2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)
October 14-18, 2024. Abu Dhabi, UAE

Efficient Multimodal Semantic Segmentation via Dual-Prompt Learning

Shaohua Dong[1], Yunhe Feng[1], Qing Yang[1], Yan Huang[1], Dongfang Liu[2], and Heng Fan[1]

***Abstract*— Multimodal (*e.g.*, RGB-Depth/RGB-Thermal) fusion has shown great potential for improving semantic segmentation in complex scenes (*e.g.*, indoor/low-light conditions). Existing approaches often fully fine-tune a dual-branch encoder-decoder framework with a complicated feature fusion strategy for achieving multimodal semantic segmentation, which is training-costly due to the massive parameter updates in feature extraction and fusion. To address this issue, we propose a surprisingly simple yet effective dual-prompt learning network (dubbed DPLNet) for *training-efficient* multimodal (*e.g.*, RGB-D/T) semantic segmentation. The core of DPLNet is to directly adapt a frozen pre-trained RGB model to multimodal semantic segmentation, reducing parameter updates. For this purpose, we present two prompt learning modules, comprising multimodal prompt generator (MPG) and multimodal feature adapter (MFA). MPG works to fuse the features from different modalities in a compact manner and is inserted from shallow to deep stages to generate the multi-level multimodal prompts that are injected into the frozen backbone, while MFA adapts prompted multimodal features in the frozen backbone for better multimodal semantic segmentation. Since both the MPG and MFA are lightweight, only a few trainable parameters (3.88M, 4.4% of the pre-trained backbone parameters) are introduced for multimodal feature fusion and learning. Using a simple decoder (3.27M parameters), DPLNet achieves new state-of-the-art performance or is on a par with other complex approaches on four RGB-D/T semantic segmentation datasets while satisfying parameter efficiency. Moreover, we show DPLNet is general and applicable to other multimodal segmentation tasks. Without special design, DPLNet outperforms many complicated models. The source code can be found at https://github.com/ShaohuaDong2021/DPLNet.**

I. Introduction

Semantic segmentation, aiming to assign each pixel in an image with a pre-define label, is a fundamental problem in computer vision with a wide spectrum of crucial applications such as intelligent driving and robotics, and has seen considerable progress (*e.g.* [1], [2], [3], [4], [5]) in recent years. Despite this, RGB-based segmentation models may largely degenerate when applied to complex scenarios (*e.g.*, in *cluttered indoor* environment or *low-light* condition). To handle this, an auxiliary modality (*e.g.*, Depth or Thermal), which provides supplemental information to the RGB image, has been used for multimodal, RGB+Depth (RGB-D) [6], [7], [8], [9] or RGB+Thermal (RGB-T) [10], [11], [12], [8], [9], semantic segmentation, showing promising performance.

Existing multimodal methods for semantic segmentation often adopt a straight-forward dual-branch encoder-decoder architecture (*e.g.*, [13], [12], [14], [15], [8]), where one branch is for feature extraction of the RGB modality and the other for feature of the auxiliary modality (see Fig. 1 (a)). Afterwards, a complex fusion strategy is used to merge multimodal features to achieve semantic segmentation. Albeit simple, such a framework needs to *fully fine-tune* the entire network, which is *training-costly* due to massive parameter updates in feature extraction and fusion. In addition, it requires to retain two heavy encoders after training, and thus increases deployment memory of multimodal semantic segmentation. Moreover, the auxiliary encoder branch is often initialized with a pre-trained RGB model (*e.g.*, [16], [1]) and then fully fine-tuned, which may lead to suboptimal performance due to domain gap between different modalities [17]. Noticing these issues, we ask: *Is there a better way that is* ***training-efficient***, ***memory-deployment-friendly***, *and* ***effective*** *for multimodal semantic segmentation?*

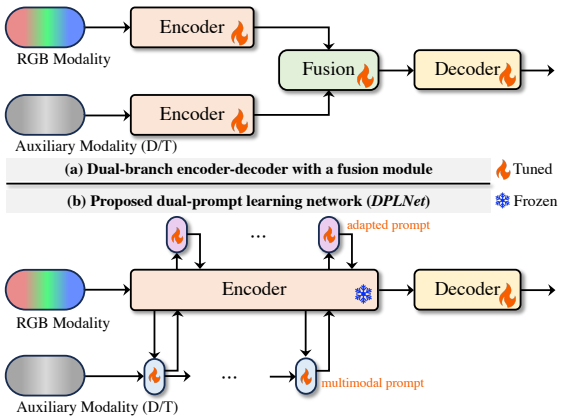


Fig. 1: Compared to current dual-branch encoder-decoder framework with a fusion module (a), our dual-prompt learning paradigm (b) does not require tuning the weighty encoder pre-trained on RGB data and only introduces a few trainable parameters for multimodal feature fusion and adaption, achieving parameter-efficient training for multimodal semantic segmentation. *Best viewed in color and by zooming in.*

We answer *yes* to the above question, and show a solution from the *prompt learning view*. The idea of prompt learning (or called prompt tuning) has originated from natural language processing (NLP) (*e.g.*, [18], [19]) and aims to transfer knowledge from frozen language models to downstream tasks by injecting textual prompts in an efficiency way. Motivated by this, researchers have recently applied prompt tuning to vision tasks (*e.g.*, [20], [21], [22], [23]) and exhibited promising results. Despite this, these prompting learning methods are *not* applicable to the complicated multimodal dense prediction tasks demanding effective feature fusion and adaption, which thus motivates us to seek a new prompt tuning approach for multimodal semantic segmentation.

