non-RGB modalities to prompt the RGB modality. DPLNet [25] takes a further step by leveraging both RGB and non-RGB modalities to prompt the RGB foundation model simultaneously. Despite the differences in prompting methods, these two approaches have only explored the potential of the foundation model within the RGB modality.

Our approach aims to leverage a cross-prompting fine-tuning method, fully exploiting the potential of pre-trained LVFMs in both RGB and non-RGB modalities. This enables the model to effectively utilize information across different modalities, achieving complementary performance in complex environments. Besides, unlike previous methods [19], [21], [34], [35] that focus on fusion with a specific modality, our designed CPAL exhibits universality and flexibility. It can adaptively adopt fusion strategies based on the perceived modality environment and utilize a gated structure to filter out noisy information.

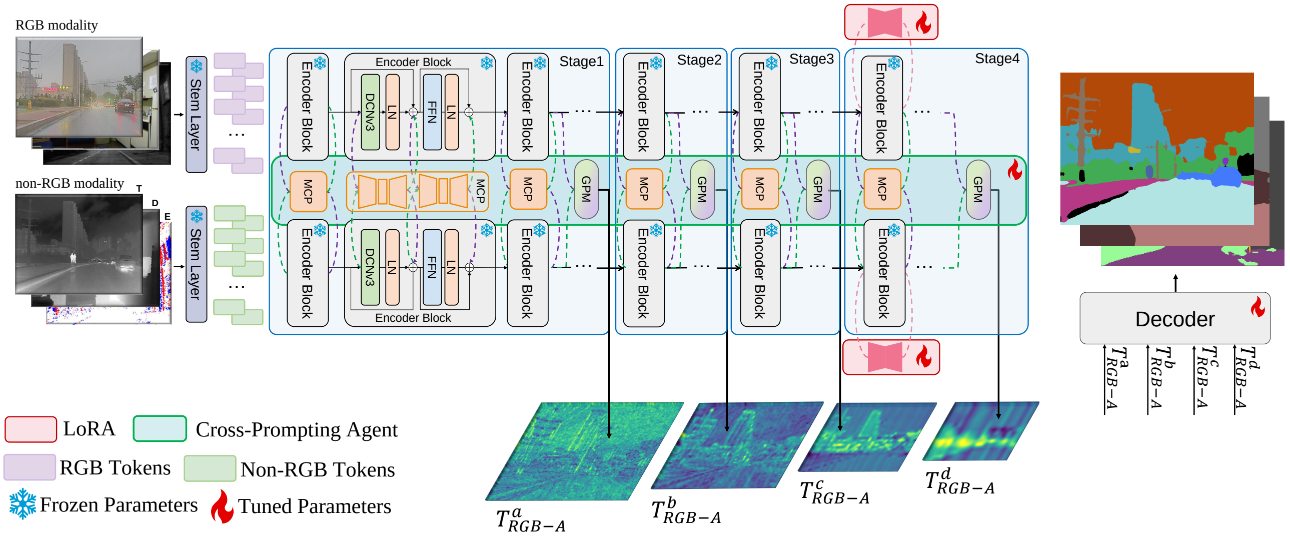


Fig. 2. Overview of CPAL for multi-modal semantic segmentation. We adapt an off-the-shelf pre-trained RGB-based foundation by incorporating a cross-prompting adapter with LoRAs for multi-modal tasks. The cross-prompting adapter is paralleled with the dual-stream encoder layer, enabling it to perceive the task environment and generate bi-directional prompts to motivate the environment for modal complementary decision-making. LoRA is embedded in each foundation model, making them learn modality specific semantics. Ultimately, the merged multi-scale staged prompted results are fed into the decoder head for mask prediction.

B. Parameter-Efficient Fine-Tuning

Pre-trained large-scale models can acquire robust feature representation capabilities from extensive data. Consequently, how to effectively utilize these large foundation models to adapt them to downstream tasks has become a significant problem [40]–[42]. Parameter-efficient tuning [43]–[46] (PEFT) is specifically designed to enhance the efficiency of fine-tuning process. Recently, prefix-tuning, a novel approach to parameter-efficient fine-tuning, has gained significant traction in many down-stream NLP tasks [44], [47]. VPT [48] introduces prompt-tuning into the vision task, which proposes to fine-tune the prompt tokens and the head to attain outstanding performance and diminish training costs. Low rank adaptation [45] (LoRA) reparameterises some certain layers in the pre-trained model, which designs the low rank approximation to fine-tune the large-scale language model.

Unlike previous approach [24] where non-RGB modalities prompt the main modality, in this work we propose the cross-prompt tuning framework and introduce LoRA tuning to multi-modal semantic segmentation. We utilize the MCP block within cross-prompting adapter to generate cross-modal prompts, further activating high-level features through the LoRA embedded in the last encoder stage.

III. Method

In this work, we propose CPAL for effectively and efficiently adapting the pre-trained large-scale RGB-based semantic segmentation model to multi-modal tasks. Instead of fully fine-tuning the whole foundation model, CPAL leverages a relatively lightweight cross-prompting adapter and low-rank adaptation (LoRA) to achieve parameter-efficient tuning, resulting in exceptional multi-modal complementarity and superior segmentation accuracy. Within this framework, the cross-prompting adapter consists of two components: multi-modal cross promoter (MCP) and gated perception module (GPM). It is worth noting that, for enhanced modality perception, all the modules within the adapter are designed with a multi-scale approach. The overall architecture of our CPAL is presented in Fig. 2.

A. Multi-modal Semantic Segmentation and Foundation Model.

Problem Definition. For RGB semantic segmentation tasks, the objective is to learn a segmentation model S_{RGB}: I_{RGB} $\rightarrow O_{Mask}$, where I_{RGB} is the input RGB image, and O_{Mask} is the predicted category of each pixel in the image. For the task of multi-modal semantic segmentation, an additional input is incorporated, expanding the model input to (I_{RGB}, I_A), with the subscript A denoting non-RGB modalities such as depth, thermal infrared, or event information. Consequently, the formulation of the multi-modal segmentation model can be expressed as S_{RGB-A} : (I_{RGB}, I_A) $\rightarrow O_{Mask}$, wherein S_{RGB-A} represents the multi-modal segmentation model.

Foundation Model. Although the PEFT methods have achieved success in language models [49], its application in multi-modal dense prediction tasks has not been fully explored. Unlike language models, for better adaptation to dense prediction tasks, we meticulously select recent LVFM, Intern-image [50], as our foundation model. It features a multi-stage encoder structure designed for semantic segmentation, capable of extracting features at different scales, with a parameter count of 1.12 billion. Generally, the foundation segmentation model S_{RGB} can be decomposed into *Enc•Dec*, where *Enc* : $I_{RGB} \rightarrow T_{RGB}$ denotes the encoder foundation, which serves as the feature extraction function, T_{RGB} represents the output features of input images, and the decoder head *Dec* : T_{RGB} $\rightarrow O_{Mask}$ is adopted to upsample the feature maps output by the encoder and achieve pixel-level prediction.

The initial step involves introducing the RGB modality, namely I_{RGB}, to the stem layer that includes a patch embedding layer and a position dropout layer, generating RGB

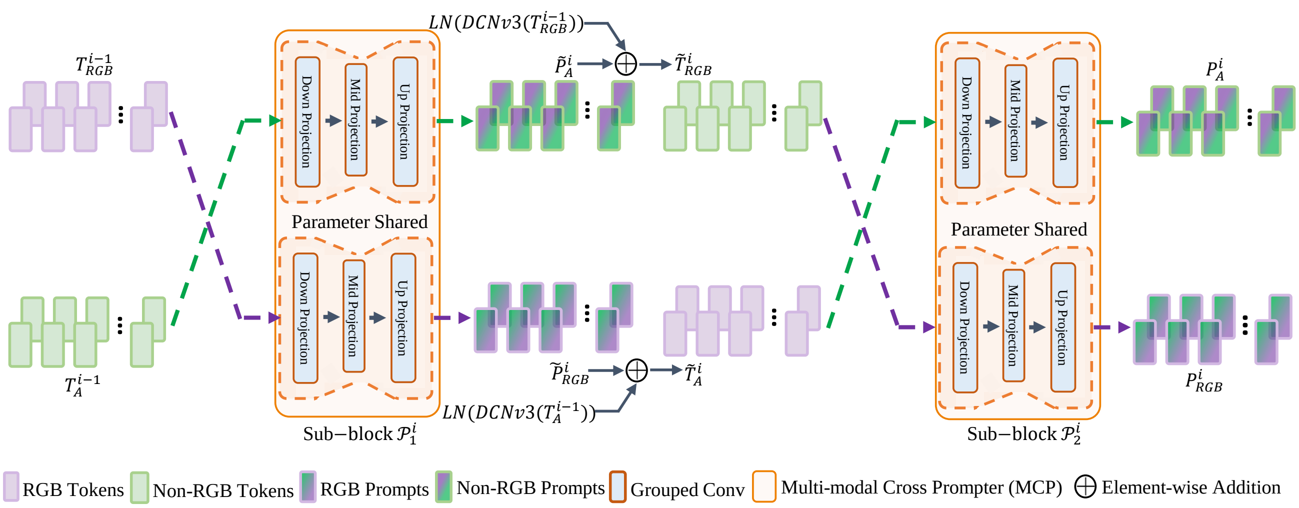


Fig. 3. Detailed design of Multi-modal Cross Prompter (MCP) which consists of two sub-blocks.

modality tokens T^0_{RGB}. Subsequently, these tokens are fed into the multi-stage encoder with l layers. In this context, we denote T^{i-1}_{RGB} as input to the i-th encoder layer E_i. Formally, the computational process of the forward operation in the i-th layer of the encoder can be represented as:

$$T^i_{RGB} = E^i(T^{i-1}_{RGB}), \quad i = 1, 2, \ldots, l \tag{1}$$

$$O_{Mask} = Dec(T^a_{RGB}, T^b_{RGB}, T^c_{RGB}, T^d_{RGB}), \tag{2}$$

where the encoder layer E_i comprises a deformable convolution v3 (DCNv3) and a feed-forward network (FFN). Each sub-layer is structured as a residual connection, followed by a normalization operation. The multi-scale feature maps (T^a_{RGB} , T^b_{RGB} , T^c_{RGB} , T^d_{RGB}) of each stage are taken as the input for the decoder head *Dec* to predict pixel-wise masks.