[1]Dept. of Computer Science and Engineering, University of North Texas
[2]Dept. of Engineering, Rochester Institute of Technology

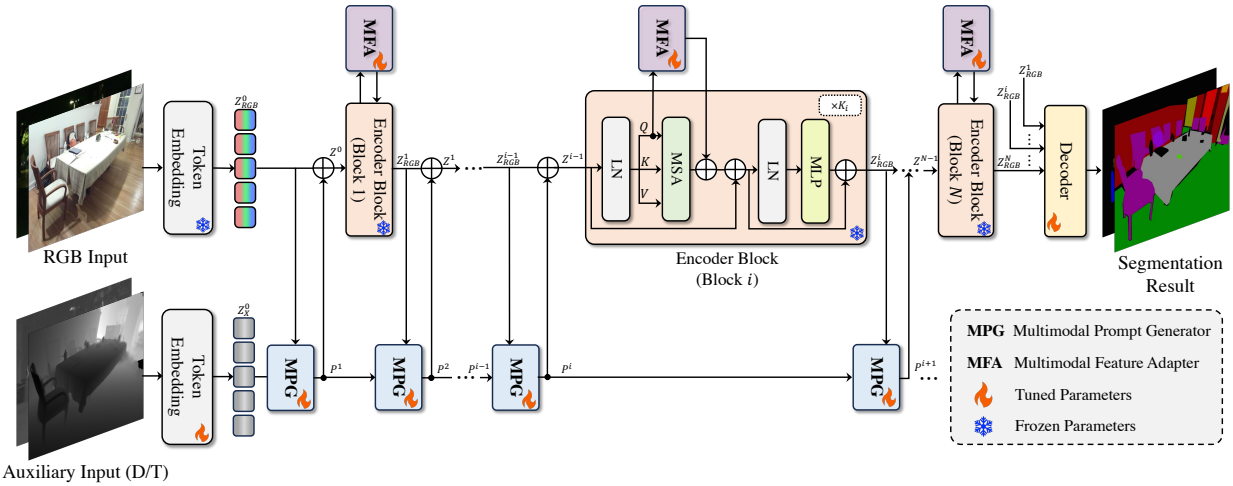


Fig. 2: Overview architecture of the proposed *DPLNet*, which adapts a frozen pre-trained model using two specially designed prompting learning modules, *MPG* for multimodal prompt generation and *MFA* for multimodal feature adaption, with only a few learnable parameters to achieve multimodal semantic segmentation in a training-efficient way.

Considering multimodal semantic segmentation involves both modality fusion and feature adaption in the frozen backbone, we propose ***DPLNet***, a novel ***D**ual-**P**rompt **L**earning **Net**work* that is able to adapt a *frozen* pre-trained model for multimodal semantic segmentation (see Fig. 1 (b)), and thus avoids massive parameter update. Specifically, DPLNet consists of two simple but crucial prompt learning modules, *i.e.*, a *multimodal prompt generator* (MPG) and a *multimodal feature adapter* (MFA). MPG aims to fuse important auxiliary modality feature into RGB feature to generate the complementary multimodal feature prompt, which can alleviate the domain gap issue. To leverage different semantic-level features, multiple MPGs are inserted from shadow to deep stages, connected in a progressive fashion, to generate multi-level multimodal feature prompts, which are injected into the frozen network to guide semantic segmentation. To adapt the frozen single-modal backbone for better multimodal feature extraction, MFA is applied in each stage by introducing a few learnable tokens, which interact with multimodal features via cross-attention for adaption to a specific task, significantly enhancing performance. Notice that, different from VPT [20] directly prepending the learnable tokens to each encoder layer, MFA applies an independent cross-attention to adapt multimodal features and obtains better results. Please note, since both MPG and MFA are lightweight, only a few minimal number of trainable parameters (equivalent to 4.4% of the original model parameters) are introduced in DPLNet. Using a simple decoder, DPLNet achieves superior results. Fig. 2 illustrates the architecture of DPLNet.

Compared with existing multimodal methods for semantic segmentation, DPLNet has several attractive properties: **First**, it is training-efficient architecture because only a few parameters requires to be tuned. **Second**, it largely reduces the deployment model size in real-world applications as it does not retrain two heavy pretrained encoder. **Third**, our DPLNet can be directly applied for multiple different multi-modal (RGB-D/T) task (*e.g.*, semantic segmentation, salient object detection), which ***unifies*** the architecture of different tasks and avoids complex task-specific model design.

To validate effectiveness of DPLNet, we conduct extensive experiments on four challenging datasets, including NYUD-v2 [24] and SUN-RGBD [25] for RGB-D semantic segmentation and MFNet [10] and PST900 [26] for RGB-T semantic segmentation. DPLNet achieves state-of-the-art performance on NYUD-v2, SUN-RGBD, and PST900 and comparable results on MFNet while being much more training-efficient. Furthermore, we show that DPLNet is general and applicable to other multimodal tasks such as salient object detection and video semantic segmentation, as shown in our experiments.

In summary, our **contributions** are as follows:

- ♠ *We propose a novel and simple yet effective dual-prompt learning network, named DPLNet, for training-efficient multimodal semantic segmentation.*
- ♡ *We introduce a multimodal prompt generator (MPG) to fuse different modalities in a compact manner, eliminating complex multimodal fusion in previous methods.*
- ♣ *We present a multimodal feature adapter (MFA) that adapts the frozen pre-trained backbone for better multimodal feature extraction, greatly boosting performance.*
- ♢ *In extensive experiments on four benchmarks, DPLNet achieves state-of-the-art results or is on par with other methods while satisfying parameter efficiency, validating its effectiveness. Moreover, we show that DPLNet is general and applicable to other tasks.*

II. Related Work

A. Multimodal Semantic Segmentation

RGB-D semantic segmentation aims to improve segmentation using depth information which consists of abundant geometric properties. Significant progress has recently been made [27], [6], [28], [29], [30], [9]. For example, the approach of [27] leverages an encoder-decoder architecture and proposes an effective fusion module to fuse the information

between two modalities. The work of [28] dynamically fuses modality features by leveraging a transformer architecture. The work of [9] adopts a similar spirit to utilize transformer as the backbone to extract features, and then fuse different modalities to achieve RGB-D semantic segmentation.

RGB-T semantic segmentation improves segmentation in low-light conditions using RGB and thermal cues. Similar to RGB-D semantic segmentation, RGB-T-based methods mainly add an auxiliary branch based on RGB-based models to extract thermal features and then fuse multimodal features for segmentation. Many methods [10], [11], [17], [12], [13], [15], [9] have been recently proposed. The approach of [11] and its extensions [13], [17] develop a dual-branch encoder-decoder architecture with a simple fusion strategy for RGB-T semantic segmentation. Interestingly, the method in [9] for RGB-D semantic segmentation is also applicable to RGB-T semantic segmentation, indicating the underlying relationship between two tasks and the demand for a unified framework.

Different from the above multimodal RGB-D/T methods for semantic segmentation that fully fine-tune a dual-branch encoder-decoder network, DPLNet aims to learn a parameter-efficient paradigm with dual-prompt tuning. In addition, DPLNet is a unified framework and applicable to various multimodal segmentation tasks with promising performance.

B. Visual Prompt Learning

Prompt learning has recently attracted extensive attention by significantly decreasing the number of trainable parameters, providing an efficient manner to leverage pre-trained models. It originates from NLP [18], [19] and is applied to various vision tasks [20], [31], [21], [22], [23], [32]. The work of [20] prepends a set of trainable parameters to adapt vision transformer (ViT) [33] to various downstream visual recognition tasks, exhibiting remarkable performance. Unlike [20], the method of [21] proposes to insert a lightweight module into ViT and achieves superior results over the full fine-tuning models. The work of [22] adds a few learnable parameters to self-attention in transformer model and applies pruning methods to reduce parameters, achieving effective and efficient prompt learning. Besides visual recognition, prompt tuning has been successfully applied in other visual tasks including tracking [31], [32], style transfer [34], etc.

Different from the aforementioned visual prompting tuning approaches, DPLNet is specifically focused on the task of multimodal dense prediction for RGB-D/T semantic segmentation, which to our knowledge has not been studied before. In addition to the difference in task, the proposed DPLNet introduces the novel dual-prompt learning, which differs from the above visual single-prompt tuning methods and displays superior performance for multimodal semantic segmentation, as can be seen in our experiments.

III. The Proposed Approach

Overview. In this paper, we introduce the simple yet effective DPLNet, which tunes a RGB-based pre-trained model for multimodal (MM) semantic segmentation in a *parameter-efficient* way. The key of DPLNet contains two prompt tuning modules, including MPG and MFA. As in Fig. 2, MPG fuses features from multiple modalities and returns a multimodal feature prompt to the frozen model, while MFA works for adapting the prompted feature to the specific task. Since the main backbone is frozen, only a few additional trainable parameters are added, leading to high efficiency in training.

A. Preliminaries

Before elaborating on our method, we first present necessary preliminaries for the used RGB-based pre-trained model and the problem definition for our task.

Pre-trained RGB Model. We focus on tuning the vision Transformer [1] (called $\mathbf{T}_{\text{RGB}}$), which has been pre-trained with a large amount of data for RGB segmentation, for MM semantic segmentation. $\mathbf{T}_{\text{RGB}}$ has two major components, including an encoder for feature extraction and a head for the specific task. The encoder consists of a stack of blocks, with each containing a set of self-attention layers. Specifically, given an RGB image I_{RGB}, $\mathbf{T}_{\text{RGB}}$ first projects it into the token embedding $Z^0 = \texttt{Emb}(I_{\text{RGB}})$. Then it extracts features via N encoder blocks $B_i(\cdot)$ (i=1,$\cdots$,N), followed by a decoder $\texttt{Dec}(\cdots)$ for generating result Y_{RGB}, as follows,

$$\begin{aligned} Z^i &= B_i(Z^{i-1}) \quad i = 1, 2, \cdots, N \\ Y_{\text{RGB}} &= \texttt{Dec}(Z^1, Z^2, \cdots, Z^N) \end{aligned} \tag{1}$$

where Z^i ($1 \leq i \leq N$) is the output from encoder block B_i, and fed to B_{i+1}. Each encoder block B_i is formed by a set of K_i self-attention layers, with each defined as follows,

$$\texttt{SAL}_k(z) = \texttt{FFN}(\texttt{MSA}(z)) \tag{2}$$

where $\texttt{SAL}_k(\cdot)$ ($1 \leq k \leq K_i$) denotes the k^{th} self-attention layer, containing an efficient multi-head self-attention module $\texttt{MSA}(\cdot)$ and a feed-forward network $\texttt{FFN}(\cdot)$. We omit the layer normalization in self-attention layer for simplicity.

Problem Definition. We aim to adapt a pre-trained $\mathbf{T}_{\text{RGB}}$ to learn a multimodal semantic segmentation model $\mathbf{T}_{\text{RGB-X}}$, which receives images I_{RGB} and I_{X} from RGB and auxiliary X (X can be Depth or Thermal) modalities and outputs the prediction $Y_{\text{RGB-X}}$, while keeping encoder of $\mathbf{T}_{\text{RGB}}$ frozen.

B. DPLNet for MM Semantic Segmentation

We propose DPLNet to adapt $\mathbf{T}_{\text{RGB}}$ for efficient learning of $\mathbf{T}_{\text{RGB-X}}$ for multimodal semantic segmentation. The core of DPLNet consists of two prompt-learning modules, including MPG and MFA. MPG aims to generate the multimodal RGB-X prompt, while MFA works on adapting the frozen encoder of $\mathbf{T}_{\text{RGB}}$ for better MM semantic segmentation.

Specifically, given a pair of RGB-X input images I_{RGB} and I_X, we first employ a patch embedding layer to obtain RGB and X token embeddings Z^0_{RGB} and Z^0_X via

$$Z^0_{RGB} = \texttt{Emb}_{RGB}(I_{\text{RGB}}) \qquad Z^0_X = \texttt{Emb}_Z(I_{\text{X}}) \tag{3}$$

where $\texttt{Emb}_{RGB}(\cdot)$ is the frozen patch embedding layer from $\mathbf{T}_{\text{RGB}}$, and $\texttt{Emb}_Z(\cdot)$ is the learnable patch embedding layer.