B. Cross-prompting Adapter with LoRAs

As illustrated in Fig. 2, our CPAL architecture features a dual-stream encoder structure catering to both the RGB modality and the supplementary modality individually. Each stream within this architecture is characterized by a shared set of frozen parameters. CPAL initially processes a pair of RGB and non-RGB input images, namely I_{RGB} and I_A, into the stem layer, obtaining the RGB tokens T^0_{RGB} and non-RGB tokens T^0_A. Then, our universal cross-prompting adapter is embedded between the dual-stream encoders. The operation of the cross-prompting adapter is divided into two phases. The first phase involves the acquisition and prompt of features from both modalities. For the i-th layer of each encoder, tokens (T^{i-1}_{RGB}, T^{i-1}_A) originating from diverse modalities are fed into the multi-modal cross prompter (MCP) to generate bi-directional cross-modal prompts (P^i_{RGB}, P^i_A). The cross-modal prompts, enriched with diverse scale information, are then added to the encoder network in a form of residual. This integration enhances the interactivity of modal information, fostering a more dynamic task environment in a systematic, layer-by-layer fashion:

$$(P^i_{RGB}, P^i_A) = \mathcal{P}^i(T^{i-1}_{RGB}, T^{i-1}_A), \quad i = 1, 2, \ldots, l \tag{3}$$

$$(T^i_{RGB}, T^i_A) = E^i(T^{i-1}_{RGB}, T^{i-1}_A) + (P^i_{RGB}, P^i_A), \tag{4}$$

where $\mathcal{P}^i$ denotes the embedded MCP block in the i-th layer of the encoder. Then, the multi-scale features from each encoder stage are fed into the gated perception module (GPM) $\mathcal{M}^i$. Leveraging the preceding prompts, this module integrates information from different modalities and filters out redundant counterparts, thereby facilitating the adapter's inferential decision-making function. Finally, the decoder head receives the multi-modal and multi-scale complementary features output by the adapter to obtain the final segmentation results.

$$T^i_{RGB-A} = \mathcal{M}^i(T^i_{RGB}, T^i_A), \quad i = a, b, c, d \tag{5}$$

$$O_{Mask} = Dec(T^a_{RGB-A}, T^b_{RGB-A}, T^c_{RGB-A}, T^d_{RGB-A}) \tag{6}$$

Multi-modal Cross Prompter (MCP). The current approaches to multi-modal semantic segmentation predominantly involve full fine-tuning, with fusion strategies between different modalities often meticulously designed specifically. These methods typically exhibit large parameter counts and lack generalizability. Recent fine-tuning method [24] has not accounted for the variations in the primary and secondary roles of different modalities. They have also failed to fully leverage the potential of pre-trained models on non-RGB modalities. To enhance the fine-tuning of frozen large pre-trained foundation model and facilitate the cross-prompting adapter's awareness of different modalities in the environment, we propose the universal MCP block, so that we can better utilize complementary information across RGB and non-RGB modalities.

As depicted in Fig. 3, the MCP employs a modular design, integrating into both the DCNv3 stage and the FFN stage individually. Each MCP block is constructed by concatenating two structurally identical sub-blocks in series. Formally, providing tokens T^{i-1}_{RGB}, T^{i-1}_A of two modalities from the i-th encoder layer equipped with a MCP module $\mathcal{P}^i$. We denote P^i_{RGB} and P^i_A as the output to $\mathcal{P}^i$, modality prompts are generated as follows:

$$\widetilde{T}^i_A = LN(DCNv3(T^{i-1}_A)) + T^{i-1}_A + \widetilde{P}^i_{RGB}, \tag{7}$$

$$\widetilde{T}^i_{RGB} = LN(DCNv3(T^{i-1}_{RGB})) + T^{i-1}_{RGB} + \widetilde{P}^i_A, \tag{8}$$

$$(\widetilde{P}^i_{RGB}, \widetilde{P}^i_A) = \mathcal{P}^i_1(T^{i-1}_{RGB}, T^{i-1}_A), \tag{9}$$

where *DCNv3* and *LN* represent the deformable convolution block and layer normalization in each encoder layer. $\mathcal{P}^i_1$ refers to the first sub-block of the i-th MCP. Additionally, $\widetilde{P}^i_{RGB}$ and $\widetilde{P}^i_A$ symbolize the feature prompts extracted from RGB and non-RGB modality. We obtain $\widetilde{T}^i_A$ and $\widetilde{T}^i_{RGB}$ by adding the prompts from two modalities to the intermediate layer results of the encoder, respectively. After that, both $\widetilde{T}^i_A$ and $\widetilde{T}^i_{RGB}$ will be fed into the feed forward layer *FFN* separately, and added with feature prompts ($\widetilde{P}^i_{RGB}$, $\widetilde{P}^i_A$) and ($\widetilde{T}^i_{RGB}$, $\widetilde{T}^i_A$) together to obtain the output (T^i_{RGB}, T^i_A) of the i-th layer,

$$T^i_A = LN(FFN(\widetilde{T}^i_A)) + \widetilde{T}^i_A + P^i_{RGB}, \tag{10}$$

$$T^i_{RGB} = LN(FFN(\widetilde{T}^i_{RGB})) + \widetilde{T}^i_{RGB} + P^i_A, \tag{11}$$

$$(P^i_{RGB}, P^i_A) = \mathcal{P}^i_2(\widetilde{T}^i_{RGB}, \widetilde{T}^i_A), \tag{12}$$

The structure of our MCP is illustrated in Fig. 3, which is designed to perceive information from two modalities and provide bi-directional cross-modal prompts for the task environment. Each MCP layer consists of two sub-blocks, each comprising three layers of projection: down-projection, mid-projection, and up-projection. Each projection layer adopts a grouped convolutional structure, which is simple yet exhibits strong representational capability while simultaneously reducing parameter count and enhancing computational speed. Tokens from each modality sequentially pass through these layers, generating modality-complementary prompts. These prompts are then added to corresponding encoder of the other modality for bi-directional cross-modal prompting.

Gated Perception Module (GPM). When the MCP is provided with cross-modal tokens through the dual-stream encoder, the cross-prompting adapter comprehensively learns features from different modalities. Subsequently, the frozen encoder produces features of various scales at each stage. Our designed GPM is seamlessly connected to the output layer of each stage, thereby playing a crucial role in the decision-making of cross-modal representations for the cross-prompting adapter.

Although the prompt generated by MCP enable the encoder to represent features from different modalities, it still contains a substantial amount of redundant and noisy information. These pieces of information may introduce interference to the crucial cross-modal interaction signals, maintaining the dominance of the original information from each modality on the final segmentation results, which is a challenge that existing modality fusion methods have not effectively addressed. To endow the cross-prompting adapter with the capability to selectively filter out redundant information and activate pertinent cross-modality signals, we introduce the GPM.

As shown in Fig. 4, we assign a GPM to receive bi-directional adaption information (T_{RGB} and T_A) jointly generated by MCP and frozen encoder. In order to enable the cross-prompting adapter to perceive various modalities of the surrounding environment and formulate corresponding fusion strategies, GPM constructs a memory vector and a forget gate to regulate the influx of noise and mismatched information.

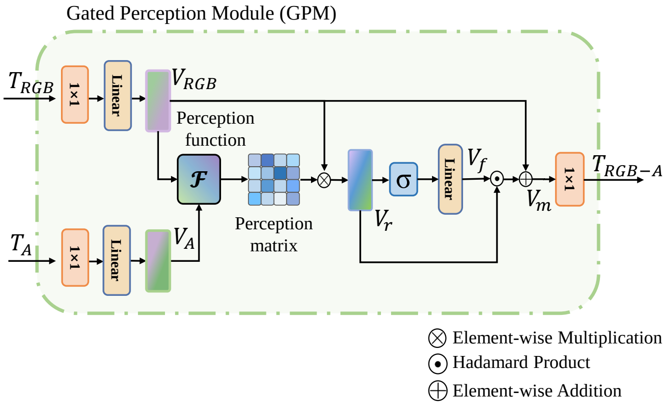


Fig. 4. Structure of the proposed Gated Perception Modul (GPM)

GPM consists of five primary stages: (i) projecting multi-modal tokens into lower-dimensional latent embeddings; (ii) computation of the perception matrix for cross fusion and remember vector; (iii) calculation of the forget vector; (iv) generation of the cross-modal merged feature; (v) projecting to original dimension. Specifically, T_{RGB} and T_A are firstly transformed with the 1×1 convolutional layer and linear layer. Then, they are fed into the perception function $\mathcal{F}$ to calculate the perception matrix m. Here, $\mathcal{F}$ signifies a simple yet effective cross-attention operation, while m represents the attention score map. Using this matrix, a modal fusion strategy is applied to derive the remember vector V_r, forget vector V_f, and merged vector V_m,

$$V_{RGB} = Linear(conv(T_{RGB})), \tag{13}$$

$$V_A = Linear(conv(T_A)), \tag{14}$$

$$V_r = \mathcal{F}(V_{RGB}, V_A) \tag{15}$$

$$V_f = \sigma(Linear(V_r)), \tag{16}$$

$$V_m = V_r \odot V_f + V_{RGB}, \tag{17}$$

where $\odot$ represents element-wise dot production. In Equation 16, we keep the raw feature V_{RGB} to prevent the forget gate from excessively filtering out valuable information. Finally, we obtain complementary cross-modal features by:

$$T_{RGB-A} = conv(V_m) \tag{18}$$

LoRA in frozen encoder. First, let's briefly recap the design of low-rank adaptation (LoRA). LoRA is a method for efficiently fine-tuning pre-trained models, leveraging a low-rank approach to modify supplementary model weights. As opposed to full fine-tuning, this approach significantly reduces the overall parameter count, and alleviates training costs.

The LoRA strategy in CPAL is illustrated in Fig. 5. Considering that the prompts from MCP for each layer of the encoder are already sufficiently informative, and the last stage contains higher-level semantic information, providing more substantial assistance to the decoder. Therefore, we exclusively apply LoRA to the core DCNv3 block of the final stage. In DCNv3, the offset layer determines the displacement of convolutional kernels across the input feature map. By adjusting these offsets, deformable convolution can capture

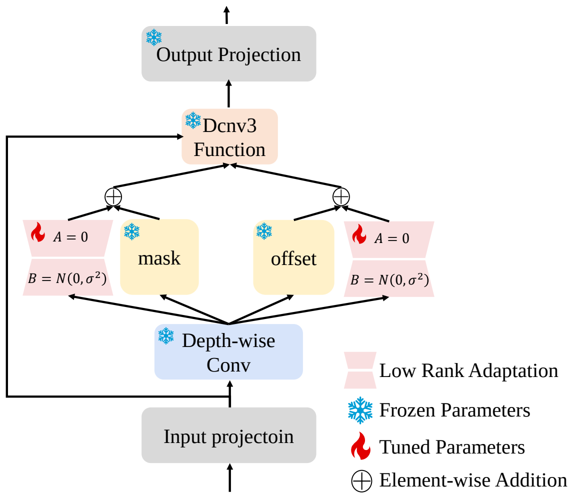


Fig. 5. Low Rank Adaptation (LoRA) in DCNv3

more flexible features, which is crucial for handling deformed objects, varying viewpoints, and complex backgrounds. The mask layer generates weights for each sampling position, which are used to weight the contributions of the sampling points. By adjusting the weights of individual sampling points, the mask layer further enhances the model's selectivity towards features, providing greater flexibility when processing diverse features. Therefore, LoRA fine-tuning is applied to these two most crucial components. Specifically, For the frozen weight matrix $W_{mask} \in \mathbb{R}^{d \times k}$ of mask layer, LoRA constrains the update during fine-tuning by representing the parameter update with a low-rank decomposition:

$$W_{mask} + \varDelta W_{mask} = W_{mask} + B_{mask} A_{mask}, \tag{19}$$

where B_{mask} and A_{mask} are learnable LoRA parameters, $B_{mask} \in \mathbb{R}^{d \times r}$, $A_{mask} \in \mathbb{R}^{r \times k}$, and the rank r $\leqslant$ min(d, k). In a similar vein, the treatment of the offset layer follows the same approach:

$$W_{offset} + \varDelta W_{offset} = W_{offset} + B_{offset} A_{offset}, \tag{20}$$

In this way, we activate deeper-level features on top of cross-modal prompts, fully leveraging the frozen large-scale foundation model and exploring information from different modalities.

IV. Experiments

A. Datasets

Following the common experiment settings of multi-modal semantic segmentation methods, we conduct extensive experiments on five widely used datasets, including RGB-depth, RGB-thermal and RGB-event.

NYU Depth V2 dataset [51] consists of 1449 RGB-depth images with the size 480×640. These images are divided into 795 training images and 654 testing images, with annotations available for 40 semantic categories.

SUN-RGBD dataset [52] comprises 10335 RGB-depth images across 37 classes. For our experiment, we adhere to the same training/test split as outlined in [52].

FMB dataset [53] presents a novel and formidable collection comprising 1500 pairs of meticulously calibrated RGB-thermal image pairs. Within this dataset, the training set encompasses 1220 image pairs, and the test set comprises 280 pairs. Notably, the dataset covers a diverse array of challenging scenes, including those with the Tyndall effect, rain, fog, and intense light.

PST900 dataset [14] provides 894 RGB-thermal images captured at a resolution of 720×1280 in cave and subterranean environments for the DARPA Subterranean Challenge. This dataset includes annotated segmentation labels for five classes, encompassing one background class (unlabeled) and four object classes.

DDD17 dataset [16], [54] incorporates more than 12 hours of DAVIS sensor recordings, each with a resolution of 260×346 pixels, capturing diverse scenarios of both highway and city driving.

B. Implementation Details

We conducted all experiments on a single NVIDIA 3090 GPU with a global batch size of 2 and iteration of 40k. The AdamW optimizer [84] with a weight decay of 0.05 is adopted. The initial learning rate is set to 2×10^{-5} and decayed following the polynomial decay schedule with a power of 1.0. The fixed parameters in our model are initialized by the pre-trained foundation. Additionally, we utilize the same loss function as [50]. Some recent works [37], [56], [58], [67], [69] employ multi-scale inference strategy for data augmentation. While in our work, to showcase the effectiveness of CPAL, we only adopt the single scale test. Following existing methods, we report the popular mean intersection over union (mIoU), pixel accuracy (pAcc) and mean accuracy (mAcc) as the primary evaluation metrics to measure the segmentation performance. Additionally, we selected two common decoders, namely Uper Head [85] and a more lightweight Hamburger Head [86], to serve as our decoder heads, resulting in our model CPAL-L and CPAL-T, respectively.

C. Comparison with State-of-the-arts

Results on RGB-depth Datasets. We conduct a comparative analysis between our CPAL and 22 contemporary RGB-depth semantic segmentation methods using NYUDepthv2 dataset [51] and SUN-RGBD dataset [52]. As indicated in Table I, our method, even with single-scale inference, outperforms previous state-of-the-art methods by a large margin. The most recent fine-tuning method GoPT [24] utilizes a frozen pre-trained model to extract RGB features, fine-tunes non-RGB branches, and employs non-RGB modalities to prompt the RGB modality, overlooking the pre-trained model's capability to extract features from the non-RGB modality. Dformer [67] initially undergoes pre-training on RGB-D datasets, followed by a comprehensive full fine-tuning process, which appears to be intricate. In contrast, our approach directly utilizes a pre-trained RGB-based model, incorporating a lightweight cross-prompting adapter with LoRA that seamlessly embeds cross-modal prompts from both RGB and depth modality. By incorporating the adapter's modality perception along with

TABLE I

Results on NYU Depth V2 [51] and SUN-RGBD [52]. The top three results in each column are highlighted in **RED**, **BLUE** and **GREEN**, ♠ represents the method of fine-tuning pre-trained model.

Model	Publication	NYUDepthv2			SUN-RGBD			Parameters	
		Input size	pAcc(%)	mIoU(%)	Input size	pAcc(%)	mIoU(%)	Trainable	Inference
ACNet [55]	2019-ICIP	480 × 640	—	48.3	530 × 730	—	48.1	116.6M	116.6M
SA-Gate [56]	2020-ECCV	480 × 640	77.9	52.4	530 × 730	82.5	49.4	110.9M	116.6M
CEN [57]	2020-NeurIPS	480 × 640	77.7	52.5	530 × 730	83.5	51.1	133.9M	133.9M
SGNet [58]	2021-TIP	480 × 640	76.8	51.1	530 × 730	82.0	48.6	64.7M	64.7M
ShapeConv [21]	2021-ICCV	480 × 640	76.4	51.3	530 × 730	82.2	48.6	86.8M	86.8M
ESANet [59]	2021-ICRA	480 × 640	–	50.3	480 × 640	–	48.2	31.2M	31.2M
FRNet [60]	2022-JSTSP	480 × 640	–	53.6	530 × 730	–	51.8	85.5M	85.5M
PGDENet [61]	2022-TMM	480 × 640	–	53.7	530 × 730	–	51.0	100.7M	100.7M
EMSANet [62]	2022-IJCNN	480 × 640	–	51.0	–	–	–	46.9M	46.9M
TokenFusion [63]	2022-CVPR	480 × 640	79.0	54.2	530 × 730	84.7	53.0	45.9M	45.9M
MultiMAE [64]	2022-ECCV	640 × 640	–	56.0	640 × 640	–	51.1	95.2M	95.2M
Omnivore [34]	2022-CVPR	480 × 640	–	54.0	–	–	–	95.7M	95.7M
NAS [65]	2023-ACMMM	–	79.4	55.1	–	82.9	50.3	48.9M	48.9M
PDCNet [66]	2023-TCSVT	480 × 480	78.4	53.5	480 × 480	83.3	49.6	–	–
CMX [37]	2023-TITS	480 × 640	**80.1**	56.9	530 × 730	83.8	52.4	181.1M	181.1M
CMNext [38]	2023-CVPR	480 × 640	–	56.9	530 × 730	–	51.9	119.6M	119.6M
DFormer [67]	2024-ICLR	480 × 640	–	57.2	530 × 730	–	52.5	39.0M	39.0M
OmniVec [68]	2024-WACV	480 × 640	–	60.8	–	–	–	95.7M	95.7M
GeminiFusion [69]	2024-ICML	480 × 640	–	**60.9**	530 × 730	–	52.5	–	–
Sigma [39]	2024-arXiv	480 × 640	–	57.0	480 × 640	–	52.4	69.8M	69.8M
DPLNet♠ [25]	2023-arXiv	480 × 640	–	59.3	530 × 730	–	**52.8**	7.15M	88.6M
GoPT♠ [24]	2024-AAAI	480 × 640	**80.1**	54.3	530 × 730	**85.5**	52.3	0.97M	112.6M
CPAL-T♠(Ours)	2024	480 × 640	**84.6**	**66.2**	530 × 730	**85.9**	**56.4**	6.2M	1.1B
CPAL-L♠(Ours)	2024	480 × 640	**84.5**	**65.8**	530 × 730	**86.3**	**58.2**	51.2M	1.1B