After that, a set of MPGs is used in DPLNet, as in Fig. 2, to learn a cascaded multimodal prompt, which is added to the original RGB flow with residual connection for the encoder

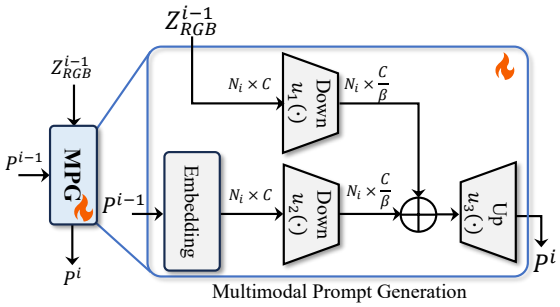


Fig. 3: Illustration of MPG, which aims to fuse RGB features with a multimodal prompt from the last stage for generating a new multimodal prompt.

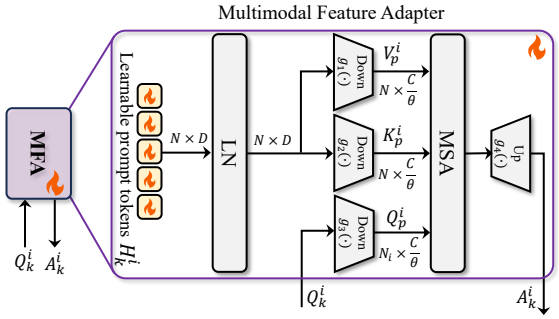


Fig. 4: Illustration of MFA, which is leveraged to adapt query features in frozen backbone with a set of learnable tokens.

block. In particular, for the i^{th} encoder B_i, its input feature Z^{i-1} is obtained by

$$Z^{i-1} = Z_{RGB}^{i-1} + P^i \quad i = 1, 2, \cdots, N \tag{4}$$

where Z_{RGB}^{i-1} represents the output feature of the $(i-1)^{\text{th}}$ encoder B_{i-1} (here $i > 1$) and P^i is the multimodal prompt generated from the i^{th} MPG as follows,

$$P^i = \texttt{MPG}(Z_{RGB}^{i-1}, P^{i-1}) \tag{5}$$

where $\texttt{MPG}(\cdot,\cdot)$ is our MPG (as in Sec. C), and P^{i-1} multimodal prompt generated from last MPG, where $P^0 = Z_X^0$.

With the multimodal prompt Z^{i-1}, we send it to B_i for further feature learning. However, because B_i is pre-trained only with RGB modality and frozen in DPLNet, it may not be suitable for multimodal feature learning. To address this issue, we introduce MFA to adapt B_i for multimodal feature learning. The key idea is to leverage a small set of learnable prompts to generate the query-adaptive prompt and insert it into each self-attention layer of B_i. Specifically, for the k^{th} self-attention layer $\texttt{SAL}_k$ of B_i, a set of learnable prompt tokens H_k^i are used to generate the adaption prompt via

$$A_k^i = \texttt{MFA}(H_k^i, Q_k^i) \tag{6}$$

where $\texttt{MFA}(\cdot,\cdot)$ represents our MFA as describe late, $Q_k^i = W_q \texttt{SAL}_{k-1}(z)$ is the query generated by last self-attention layer, and A_k^i is the generated adaption prompt. With A_k^i, we insert it into $\texttt{SAL}_k$ with residual addition via

$$\texttt{SAL}_k(z) = \texttt{FFN}(\texttt{MSA}(z) + A_k^i) \tag{7}$$

With the above equation, we can adapt B_i to better multimodal feature learning by $Z_{RGB}^i = B_i(Z^{i-1})$, where Z_{RGB}^i is the output feature after the encoder block B_i, and will be fed to the next encoder block.

After the N^{th} encoder block, all features Z_{RGB}^i are sent to the decoder for segmentation as in $\mathbf{T}_{\text{RGB}}$ via

$$Y_{\text{RGB-X}} = \texttt{Dec}(Z_{RGB}^1, Z_{RGB}^2, \cdots, Z_{RGB}^N) \tag{8}$$

Notice that, Z_{RGB}^i $(1 \le i \le N)$ here contains information from both RGB and X modalities. Due to the domain gap between RGB and RGB-X tasks, the decoder used for RGB-X semantic segmentation in this work is set to be learnable. During training, we use a simple cross-entropy loss to update parameters in DPLNet. Fig. 2 illustrates our DPLNet.

C. Multimodal Prompt Generation (MPG)

A key of DPLNet is to generate appropriate multimodal prompts to guide the network for effective MM semantic segmenting. To this end, a multimodal prompt generator (MPG) is designed to fuse multimodal features in a compact way, which is thus more parameter-efficient than existing heavy fusion strategies. Fig. 3 illustrates MPG. More concretely, given RGB features $Z_{RGB}^{i-1} \in R^{N_i \times C}$ from encoder B_{i-1} and multimodal prompt $P^{i-1} \in R^{N_i \times C}$ from last MPG module, where N_i is the token number and C the token dimension, we follow the work of [1] to adopt a patch embedding strategy (which is implemented using a light 3×3 convolutional layer) to obtain more robust multimodal prompt information via

$$\hat{P}^{i-1} = \texttt{Emb}_\texttt{i}(P^{i-1}) \quad i = 2, 3, \cdots, N \tag{9}$$

where $\texttt{Emb}_\texttt{i}(\cdot)$ denotes the learnable patch embedding layer. We omit the patch embedding layer in the first MPG module because the I_X is tokenized by $\texttt{Emb}_Z(\cdot)$ in Eq. 3.

Inspired by [21], we first downsample features from different modalities in channel dimenson for fusion and then upsample fused feature back to original dimension for better adaption. Mathematically, this can be formulated as follows,

$$P^i = u_3(u_1(Z_{RGB}^{i-1}) + u_2(\hat{P}^{i-1})) \quad i = 1, 2, \cdots, N \tag{10}$$

where $u_1(\cdot)$ and $u_2(\cdot)$ are two projection layers to reduce feature dimension, while $u_3(\cdot)$ works to increase the feature dimension. All the parameters are learnable in MPG module and the reducing factor β is set to 4.