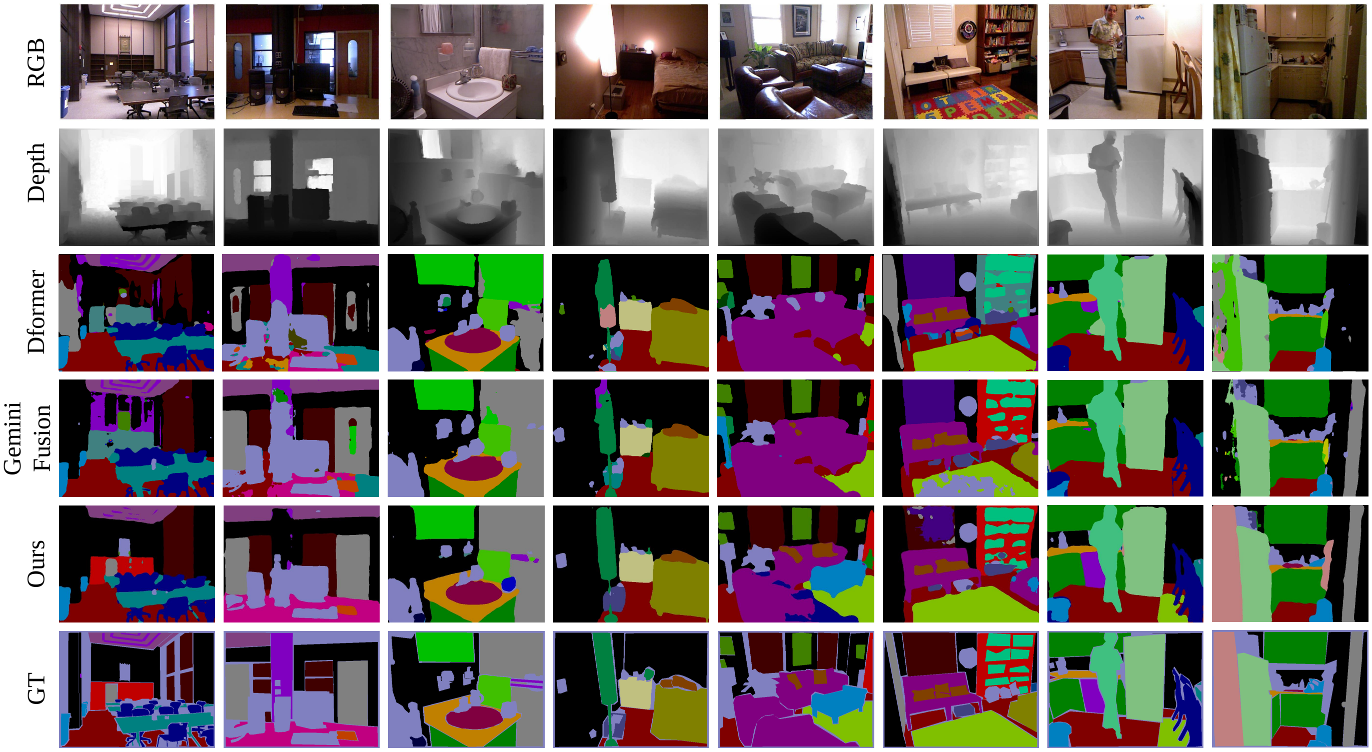


Fig. 6. Qualitative visual comparisons between our method (CPAL-L) and the state-of-the-art networks on the NYU Depth V2 test set.

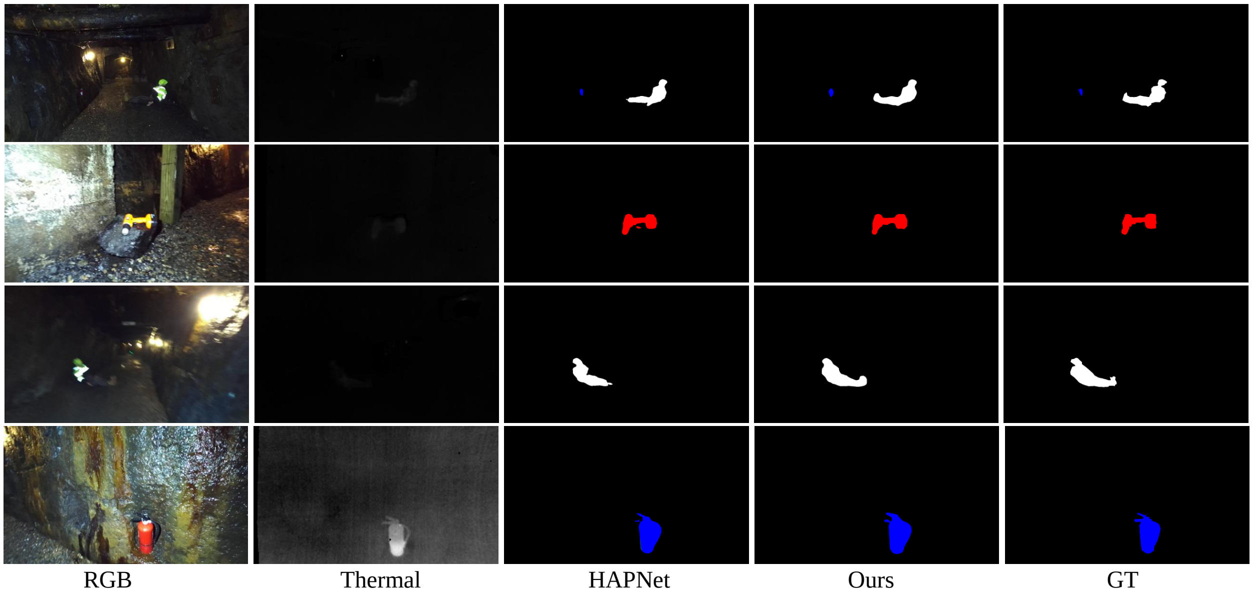


Fig. 7. Visual comparisons between our CPAL-L and the state-of-the-art method (HAPNet) on the PST900 dataset. Please zoom-in for the best view.

TABLE II
OVERALL PERFORMANCE ON FMB [53].

Model	Publication	mAcc(%)	mIoU(%)	Parameters Trainable	Parameters Inference
DIDFuse [70]	2020-IJCAI	73.0	50.6	–	–
GMNet [71]	2021-TIP	64.4	49.2	149.8M	149.8M
FEANet [72]	2021-IROS	64.5	46.8	255.2M	255.2M
ReCoNet [73]	2022-ECCV	71.4	50.9	–	–
U2Fusion [74]	2022-TPAMI	70.1	47.9	–	–
TarDAL [75]	2022-CVPR	74.8	48.1	–	–
LASNet [76]	2023-TCSVT	56.9	42.5	93.6M	93.6M
EGFNet [77]	2023-AAAI	63.0	47.3	201.3M	201.3M
SegMiF [53]	2023-ICCV	**74.5**	**54.8**	–	–
CPAL-T(Ours)	2024	**74.6**	**64.2**	6.2M	1.1B
CPAL-L(Ours)	2024	**76.2**	**64.6**	51.2M	1.1B

cross-modal prompts, our method efficiently harnesses the latent potential of the frozen large-scale foundational model, streamlining the training pipeline. In terms of the mIoU, CPAL-L outperforms the second-ranked method by 4.9% on NYUDepthv2, and by 5.7% on SUN-RGBD, respectively. Moreover, surprisingly, our CPAL-T model achieved superior performance on the NYUDepthv2 dataset with fewer training parameters. Qualitative visual comparisons with state-of-the-art networks on the NYU Depth V2 dataset are illustrated in Fig. 6.

Results on RGB-thermal Datasets. For RGB-T semantic segmentation, we conduct experiments on PST900 [14] and FMB [53]. The quantitative performance of CPAL and the compared methods on the FMB dataset is reported in Table II. CPAL-L outperforms previous SOTA algorithms by a substantial margin, surpassing the SegMiF [53] by 8.48% mIoU. This notable performance differential underscores the effectiveness of CPAL in leveraging thermal information, showcasing its adaptability in challenging scenarios, such as night-time conditions. Table III compares our CPAL with 15 SOTA RGB-thermal semantic segmentation models on the PST900. Our method and HAPNet [83] both achieve a tie for first place with an mIoU of 89.0%. As can be seen from the visualization results in Fig. 7, our model does a better job of handling details and has superior segmentation performance. Furthermore, our approach significantly outperforms the recent fine-tuning method GoPT [24], demonstrating its superior ability to address the shortcomings of previous technique.

Results on RGB-E Datasets. We compare our methods with cutting-edge event-based semantic segmentation methods, including Spiking-Deeplab/FCN [87], Ev-SegNet [16], ESS [36], Evdistill [18], SSAM [88], CMX [37] and CMNeXt [38], as shown in Table IV. It is noteworthy that we just utilize event images derived from raw event data rather than event stream, our CPAL-L achieves remarkable 77.42% mIoU, surpassing the second-best model CMNeXt [88] by 4.75%. In addition, the dataset collected are grayscale images rather than RGB images. This indicates that, despite the absence of color information, our CPAL fine-tuning process has enabled the pre-trained RGB-based model to exhibit generalization capabilities even on grayscale images.

D. Generalization to Other Multi-modal Task

To validate the generalization capability of our approach, we conducted experiments on a multi-spectral video object segmentation benchmark MVseg [91] and four RGB-D salient object detection benchmarks NJU2K [97], NLPR [98], DES [99] and SIP [100].

Multi-spectral Video Object Segmentation. We employed the same experimental setup as described in subsection B. The results, as depicted in Table V, demonstrate that although we did not utilize temporal information from consecutive frames and only performed semantic segmentation on individual images, our fine-tuning method achieved remarkable performance, surpassing the second-best result by 2.29% in mIoU.

TABLE III
Quantitative comparisons on PST900 [14] dataset, results are reported in percentage (%). The top three results in each column are highlighted in **red**, **blue** and **green**, ♠ represents the method of fine-tuning pre-trained model.

Methods	Publication	Background		Hand-Drill		Backpack		Fire-Extinguisher		Survivor		mAcc(%)	mIoU(%)	**Parameters**	
		Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU	Acc	IoU			Trainable	Inference
MFNet [13]	2017-IROS	-	98.6	-	41.1	-	64.2	-	60.4	-	20.7	-	57.0	8.4M	8.4M
PSTNet [14]	2020-ICRA	-	98.9	-	53.6	-	69.2	-	70.1	-	50.0	-	68.3	105.8M	105.8M
EGFNet [77]	2022-AAAI	99.5	99.3	98.0	64.7	94.2	83.05	95.2	71.3	83.3	74.3	94.0	78.5	201.3M	201.3M
MTANet [78]	2022-TIV	-	99.3	-	62.1	-	87.5	-	65.0	-	79.1	-	78.6	121.9M	121.9M
MFFENet [79]	2022-TIP	-	99.4	-	72.50	-	81.0	-	66.4	-	75.6	-	79.0	–	–
GMNet [71]	2021-TIP	99.8	99.4	90.3	85.2	89.0	83.8	88.3	73.8	80.9	78.4	89.6	84.1	149.8M	149.8M
DSGBINet [80]	2022-TCSVT	99.7	99.4	94.5	75.0	88.7	85.1	94.8	79.3	81.4	75.6	91.8	82.9	–	–
LASNet [76]	2023-TCSVT	99.8	99.5	91.8	82.8	90.8	86.5	92.4	77.8	83.4	75.5	91.6	84.4	–	–
CAINet [19]	2023-TMM	99.7	99.5	95.9	80.3	96.1	88.0	88.4	77.2	91.4	78.7	**94.3**	84.7	12.16M	12.16M
FDCNet [81]	2023-TCSVT	99.7	99.2	82.5	70.4	77.5	72.2	91.8	71.5	78.4	72.4	85.96	77.1	–	–
SGFNet [82]	2023-TCSVT	99.8	99.4	94.0	76.7	90.4	85.4	89.4	75.6	82.7	76.7	91.2	82.8	–	–
EAEFNet [15]	2023-RAL	99.8	99.5	93.0	83.9	91.0	87.7	92.2	80.4	79.3	75.6	91.1	**85.4**	–	–
HAPNet [83]	2024-arXiv	99.8	99.6	95.5	89.3	95.1	92.0	93.9	81.3	85.6	82.4	**94.0**	**89.0**	–	–
DPLNet♠ [25]	2023-arXiv	–	–	–	–	–	–	–	–	–	–	–	86.7	7.15M	88.6M
GoPT♠ [24]	2024-AAAI	–	–	–	–	–	–	–	–	–	–	–	81.5	0.97M	112.6M
CPAL-T (Ours) ♠	2024	99.8	99.6	93.9	85.1	91.2	88.7	90.1	81.9	84.8	81.3	92.0	**87.3**	6.2M	1.1B
CPAL-L (Ours) ♠	2024	99.8	99.6	94.2	87.2	93.6	90.8	93.7	83.1	89.3	84.1	**94.1**	**89.0**	51.2M	1.1B

TABLE IV
Results on the DDD17 [16], [54] dataset. E denotes events, G denotes grayscale images.

Model	**Type**	**Input**	**mIoU(%)**	**Parameters**	
				Trainable	**Inference**
Evdistill [18]	ANN	E	58.02	–	5.8M
ESS [36]	ANN	E	61.37	–	6.7M
Ev-SegNet [16]	ANN	E	54.81	29.1M	29.1M
Spiking-DeepLab [87]	SNN	E	33.7	4.1M	4.1M
Spiking-FCN [87]	SNN	E	34.2	13.6M	13.6M
SSAM [88]	SNN	E	53.15	8.6M	8.6M
ESS [36]	ANN	E+G	60.43	–	6.7M
HALSIE [89]	ANN+SNN	E+G	60.66	1.8M	1.8M
EDCNet-S2D [90]	ANN	E+G	61.99	17.0M	17.0M
Ev-SegNet [16]	ANN	E+G	68.36	29.1M	29.1M
SSAM [88]	SNN	E+G	72.57	8.6M	8.6M
CMX [37]	ANN	E+G	71.88	181.1M	181.1M
CMNeXt [38]	ANN	E+G	**72.67**	119.6M	119.6M
CPAL-T(Ours)	ANN	E+G	**76.39**	6.2M	1.1B
CPAL-L(Ours)	ANN	E+G	**77.42**	51.2M	1.1B

TABLE V
Video object segmentation results on the MVseg [91] dataset .

Model	**Parameters**		**mIoU(%)**
	Trainable	Inference	
CCNet [92]	–	–	51.70
OCRNet [93]	–	–	52.38
STM [94]	–	–	52.51
LMANet [95]	–	–	52.73
MFNet [13]	8.4M	8.4M	51.63
RTFNet [96]	337.1M	337.1M	52.77
EGFNet [77]	201.3M	201.3M	53.44
MVNet [91]	88.4M	88.4M	**54.52**
CPAL-T(Ours)	6.2M	1.1B	**55.39**
CPAL-L(Ours)	51.2M	1.1B	**56.81**

RGB-D Salient Object Detection. In order to evaluate the model's robustness to domain shift, we fine-tune and test CPAL-T on four popular RGB-D salient object detection datasets. The finetuning dataset consists of 2,195 samples, where 1,485 are from NJU2K-train [97] and the other 700 samples are from NLPR-train [98]. Our model is evaluated on four datasets, i.e., DES [99], NLPR-test [98], NJU2K-test [97], and SIP [100]. For performance evaluation, we adopt four golden metrics of this task, i.e., Structure-measure (S) [106], mean absolute error (M) [107], max F-measure (F) [108], and max E-measure (E) [109]. To adapt the semantic segmentation model for salient object detection tasks, we adopt the same experimental configuration as DPLNet [25] and set the output channel of the decoder head to 1. As shown in Table VI, our method ranks first across most metrics, demonstrating its transferability and ability to handle domain shifts.

E. Ablation Study

To verify the effectiveness of our CPAL, we employ our CPAL-L model to conduct a detailed ablation study across three modalities, i.e., RGB-T, RGB-D, and RGB-E, corresponding to the FMB, SUN-RGBD, and DDD17 datasets.

1) Component Analysis: Our CPAL is composed of cross-prompting adapter (MCP + GPM) and low rank adaptation (LoRA). To better understand the impact and contribution of each component, We take out these three components from CPAL respectively. As shown in Table VII, (1) denotes RGB-based foundation without thermal branch, (2) denotes the model that adapts the RGB-based model to RGB-thermal segmentation by MCP, incorporating additive feature fusion at each stage. Compared with (2), the enhancement observed in (2) demonstrates the efficacy of cross-modal prompts. (3) indicates the dual-stream model with GPM. (4) represents the method for fine-tuning the RGB foundation through LoRA. (5) denotes that LoRA is inserted into the fourth stage for fine-tuning, based on (2). (6) denotes the model with LoRA removed. In comparison with (2), there is an increase of 1.41%

TABLE VI
RESULTS AND COMPARISON ON RGB-D SOD BENCHMARKS. ↑/↓ INDICATES THAT A LARGER/SMALLER VALUE IS BETTER.

Model	NJU2K [97]				NLPR [98]				DES [99]				SIP [100]			
	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$
CMWNet [101]	.903	.912	.880	.046	.917	.951	.872	.029	.933	.967	.899	.022	.868	.907	.851	.062
cmWS [102]	.900	.914	.886	.044	.915	.945	.870	.027	-	-	-	-	-	-	-	-
SSF [103]	.898	.912	.885	.043	.913	.949	.875	.026	.903	.946	.882	.026	-	-	-	-
BBSNet [104]	.912	.919	.893	.040	.920	.945	.870	.027	.906	.941	.866	.029	.871	.909	.850	.057
LSNet [105]	.911	.922	.900	.037	.918	.956	.885	.024	.925	.970	.910	.020	.886	.927	.884	.048
DPLNet [25]	.920	.944	.904	.035	.933	.962	.897	.020	.940	.978	.921	.017	.890	.932	.888	.045
CPAL-T (Ours)	.922	.942	.931	.033	.932	.954	.933	.023	.947	.973	.949	.014	.905	.928	.923	.039

TABLE VII
COMPONENT ANALYSIS ON FMB/SUN-RGBD/DDD17 DATASET.

Model	**MCP**	**GPM**	**LoRA**	**mIoU(%)**		
				RGB-T	RGB-D	RGB-E
(1)				58.86	52.91	72.39
(2)	✓			62.52	56.68	75.94
(3)		✓		61.49	53.16	74.58
(4)			✓	59.12	54.33	74.19
(5)	✓		✓	63.10	57.31	77.31
(6)	✓	✓		63.92	58.02	77.39
FFT	✓	✓	✓	63.12	58.14	76.98
CPAL	✓	✓	✓	**64.69**	**58.23**	**77.42**

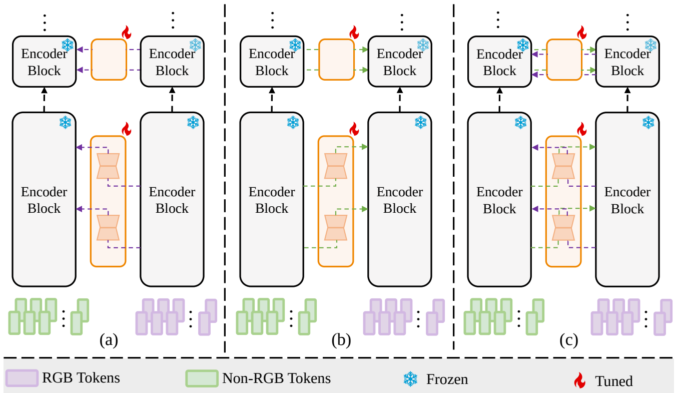


Fig. 8. Different variants of multi-modal cross prompter for dual-stream encoder framework.

TABLE VIII
RESULTS OF DIFFERENT MULTI-MODAL CROSS PROMPTER (MCP) VARIANTS IN FMB/SUN-RGBD/DDD17 DATASET. ◄ DENOTES A VARIANT OF THE MCP FOR UNIDIRECTIONAL PROMPT FROM RGB TO X MODALITY (THERMAL/DEPTH/EVENT), WHILE ► DENOTES A VARIANT OF THE MCP FOR UNIDIRECTIONAL PROMPT FROM X MODALITY TO RGB MODALITY, AND ◄► REPRESENTS OUR CROSS-PROMPTING METHOD.