D. Multimodal Feature Adapter (MFA)

In our DPLNet, the generated multimodal prompt is sent to the pre-trained RGB Transformer block for feature learning. Due to domain gap between different tasks, this may not be optimal. To deal with this, we propose a multimodal feature adapter (MFA) that inserts a set of learnable prompt tokens to improve feature learning for pre-trained RGB Transformer blocks. Please notice, compared to the visual prompt tuning (VPT) [20] in token space for visual recognition, the proposed MFA differs by adapting queries only for dense prediction, which is more effective. More specifically, as shown in Fig. 4, we generate adaption prompt $A_k^i \in R^{B \times N_i \times C}$ based on original self-attention query $Q_k^i \in R^{B \times N_i \times C}$, which can learn query-adaptive information from the learnable prompt

	ACNet -SS [27]	SGNet -MS [35]	SA-Gate -MS [6]	PGDENet -SS [36]	TokenFusion-B3 -SS [28]	MultiMAE -SS [29]	Omnivore-B -SS [30]	CMX-B5 -MS [8]	CMXNeXt -MS [9]	DFormer-L -MS [37]	**DPLNet** -SS (Ours)	**DPLNet** -MS (Ours)
Backbone	ResNet-50	ResNet-101	ResNet-101	ResNet-34	MiT-B3	ViT-B	Swin-B	MiT-B5	MiT-B4	DFormer-L	MiT-B5	MiT-B5
Learnable Params (M)	116.6	64.7	110.9	100.7	45.9	95.2	95.7	181.1	119.6	39.0	**7.15**	**7.15**
Total Params (M)	116.6	64.7	110.9	100.7	45.9	95.2	95.7	181.1	119.6	39.0	88.58	88.58
mIoU (%)	48.3	51.1	52.4	53.7	54.2	56.0	54.0	56.9	56.9	57.2	**58.3**	**59.3**

TABLE I: RGB-D semantic segmentation on NYUDv2. "SS" and "MS" denote single- or multi-scale for testing. The best two results are highlighted in **red** and **blue** in all comparison tables.

	ACNet -SS [27]	SGNet -MS [35]	SA-Gate -MS [6]	PGDENet -SS [36]	TokenFusion-B3 -SS [28]	CMX-B4 -MS [8]	CMX-B5 -MS [8]	CMXNeXt -MS [9]	DFormer-B -MS [37]	DFormer-L -MS [37]	**DPLNet** -SS (Ours)	**DPLNet** -MS (Ours)
Backbone	ResNet-50	ResNet-101	ResNet-101	ResNet-34	MiT-B3	MiT-B4	MiT-B5	MiT-B4	DFormer-B	DFormer-L	MiT-B5	MiT-B5
Learnable Params (M)	116.6	64.7	110.9	100.7	45.9	139.9	181.1	119.6	29.5	39.0	**7.15**	**7.15**
Total Params (M)	116.6	64.7	110.9	100.7	45.9	139.9	181.1	119.6	29.5	39.0	88.58	88.58
mIoU (%)	48.1	48.6	49.4	51.0	51.0†	52.1	52.4	51.9†	51.2	**52.5**	52.1	**52.8**

TABLE II: RGB-D semantic segmentation on SUN RGB-D. † indicates that we follow the results from DFormer [37].

tokens $H_k^i \in R^{B\times N\times D}$, where N and D are the number and dimension of learnable prompt tokens, respectively.

Since a pre-trained model resides on low intrinsic dimension [38], we obtain the prompt query Q_p^i by projecting self-attention query Q_k^i with linear projection to reduce redundant features and parameters through a reducing factor θ via

$$Q_p^i = g_3(Q_k^i) \quad i = 1, 2, \cdots, N \tag{11}$$

Accordingly, we also reduce the dimension of K_p^i and V_p^i by linear projection layer through the reducing factor θ after a normalization layer as follows,

$$V_p^i = g_1(\texttt{LN}(H_k^i)) \qquad K_p^i = g_2(\texttt{LN}(H_k^i)) \tag{12}$$

where $g_1(\cdot)$, $g_2(\cdot)$ and $g_3(\cdot)$ are liner projection layers, $\texttt{LN}(\cdot)$ is the normalization layer and the reducing factor θ is set to 32. After that, we expand the dimension into original embedding space using $g_4(\cdot)$ after the MSA layer to generate the adaption prompt A_k^i as follows,

$$A_k^i = g_4(\texttt{MSA}(Q_p^i, K_p^i, V_p^i)) \tag{13}$$

IV. Experiments

Datasets and Evaluation Metric. To verify effectiveness of DPLNet, we conduct experiments on four datasets, including two RGB-D datasets NYUDv2 [24] and SUN RDB-D [25] and two RGB-T datasets MFNet [10] and PST900 [26].

NYUDv2 [24] has 1,449 RGB-D images from 41 classes including background, among which 795 are used for training and the rest for testing. SUN RGB-D [25] comprises 10,335 labeled RGB-D images which are divided into 5,285 and 5,050 RGB-D pairs for training and testing, respectively. MFNet [10] has 1,569 RGB-T images including 820 daytime image pairs and 749 nighttime image pairs from eight foreground classes and one background class. PST900 [26] contains 894 aligned pairs of RGB and thermal images from five categories. We follow [26] for the training/test split.

Following existing methods, we use the popular metric of Intersection over Union (IoU) for evaluation and comparison.

Implementation. We conduct all experiments on a single Nvidia A6000 GPU using PyTorch. As each dataset heavily varies in distributions, we use different hyper-parameters (*e.g.*, learning rate) for them as in existing works. Specifically, for NYUDv2 [24], the learning rate is 4e-2 with a weight decay 5e-4. For SUN RGB-D [25], we set the learning rate as 1e-2 and the weight decay as 5e-4. Similar to many RGB-D semantic segmentation works (*e.g.*, [37]), we adopt the multi-scale (MS) flip inference strategy with scales {0.5, 0.75, 1, 1.25, 1.5}, but also report the result of our single-scale (SS) version. For MFNet [10], the learning rate is 5e-3 with a weight decay of 5e-4. For PST900 [26], the learning rate is 1e-3 with a weight decay of 5e-4. For all experiments, we utilize the same network architecture and it adds only a few learnable parameters of 7.15M, with 3.88M for adapting the backbone which is only 4.4% of the original backbone, and 3.27M for the decoder. Our code will be released.