Model	**MCP**	**GPM**	**LoRA**	**mIoU(%)**		
				RGB-T	RGB-D	RGB-E
RGB-only				58.86	52.91	72.39
X-only				54.23	47.61	63.87
ⓐ RGB+X	◄	✓	✓	63.57	56.12	73.61
ⓑ RGB+X	►	✓	✓	64.03	57.89	76.88
ⓒ RGB+X	◄►	✓	✓	**64.69**	**58.23**	**77.42**
ⓓ RGB+RGB	◄►	✓	✓	58.87	52.73	72.27
ⓔ X+X	◄►	✓	✓	54.19	47.61	63.91

mIoU in (6), highlighting the crucial role played by the GPM within cross-prompting adapter in modality perception and decision-making. After that, the final model (CPAL) achieves 64.69% mIoU, indicating that leveraging the cross-prompting adapter allows LoRA to activate deeper and higher-level information. Additionally, we employed the full fine-tuning (FFT) method, training all parameters. The results indicated that the large number of parameters in the LVFM constrained the effectiveness of FFT, thereby validating the correctness of our CPAL. We observe that the improvement brought by LoRA here is relatively small compared with MCP and GPM. The possible reason is that MCP and GPM have already fully tapped the potential of the foundation model, leaving limited room for further improvement, which is then exploited by LoRA.

2) The Effectiveness of Multi-modal Cross Prompter (MCP): To evaluate the effectiveness of MCP and modality complementarity, we present the foundational results with only RGB/Thermal input and three variants of MCP in Table VIII. Models ⓐ, ⓑ, and ⓒ correspond to (a), (b), and (c) in Fig. 8, respectively. ⓒ represents the cross-prompting strategy employed in our model. Results from using only RGB or thermal modalities reveal the inadequacy of single-modal information. Besides, using a single modality to prompt another modality proves to be ineffective. This approach fails to account for the dynamic relationships between different modalities. Additionally, methods that use a pre-trained model to extract features from only one modality do not fully exploit the potential of the pre-trained model. In contrast, our proposed cross-prompting strategy applies the pre-trained model to both modality branches, effectively leveraging complementary information between modalities.

3) Inserting Stages of CPAL Block: We experimentally investigate the effect of inserting stages of each component block in CPAL and summarize the results in Table IX, which shows the parameter quantities of each module and mIoU scores on the FMB dataset. The results from both MCP and GPM indicate that as the number of inserted stages increases, the mIoU rises. In contrast, the results from LoRA demonstrate that embedding only into the fourth stage not only reduces

TABLE IX
ABLATION ON THE NUMBER OF EACH COMPONENT BLOCK.

Stage	MCP		GPM		LoRA	
	Param	mIoU	Param	mIoU	Param	mIoU
4	0.15M	63.39%	2.62M	63.56%	0.34M	64.69%
3-4	0.56M	64.42%	3.28M	63.92%	1.28M	64.53%
2-4	0.60M	64.51%	3.48M	64.21%	1.36M	64.42%
1-4	0.62M	64.69%	3.62M	64.69%	1.41M	64.67%

TABLE X
AN ABLATION STUDY (%) ON DIFFERENT FOUNDATIONS.

Foundation	ViT-B	ViT-L	ViT-H	MiT-B5	Internimage(Ours)
Base Result	58.09	58.12	58.85	57.65	58.86
Final Result	62.65	62.73	62.98	61.10	64.69

TABLE XI
AN ABLATION STUDY (%) ON DIFFERENT ADAPTION METHODS.

Adapter	Series Adapter [110]	Parallel Adapter [111]	LoRA(Ours)
Result	64.21	64.13	64.69

TABLE XII
AN ABLATION STUDY (%) ON DIFFERENT PROMPT-TUNING METHODS.

Method	VPT [48]	Prefix Tuning [112]	Cross-Prompt Tuning(Ours)
Result	63.31	63.35	64.69

parameter count but also yields optimal performance. This validates the notion that the high-level semantic information in the last stage of the encoder is the most beneficial.

4) Different Foundation and Fine-tuning Method: To further validate the effectiveness of our method, we conducted ablation experiments on the foundation, adapter, and prompt tuning methods using the FMB dataset, as shown in Table X, Table XI and Table XII. The results validate the effectiveness of our method. Table XI presents a comparison of three adapter methods. The Series Adapter [110] sequentially integrates additional adapter layers after the original encoder layers, facilitating fine-tuning while preserving the original parameters. In contrast, the Parallel Adapter [111] connects multiple adapter modules in parallel to the original layers, enabling independent feature extraction and aggregation. Additionally, LoRA utilizes low-rank factorization of weight matrices, allowing for efficient fine-tuning by updating only a limited number of low-rank parameters, making it more suitable for resource-constrained scenarios compared to other methods.

5) Visualization Results: To further investigate the role of CPAL in modality complementarity, we visualiz the segmentation results and feature maps, as depicted in Fig. 9 and 10 respectively. Specifically, we select two pairs of RGB-thermal images in low-light environment and visualize the feature maps. We compare the feature maps of the RGB-only foundation with those augmented by CPAL. The results indicate that the latter's feature maps, integrating both RGB and thermal information, exhibit clearer contours and richer textures.

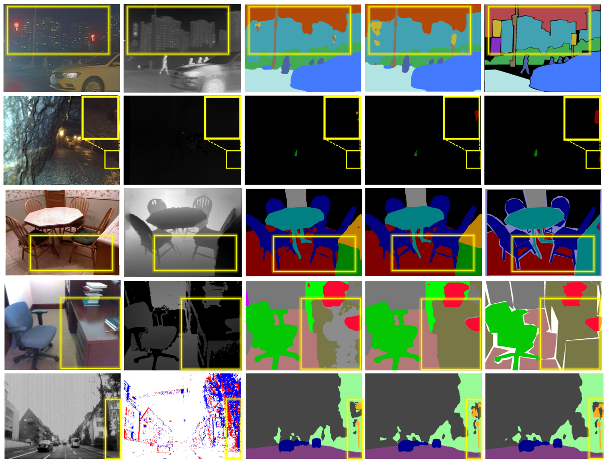

(a) RGB Modal (b) Non-RGB Modal (c) Foundation Results (d) CPAL Results (e) Ground Truth

Fig. 9. Visualization results of foundation and our CPAL-L. From top to bottom: FMB, PST900, NYU Depth V2, SUN-RGBD and DDD17 semantic segmentation.

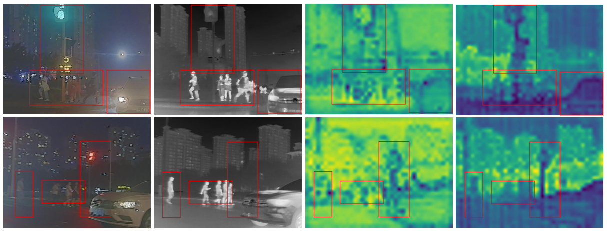

(a) RGB (b) Thermal (c) Feature Map of Foundation (d) Feature Map of CPAL

Fig. 10. Visualization of the features of foundation and our CPAL-L.

F. Discussion on Limitations

In this paper, we have successfully explored the potential of LVFM in multi-modal semantic segmentation, achieving state-of-the-art performance. Furthermore, our method requires at least 6.26M trainable parameters which is only 0.58% of the amount of parameters (1.12B) in LVFM, demonstrating its a parameter-efficient fine-tuning approach. However, due to the adoption of a large foundation model, our approach has a relatively large number of parameters for inference and a longer inference time compared to previous methods, which may not be as favorable for edge computing or applications requiring high inference speeds. Our future work will focus on adopting techniques such as knowledge distillation to lighten the model while maintaining its performance.

V. CONCLUSION

In this work, we introduce large-scale pre-trained RGB model into multi-modal dense prediction tasks and propose a novel general multi-modal parameter efficient fine-tuning paradigm: CPAL. CPAL integrates LoRA tuning and bi-directional cross-prompt tuning, tailored for enhanced adaptation to multi-modal tasks. This approach fully leverages the potential of pre-trained LVFM in both RGB and non-RGB branches, as well as the complementary information between modalities. Extensive experimentation validates the effectiveness and generalization of our approach. We expect this work can attract more attention to prompt tuning of LVFM for multi-modal semantic segmentation.

References

[1] J. Ji, R. Shi, S. Li, P. Chen, and Q. Miao, “Encoder-decoder with cascaded crfs for semantic segmentation,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1926–1938, 2020.

[2] W. Kim and J. Seok, “Indoor semantic segmentation for robot navigating on mobile,” in *Proc. Int. Conf. Ubiquitous Future Networks (ICUFN)*. IEEE, Jul. 2018, pp. 22–25.

[3] B. Pan, J. Sun, H. Y. T. Leung, A. Andonian, and B. Zhou, “Cross-view semantic segmentation for sensing surroundings,” *IEEE Robot. Autom. Lett.*, vol. 5, no. 3, pp. 4867–4873, Jun. 2020.

[4] M. Siam, M. Gamal, M. Abdel-Razek, S. Yogamani, M. Jagersand, and H. Zhang, “A comparative study of real-time semantic segmentation for autonomous driving,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2018, pp. 587–597.

[5] D. Feng, C. Haase-Schütz, L. Rosenbaum, H. Hertlein, C. Glaeser, F. Timm, W. Wiesbeck, and K. Dietmayer, “Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges,” *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2020.

[6] W. Shi, J. Xu, D. Zhu, G. Zhang, X. Wang, J. Li, and X. Zhang, “Rgb-d semantic segmentation and label-oriented voxelgrid fusion for accurate 3d semantic mapping,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 1, pp. 183–197, 2021.

[7] H. Sheng, R. Cong, D. Yang, R. Chen, S. Wang, and Z. Cui, “Urbanlf: A comprehensive light field dataset for semantic segmentation of urban scenes,” *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7880–7893, 2022.

[8] W. Zhou, B. Jian, M. Fang, X. Dong, Y. Liu, and Q. Jiang, “Dgpinet-kd: Deep guided and progressive integration network with knowledge distillation for rgb-d indoor scene analysis,” *IEEE Trans. Circuits Syst. Video Technol.*, 2024.

[9] D. Lin, G. Chen, D. Cohen-Or, P.-A. Heng, and H. Huang, “Cascaded feature network for semantic segmentation of rgb-d images,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 1311–1319.

[10] S.-J. Park, K.-S. Hong, and S. Lee, “Rdfnet: Rgb-d multi-level residual feature fusion for indoor semantic segmentation,” in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4980–4989.

[11] J. Jiang, L. Zheng, F. Luo, and Z. Zhang, “Rednet: Residual encoder-decoder network for indoor rgb-d semantic segmentation,” 2018, arXiv:1806.01054.

[12] E. Yang, W. Zhou, X. Qian, J. Lei, and L. Yu, “Drnet: Dual-stage refinement network with boundary inference for rgb-d semantic segmentation of indoor scenes,” *Eng. Appl. Artif. Intell.*, vol. 125, p. 106729, Jul. 2023.

[13] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, “Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes,” in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Systems (IROS)*. IEEE, Sep. 2017, pp. 5108–5115.

[14] S. S. Shivakumar, N. Rodrigues, A. Zhou, I. D. Miller, V. Kumar, and C. J. Taylor, “Pst900: Rgb-thermal calibration, dataset and segmentation network,” in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, May 2020, pp. 9441–9447.

[15] M. Liang, J. Hu, C. Bao, H. Feng, F. Deng, and T. L. Lam, “Explicit attention-enhanced fusion for rgb-thermal perception tasks,” *IEEE Robot. Autom. Lett.*, vol. 8, no. 7, Jul. 2023.

[16] I. Alonso and A. C. Murillo, “Ev-segnet: Semantic segmentation for event-based cameras,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2019, pp. 0–0.

[17] R. Xia, C. Zhao, M. Zheng, Z. Wu, Q. Sun, and Y. Tang, “Cmda: Cross-modality domain adaptation for nighttime semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 21 572–21 581.

[18] L. Wang, Y. Chae, S.-H. Yoon, T.-K. Kim, and K.-J. Yoon, “Evdistill: Asynchronous events to end-task learning via bidirectional reconstruction-guided cross-modal knowledge distillation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 608–619.

[19] Y. Lv, Z. Liu, and G. Li, “Context-aware interaction network for rgb-t semantic segmentation,” *IEEE Trans. Multimedia*, vol. 26, pp. 6348–6360, Jan. 2024.

[20] B. Lin, Z. Lin, Y. Guo, Y. Zhang, J. Zou, and S. Fan, “Variational probabilistic fusion network for rgb-t semantic segmentation,” 2023, arXiv:2307.08536.

[21] J. Cao, H. Leng, D. Lischinski, D. Cohen-Or, C. Tu, and Y. Li, “Shapeconv: Shape-aware convolutional layer for indoor rgb-d semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7088–7097.

[22] R. K. Mahabadi, S. Ruder, M. Dehghani, and J. Henderson, “Parameter-efficient multi-task fine-tuning for transformers via shared hypernetworks,” 2021, arXiv:2106.04489.

[23] X. Liu, K. Ji, Y. Fu, W. L. Tam, Z. Du, Z. Yang, and J. Tang, “P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks,” 2021, arXiv:2110.07602.

[24] Q. He, “Prompting multi-modal image segmentation with semantic grouping,” in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, no. 3, Feb. 2024, pp. 2094–2102.

[25] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan, “Efficient multimodal semantic segmentation via dual-prompt learning,” *arXiv:2312.00360*, 2023.

[26] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 3213–3223.

[27] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, “Scene parsing through ade20k dataset,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 633–641.

[28] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, “The pascal visual object classes challenge: A retrospective,” *Int. J. Comput. Vis.*, vol. 111, pp. 98–136, Jun. 2015.

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

[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 30, Dec. 2017.

[31] P. Wang, P. Chen, Y. Yuan, D. Liu, Z. Huang, X. Hou, and G. Cottrell, “Understanding convolution for semantic segmentation,” in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*. Ieee, Mar. 2018, pp. 1451–1460.

[32] C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, and N. Sang, “Bisenet: Bilateral segmentation network for real-time semantic segmentation,” in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, Sep. 2018, pp. 325–341.

[33] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, “Segmenter: Transformer for semantic segmentation,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 7262–7272.

[34] R. Girdhar, M. Singh, N. Ravi, L. Van Der Maaten, A. Joulin, and I. Misra, “Omnivore: A single model for many visual modalities,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 16 102–16 112.

[35] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, “Segmenting objects in day and night: Edge-conditioned cnn for thermal image semantic segmentation,” *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 7, pp. 3069–3082, Jul. 2020.

[36] Z. Sun, N. Messikommer, D. Gehrig, and D. Scaramuzza, “Ess: Learning event-based semantic segmentation from still images,” in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Oct. 2022, pp. 341–357.

[37] J. Zhang, H. Liu, K. Yang, X. Hu, R. Liu, and R. Stiefelhagen, “Cmx: Cross-modal fusion for rgb-x semantic segmentation with transformers,” *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 12, pp. 14 679–14 694, Dec. 2023.

[38] J. Zhang, R. Liu, H. Shi, K. Yang, S. Reiß, K. Peng, H. Fu, K. Wang, and R. Stiefelhagen, “Delivering arbitrary-modal semantic segmentation,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1136–1147.

[39] Z. Wan, Y. Wang, S. Yong, P. Zhang, S. Stepputtis, K. Sycara, and Y. Xie, “Sigma: Siamese mamba network for multi-modal semantic segmentation,” *arXiv:2404.04256*, 2024.

[40] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, “Adaptformer: Adapting vision transformers for scalable visual recognition,” *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 16 664–16 678, Dec. 2022.

[41] X. He, C. Li, P. Zhang, J. Yang, and X. E. Wang, “Parameter-efficient fine-tuning for vision transformers,” vol. 3, 2022, arXiv:2203.16329.

[42] Y.-C. Liu, C.-Y. Ma, J. Tian, Z. He, and Z. Kira, “Polyhistor: Parameter-efficient multi-task adaptation for dense vision tasks,” *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 36 889–36 901, Dec. 2022.

[43] N. Ding, Y. Qin, G. Yang, F. Wei, Z. Yang, Y. Su, S. Hu, Y. Chen, C.-M. Chan, W. Chen *et al.*, “Parameter-efficient fine-tuning of large-scale pre-trained language models,” *Nat. Mach. Intell.*, vol. 5, no. 3, pp. 220–235, Mar. 2023.

[44] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. De Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," in *Proc. Int. Conf. Mach. Learn. (ICML)*. PMLR, Jun. 2019, pp. 2790–2799.

[45] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021, arXiv:2106.09685.

[46] S. Hu, Z. Zhang, N. Ding, Y. Wang, Y. Wang, Z. Liu, and M. Sun, "Sparse structure search for delta tuning," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 35, pp. 9853–9865, Dec. 2022.

[47] X. Ouyang, S. M. A. Ansari, F. X. Lin, and Y. Ji, "Efficient nlp model finetuning via multistage data filtering," 2022, arXiv:2207.14386.

[48] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Oct. 2022, pp. 709–727.

[49] K. Hambardzumyan, H. Khachatrian, and J. May, "Warp: Word-level adversarial reprogramming," 2021, arXiv:2101.00121.

[50] W. Wang, J. Dai, Z. Chen, Z. Huang, Z. Li, X. Zhu, X. Hu, T. Lu, L. Lu, H. Li *et al.*, "Internimage: Exploring large-scale vision foundation models with deformable convolutions," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 14 408–14 419.

[51] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Oct. 2012, pp. 746–760.

[52] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 567–576.

[53] J. Liu, Z. Liu, G. Wu, L. Ma, R. Liu, W. Zhong, Z. Luo, and X. Fan, "Multi-interactive feature learning and a full-time multi-modality benchmark for image fusion and segmentation," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 8115–8124.

[54] J. Binas, D. Neil, S.-C. Liu, and T. Delbruck, "Ddd17: End-to-end davis driving dataset," 2017, arXiv:1711.01458.

[55] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*. IEEE, Oct. 2019, pp. 1440–1444.

[56] X. Chen, K.-Y. Lin, J. Wang, W. Wu, C. Qian, H. Li, and G. Zeng, "Bi-directional cross-modality feature propagation with separation-and-aggregation gate for rgb-d semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Aug. 2020, pp. 561–577.

[57] Y. Wang, W. Huang, F. Sun, T. Xu, Y. Rong, and J. Huang, "Deep multimodal fusion by channel exchanging," *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, vol. 33, pp. 4835–4845, Dec. 2020.

[58] L.-Z. Chen, Z. Lin, Z. Wang, Y.-L. Yang, and M.-M. Cheng, "Spatial information guided convolution for real-time rgbd semantic segmentation," *IEEE Trans. Image Process.*, vol. 30, pp. 2313–2324, Jan. 2021.

[59] D. Seichter, M. Köhler, B. Lewandowski, T. Wengefeld, and H.-M. Gross, "Efficient rgb-d semantic segmentation for indoor scene analysis," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*. IEEE, May 2021, pp. 13 525–13 531.

[60] W. Zhou, E. Yang, J. Lei, and L. Yu, "Frnet: Feature reconstruction network for rgb-d indoor scene parsing," *IEEE J. Sel. Topics Signal Process.*, vol. 16, no. 4, pp. 677–687, May 2022.

[61] W. Zhou, E. Yang, J. Lei, J. Wan, and L. Yu, "Pgdenet: Progressive guided fusion and depth enhancement network for rgb-d indoor scene parsing," *IEEE Trans. Multimedia*, vol. 25, pp. 3483–3494, Mar. 2022.

[62] S. B. Fischedick, D. Seichter, R. Schmidt, L. Rabes, and H.-M. Gross, "Efficient multi-task scene analysis with rgb-d transformers," in *Proc. Int. Joint Conf. Neural Networks (IJCNN)*. IEEE, Jul. 2023, pp. 1–10.