A. State-of-the-art Comparison

DPLNet for RGB-D semantic segmentation. For RGB-D semantic segmentation, we evaluate DPLNet on two popular, benchmarks including NYUDv2 [24] and SUN RGB-D [25], and compare it with many recent state-of-the-art (SOTA) approaches. As in Tab. I, our DPLNet, even with single-scale inference, achieves promising results with 0.583 mIoU on NYUDv2. When applying a multi-scale inference strategy, DPLNet further improves the result to 0.593 mIoU, which significantly outperforms other models. In addition, DPLNet only contains 7.15M learnable parameters, which is more efficient than other approaches in training. Likewise, on SUN RGB-D as shown in Tab. II, our DPLNet with multi-scale inference achieves the best performance with 0.528 mIOU yet has much less trainable parameters compared to existing approaches. All these show the effectiveness of our DPLNet.

DPLNet for RGB-T semantic segmentation. For RGB-T semantic segmentation, we conduct experiments on MFNet [10] and PST900 [26]. Note that, for RGB-T semantic segmentation, all methods are evaluated with single-scale inference. Tab. III shows comparison of DPLNet with 11 SOTA models on MFNet. CMNexT shows the best result with mIoU of 0.599. Compared to CMNext, DPLNet achieves comparable performance with 0.593 mIoU, while significantly reducing the number of trainable parameters by 16× (7.15M *v.s.* 119.6M), which shows the effectiveness of

	MFNet [10]	RTFNet [11]	FuseSeg-161 [39]	ABMDRNet [17]	EGFNet [12]	MTANet [15]	GEBNet [40]	GMNet [14]	CMX-B2 [8]	CMX-B4 [8]	CMNeXt [9]	**DPLNet** (Ours)
Backbone	-	ResNet-152	DenseNet-161	ResNet-50	ResNet-152	ResNet-152	ConvNeXt-S	ResNet-50	MiT-B2	MiT-B4	MiT-B4	MiT-B5
Learnable Params (M)	8.4	337.1	141.5	-	201.3	121.9	126.2	149.8	66.6	139.9	119.6	**7.15**
Total Params (M)	8.4	337.1	141.5	-	201.3	121.9	126.2	149.8	66.6	139.9	119.6	88.58
mIoU (%)	39.7	53.2	54.5	54.8	54.8	56.1	56.2	57.3	58.2	**59.7**	**59.9**	59.3

TABLE III: RGB-T semantic segmentation results on MFNet benchmark.

	RTFNet [11]	PSTNet [26]	MTANet [15]	GMNet [14]	EGFNet [12]	EGFNet-ConvNext [13]	GEBNet [40]	CACFNet [41]	**DPLNet** (Ours)
Backbone	ResNet-152	ResNet-18	ResNet-152	ResNet-50	ResNet-152	ConvNeXt-B	ConvNeXt-S	ConvNeXt-B	MiT-B5
Learnable Params (M)	337.1	105.8	121.9	149.8	201.3	340.0	126.2	198.6	**7.15**
Total Params (M)	337.1	105.8	121.9	149.8	201.3	340.0	126.2	198.6	88.58
mIoU (%)	60.5	68.4	78.6	84.1	78.5	85.4	81.2	**86.6**	**86.7**

TABLE IV: RGB-T semantic segmentation results on PST900 benchmark.

Method	Learnable Params (M)	mIoU (%)
w/o MPG	6.86	57.4
w/o MFA	5.49	57.4
Frozen decoder	3.88	55.1
Fully fine-tuning	88.58	58.1
DPLNet (Ours)	7.15	58.3

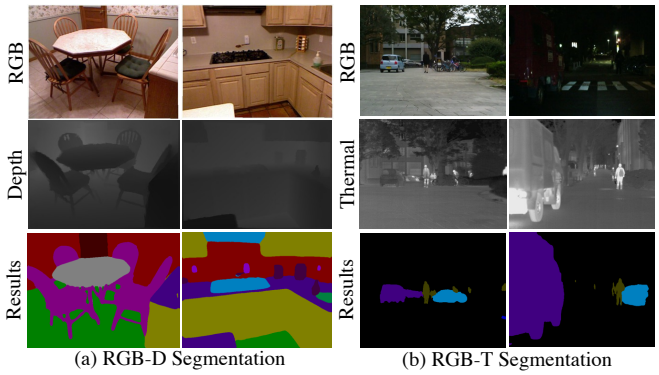
TABLE V: Ablation of key modules in DPLNet.

Backbone	Learnable Params (M)	mIoU (%)
MiT-2	5.96	54.2
MiT-3	6.37	54.7
MiT-4	6.72	57.2
MiT-5 (Ours)	7.15	58.3

TABLE VI: Ablation of different backbones.

Method	Learnable Params (M)	mIoU (%)
VPT	4.34	20.7
Adaptformer	7.49	24.0
DPLNet (Ours)	7.15	58.3

TABLE VII: Ablation of different adaptation methods.



(a) RGB-D Segmentation Results on NYUDv2

(b) RGB-T Segmentation Results on MFNet

Fig. 5: Our visualization results on NYUDv2 and MFNet.

our approach. On PST900 as shown in Tab. IV, our DPLNet is compared to 8 recent models. From Tab. IV, we can clearly observe that DPLNet achieves the best performance with mIoU of 0.867, outperforming the second best CACFNet with 0.866 mIoU yet comprising much fewer parameters, evidencing the efficacy of the proposed method.

Qualitative Evaluation. We also give the qualitative results of RGB-D and RGB-T semantic segmentation as shown in Fig. 5. As we can see, our method can segment the object accurately, which validates the effectiveness of our method.

B. Ablation Study

We ablate the proposed DPLNet with different settings on NYUDv2 [24] using **single-scale** inference.