[63] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12 186–12 195.

[64] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multimodal multi-task masked autoencoders," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Oct. 2022, pp. 348–367.

[65] W. Wang, T. Zhuo, X. Zhang, M. Sun, H. Yin, Y. Xing, and Y. Zhang, "Automatic network architecture search for rgb-d semantic segmentation," in *Proc. 31st ACM Int. Conf. Multimedia*, Oct. 2023, pp. 3777–3786.

[66] J. Yang, L. Bai, Y. Sun, C. Tian, M. Mao, and G. Wang, "Pixel difference convolutional network for rgb-d semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[67] B. Yin, X. Zhang, Z. Li, L. Liu, M.-M. Cheng, and Q. Hou, "Dformer: Rethinking rgbd representation learning for semantic segmentation," 2023, arXiv:2309.09668.

[68] S. Srivastava and G. Sharma, "Omnivec: Learning robust representations with cross modal sharing," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2024, pp. 1236–1248.

[69] D. Jia, J. Guo, K. Han, H. Wu, C. Zhang, C. Xu, and X. Chen, "Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer," 2024, arXiv:2406.01210.

[70] Z. Zhao, S. Xu, C. Zhang, J. Liu, P. Li, and J. Zhang, "Didfuse: Deep image decomposition for infrared and visible image fusion," 2020, arXiv:2003.09210.

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

[72] F. Deng, H. Feng, M. Liang, H. Wang, Y. Yang, Y. Gao, J. Chen, J. Hu, X. Guo, and T. L. Lam, "Feanet: Feature-enhanced attention network for rgb-thermal real-time semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Systems (IROS)*. IEEE, Sep. 2021, pp. 4467–4473.

[73] Z. Huang, J. Liu, X. Fan, R. Liu, W. Zhong, and Z. Luo, "Reconet: Recurrent correction network for fast and efficient multi-modality image fusion," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Oct. 2022, pp. 539–555.

[74] D. Wang, J. Liu, X. Fan, and R. Liu, "Unsupervised misaligned infrared and visible image fusion via cross-modality image generation and registration," 2022, arXiv:2205.11876.

[75] J. Liu, X. Fan, Z. Huang, G. Wu, R. Liu, W. Zhong, and Z. Luo, "Target-aware dual adversarial learning and a multi-scenario multi-modality benchmark to fuse infrared and visible for object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 5802–5811.

[76] G. Li, Y. Wang, Z. Liu, X. Zhang, and D. Zeng, "Rgb-t semantic segmentation with location, activation, and sharpening," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 3, pp. 1223–1235, Mar. 2022.

[77] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *Proc. AAAI Conf. Artif. Intell.*, vol. 36, no. 3, Feb. 2022, pp. 3571–3579.

[78] W. Zhou, S. Dong, J. Lei, and L. Yu, "Mtanet: Multitask-aware network with hierarchical multimodal fusion for rgb-t urban scene understanding," *IEEE Trans. Intell. Vehicles*, vol. 8, no. 1, pp. 48–58, 2022.

[79] W. Zhou, X. Lin, J. Lei, L. Yu, and J.-N. Hwang, "Mffenet: Multiscale feature fusion and enhancement network for rgb–thermal urban road scene parsing," *IEEE Trans. Multimedia*, vol. 24, pp. 2526–2538, Jun. 2021.

[80] C. Xu, Q. Li, X. Jiang, D. Yu, and Y. Zhou, "Dual-space graph-based interaction network for rgb-thermal semantic segmentation in electric power scene," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 4, pp. 1577–1592, 2022.

[81] S. Zhao and Q. Zhang, "A feature divide-and-conquer network for rgb-t semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2892–2905, 2023.

[82] Y. Wang, G. Li, and Z. Liu, "Sgfnet: semantic-guided fusion network for rgb-thermal semantic segmentation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 12, pp. 7737–7748, 2023.

[83] J. Li, P. Yun, Q. Chen, and R. Fan, "Hapnet: Toward superior rgb-thermal scene parsing via hybrid, asymmetric, and progressive heterogeneous feature fusion," 2024, arXiv:2404.03527.

[84] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, arXiv:1711.05101.

[85] T. Xiao, Y. Liu, B. Zhou, Y. Jiang, and J. Sun, "Unified perceptual parsing for scene understanding," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, Sep. 2018, pp. 418–434.

[86] Z. Geng, M.-H. Guo, H. Chen, X. Li, K. Wei, and Z. Lin, "Is attention better than matrix decomposition?" *arXiv:2109.04553*, 2021.

[87] Y. Kim, J. Chough, and P. Panda, "Beyond classification: Directly training spiking neural networks for semantic segmentation," *Neuromorph. Comput. Eng.*, vol. 2, no. 4, p. 044015, 2022.

[88] R. Zhang, L. Leng, K. Che, H. Zhang, J. Cheng, Q. Guo, J. Liao, and R. Cheng, "Accurate and efficient event-based semantic segmentation using adaptive spiking encoder-decoder network," 2023, arXiv:2304.11857.

[89] S. D. Biswas, A. Kosta, C. Liyanagedera, M. Apolinario, and K. Roy, "Halsie: Hybrid approach to learning segmentation by simultaneously exploiting image and event modalities," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*. IEEE, Jan. 2024, pp. 5952–5962.

[90] J. Zhang, K. Yang, and R. Stiefelhagen, "Exploring event-driven dynamic context for accident scene segmentation," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 3, pp. 2606–2622, 2021.

[91] W. Ji, J. Li, C. Bian, Z. Zhou, J. Zhao, A. L. Yuille, and L. Cheng, "Multispectral video semantic segmentation: A benchmark dataset and baseline," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 1094–1104.

[92] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, "Ccnet: Criss-cross attention for semantic segmentation," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[93] Y. Yuan, X. Chen, and J. Wang, "Object-contextual representations for semantic segmentation," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*. Springer, Aug. 2020, pp. 173–190.

[94] S. W. Oh, J.-Y. Lee, N. Xu, and S. J. Kim, "Video object segmentation using space-time memory networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 9226–9235.

[95] M. Paul, M. Danelljan, L. Van Gool, and R. Timofte, "Local memory attention for fast video semantic segmentation," in *Proc. IEEE/RSJ Int. Conf. Intell. Robots Systems (IROS)*. IEEE, Sep. 2021, pp. 1102–1109.

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

[97] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, 2014.

[98] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2014.

[99] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Proc. 14th Int. Conf. Intelligent Mobile and Cloud Computing Services (ICIMCS)*, 2014.

[100] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 32, no. 5, pp. 2075–2089, 2020.

[101] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2020.

[102] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "Rgb-d salient object detection with cross-modality modulation and selection," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2020.

[103] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2020.

[104] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao, "Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network," in *Proc. Eur. Conf. Comput. Vis.(ECCV)*, 2020.

[105] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Trans. Image Process.*, vol. 32, pp. 1329–1340, 2023.

[106] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, 2017.

[107] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2012.

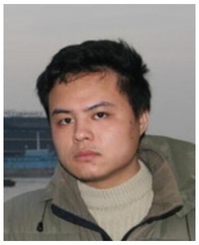
[108] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2014.

[109] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv:1805.10421*, 2018.

[110] Y. Wang, S. Mukherjee, X. Liu, J. Gao, A. H. Awadallah, and J. Gao, "Adamix: Mixture-of-adapter for parameter-efficient tuning of large language models," 2022, arXiv:2205.12410.

[111] J. He, C. Zhou, X. Ma, T. Berg-Kirkpatrick, and G. Neubig, "Towards a unified view of parameter-efficient transfer learning," 2021, arXiv:2110.04366.

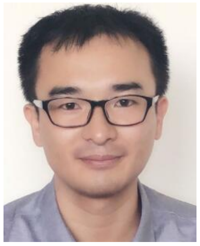
[112] X. L. Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, arXiv:2101.00190.



Ye Liu (Member, IEEE) received the Ph.D. degree in computer science from Fudan University, Shanghai, China, in 2013, and the B.Eng. degree in computer science from Tongji University, Shanghai, China, in 2007. Since 2013, he has been a faculty member with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing, China. His current research interests include computer vision and pattern recognition.

Pengfei Wu received the B.S. degree in Suzhou University of Science and Technology, Suzhou, China. He is currently pursuing the M.S. degree in Nanjing University of Posts and Telecommunications, Nanjing, China. His research interests include computer vision, multi-modal semantic segmentation.

Miaohui Wang (Senior Member, IEEE) received the Ph.D. degree from the Department of Electronic Engineering, The Chinese University of Hong Kong (CUHK), Hong Kong, China.,From 2014 to 2015, he was a Researcher working on the standardization of video coding with the Innovation Laboratory, InterDigital Inc., San Diego, CA, USA. From 2015 to 2017, he was a Senior Research Staff working on computer vision and machine learning with The Creative Life Research Institute of Hong Kong, Hong Kong, China. He is currently a tenured Associate Professor with the College of Electronics and Information Engineering, Shenzhen University, Shenzhen, China. He has authored or coauthored more than 80 peer-reviewed papers in top-tier international journals and conferences. His research interests include high-dimension visual data compression and transmission, medical image analysis, computer vision, and machine learning.,Dr. Wang was the recipient of the Best Thesis Award from the Ministry of Education of Shanghai City and Fudan University, respectively. He was the recipient of the Best Paper Award from International Conference on Advanced Hybrid Information Processing in 2018, and the Outstanding Reviewer Award from IEEE International Conference on Multimedia & Expo in 2021.



Jun Liu (Senior Member, IEEE) is a Professor and Chair in Digital Health at School of Computing and Communications in Lancaster University, UK. He received the PhD degree from Nanyang Technological University (NTU) in 2019. He was with Singapore University of Technology and Design (SUTD) from 2019 to 2024. His research interests include digital health, computer vision and machine learning. He is an associate editor of IEEE Transactions on Image Processing and IEEE Transactions on Biometrics, Behavior, and Identity Science.