Ablation on MPG. To validate effectiveness of MPG, we replace it with additive fusion at each stage (please notice, we keep the patch embedding layer). As in Tab. V, the performance degrades by 0.9% in mIoU without MPG, which shows the necessity of MPG. We further ablate the position of MPG. As in Tab. VIIIa and Tab. VIIIb, the results indicate that using all MPG modules can achieve the best result.

Ablation on MFA. Tab. V shows the ablation study on MFA. We directly remove MFA at each stage and the mIoU drops by 0.9%. Besides, we ablate the length and dimension of learnable prompt tokens H_k^i in Tab. VIIIc and Tab. VIIId, respectively. We observe that choosing 30 tokens and setting the prompt dimension to 32 lead to the best performance. Moreover, we ablate the position and numbers of MFA in DPLNet. Similar to ablation of MPG, we adopt bottom to top and top to bottom approaches as in Tab. VIIIe and in Tab. VIIIf, respectively. As shown, we observe that the result is improved when using MFA in all stages.

Ablation on different training paradigm. We further assess our method on two different training paradigms: (i) freezing segmentation decoder and (ii) fully fine-tuning the whole network. As in Tab. V, the performance drops by 3.2% in mIoU when we freeze the segmentation decoder. We argue that this is caused by the domain gap in complex dense prediction tasks. When fully fine-tuning the whole network, we observe that DPLNet with 0.583 mIoU score achieves better result than the fully fine-tuned version with 0.581 mIoU score while satisfying parameter-efficiency, this is because our method avoids using pretrained RGB network for auxiliary modality feature extraction and thus can alleviate domain gap issue, showing its effectiveness and efficiency.

Ablation on different backbone. We compare our DPLNet with different backbones. As in Tab. VI, DPLNet with more powerful backbone shows the best result.

Ablation on different adaptation strategies. We compare different adaptation methods with our DPLNet. As shown in Tab. VII, our method achieves superior performance than other two methods while only using 7.15M parameters.

(a) Ablation studies for the position of MPG module (Bottom to top).

Position	Params	mIoU (%)
Stage 1	6.86	57.6
Stage 1 $\rightarrow$ 2	6.87	57.8
Stage 1 $\rightarrow$ 3	6.95	57.6
Stage 1 $\rightarrow$ 4 (Ours)	7.15	58.3

(b) Ablation studies for the position of MPG module (Top to bottom).

Position	Params	mIoU (%)
Stage 4	7.05	57.3
Stage 4 $\rightarrow$ 3	7.13	57.2
Stage 4 $\rightarrow$ 2	7.14	57.8
Stage 4 $\rightarrow$ 1 (Ours)	7.15	58.3

(c) Ablation studies of learnable prompt length in MFA module.

Prompt length	Params	mIoU (%)
10	7.1456	57.8
20	7.1460	57.5
30 (Ours)	7.1463	58.3
40	7.1466	57.7

(d) Ablation studies of learnable prompt dimension in MFA module.

Prompt dimension	Params	mIoU (%)
8	6.41	57.7
16	6.65	57.9
32 (Ours)	7.15	58.3
64	8.13	57.8

(e) Ablation studies for the position of MFA module (Bottom to top).

Position	Params	mIoU (%)
Stage 1	5.51	57.4
Stage 1 $\rightarrow$ 2	5.57	57.7
Stage 1 $\rightarrow$ 3	6.94	57.8
Stage 1 $\rightarrow$ 4 (Ours)	7.15	58.3

(f) Ablation studies for the position of MFA module (Top to bottom).

Position	Params	mIoU (%)
Stage 4	5.69	57.7
Stage 4 $\rightarrow$ 3	7.07	57.6
Stage 4 $\rightarrow$ 2	7.13	57.7
Stage 4 $\rightarrow$ 1 (Ours)	7.15	58.3

TABLE VIII: Ablation studies on different settings for the proposed DPLNet using single-scale. Note that, $i \rightarrow j$ in (a), (b), (e), and (f) indicates the which layer the MPG module is inserted into. The parameters in each table are measured by "M".

Model	NJU2K [42]				NLPR [43]				DES [44]				SIP [45]				LFSD [46]			
	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$
CMWNet [47]	.903	.912	.880	.046	.917	.951	.872	.029	**.933**	.967	.899	.022	.868	.907	.851	.062	**.876**	.891	**.871**	**.067**
cmWS [48]	.900	.914	.886	.044	.915	.945	.870	.027	-	-	-	-	-	-	-	-	-	-	-	-
SSF [49]	.898	.912	.885	.043	.913	.949	.875	.026	.903	.946	.882	.026	-	-	-	-	.858	**.895**	**.866**	**.066**
BBSNet [50]	**.912**	.919	.893	.040	**.920**	.945	.870	.027	.906	.941	.866	.029	.871	.909	.850	.057	.843	.879	.830	.081
LSNet [51]	.911	**.922**	**.900**	**.037**	.918	**.956**	**.885**	**.024**	.925	**.970**	**.910**	**.020**	**.886**	**.927**	**.884**	**.048**	.833	.873	.852	.084
DPLNet (Ours)	**.920**	**.944**	**.904**	**.035**	**.933**	**.962**	**.897**	**.020**	**.940**	**.978**	**.921**	**.017**	**.890**	**.932**	**.888**	**.045**	**.873**	**.909**	.864	**.067**

TABLE IX: Results on RGB-D SOD benchmarks. $\uparrow/\downarrow$ indicates that a larger/smaller value is better.

Model	VT821 [52]				VT1000 [53]				VT5000 [54]			
	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	$M\downarrow$
PoolNet [55]	.751	.739	.578	.109	.834	.813	.714	.067	.769	.755	.588	.089
S2MA [56]	.811	.813	.709	.098	.918	.912	.848	.029	.853	.864	.743	.053
FMCF [57]	.760	.796	.640	.080	.873	.899	.823	.037	.814	.864	.734	.055
ADF [54]	.810	.842	.717	.077	.910	.921	.847	.034	.864	.891	.778	.048
LSNet [51]	**.877**	**.911**	**.827**	**.033**	**.924**	**.936**	**.887**	**.022**	**.876**	**.916**	**.827**	**.036**
DPLNet (Ours)	**.878**	**.908**	**.810**	**.043**	**.928**	**.951**	**.881**	**.022**	**.879**	**.916**	**.828**	**.038**

TABLE X: Results on RGB-T SOD benchmarks. $\uparrow/\downarrow$ indicates that a larger/smaller value is better.

Methods	Learnable Params (M)	mIoU (%)
LMANet [58]	-	52.7
MFNet [10]	8.4	51.6
RTFNet [11]	337.1	52.8
EGFNet [12]	201.3	53.4
MVNet [59]	88.4	**54.5**
DPLNet (Ours)	7.15	**57.9**

TABLE XI: Results on RGB-T video segmentation.

C. Generalization to Other Multimodal Tasks

To show generality of DPLNet, we conduct experiments on other tasks including RGB-D/T salient object detection and RGB-T video semantic segmentation. Note that, we only change our prediction head to adapt to different tasks.

RGB-D salient object detection. For RGB-D salient object detection, we compare our DPLNet with existing competitive methods on five datasets. For evaluation, we follow LSNet [51] and adopt four metrics (*e.g.*, Structure-measure (S) [60], Mean Absolute Error (M) [61], F-measure (F) [62] and E-measure (E) [63]). As shown in Tab. IX, DPLNet achieves SOTA performance on most of the metrics.

RGB-T salient object detection. Tab. X displays results of RGB-T salient object detection on three datasets. As in Tab. X, DPLNet achieves SOTA and competitive performance. LSNet shows better results than DPLNet on some of the metrics. However, it adopts complicated multi-loss supervision to select features, while DPLNet only employs a simple cross-entropy loss, without any special design in architecture, making it more easy to generalize to other tasks.

RGB-T video semantic segmentation. We assess DPLNet on RGB-T video semantic segmentation benchmark [59]. As shown in Tab. XI, our method achieves the best result with 0.579 mIoU, outperforming the second best MVNet with 0.545 mIoU while reducing the trainable parameters by 12$\times$.

V. Conclusion

In this paper, we introduce DPLNet, a simple yet effective framework for training-efficient multimodal semantic segmentation. In DPLNet, we propose a multimodal prompt generator (MPG) module to fuse different modalities, and present a multimodal feature adapter (MFA) module to adapt the frozen pre-trained backbone for better multimodal feature extraction. Our method achieves SOTA or comparable performance on RGB-D/T semantic segmentation. Moreover, DPLNet can be easily generalized to other multimodal tasks such as video object detection and salient object detection.

References

[1] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *NeurIPS*, 2021.

[2] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[3] H. Fan, X. Mei, D. Prokhorov, and H. Ling, "Multi-level contextual rnns with attention model for scene labeling," *IEEE TITS*, vol. 19, no. 11, pp. 3475–3485, 2018.

[4] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *ICCV*, 2021.

[5] B. Cheng, A. Schwing, and A. Kirillov, "Per-pixel classification is not all you need for semantic segmentation," *NeurIPS*, 2021.

[6] 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 *ECCV*, 2020.

[7] 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 *ICCV*, 2021.

[8] 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 TITS*, 2023.

[9] 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 *CVPR*, 2023.

[10] Q. Ha, K. Watanabe, T. Karasawa, Y. Ushiku, and T. Harada, "Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes," in *IROS*, 2017.

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

[12] W. Zhou, S. Dong, C. Xu, and Y. Qian, "Edge-aware guidance fusion network for rgb–thermal scene parsing," in *AAAI*, 2022.

[13] S. Dong, W. Zhou, C. Xu, and W. Yan, "Egfnet: Edge-aware guidance fusion network for rgb–thermal urban scene parsing," *IEEE TITS*, 2023.

[14] 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 TIP*, vol. 30, pp. 7790–7802, 2021.

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

[16] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[17] Q. Zhang, S. Zhao, Y. Luo, D. Zhang, N. Huang, and J. Han, "Abmdrnet: Adaptive-weighted bi-directional modality difference reduction network for rgb-t semantic segmentation," in *CVPR*, 2021.

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

[19] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," *arXiv*, 2021.

[20] M. Jia, L. Tang, B.-C. Chen, C. Cardie, S. Belongie, B. Hariharan, and S.-N. Lim, "Visual prompt tuning," in *ECCV*, 2022.

[21] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *NeurIPS*, 2022.

[22] C. Han, Q. Wang, Y. Cui, Z. Cao, W. Wang, S. Qi, and D. Liu, "Eˆ 2vpt: An effective and efficient approach for visual prompt tuning," in *CVPR*, 2023.

[23] L. Liu, J. Chang, B. X. Yu, L. Lin, Q. Tian, and C.-W. Chen, "Prompt-matched semantic segmentation," *arXiv*, 2022.

[24] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *ECCV*, 2012.

[25] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *CVPR*, 2015.

[26] 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 *ICRA*, 2020.

[27] X. Hu, K. Yang, L. Fei, and K. Wang, "Acnet: Attention based network to exploit complementary features for rgbd semantic segmentation," in *ICIP*, 2019.

[28] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *CVPR*, 2022.

[29] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *ECCV*, 2022.

[30] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," in *CVPR*, 2022.

[31] J. Zhu, S. Lai, X. Chen, D. Wang, and H. Lu, "Visual prompt multi-modal tracking," in *CVPR*, 2023.

[32] J. Yang, Z. Li, F. Zheng, A. Leonardis, and J. Song, "Prompting for multi-modal tracking," in *ACM MM*, 2022.

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

[34] K. Sohn, H. Chang, J. Lezama, L. Polania, H. Zhang, Y. Hao, I. Essa, and L. Jiang, "Visual prompt tuning for generative transfer learning," in *CVPR*, 2023.

[35] 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 TIP*, vol. 30, pp. 2313–2324, 2021.

[36] 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 TMM*, 2022.

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

[38] 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," *arXiv*, 2021.

[39] Y. Sun, W. Zuo, P. Yun, H. Wang, and M. Liu, "Fuseseg: Semantic segmentation of urban scenes based on rgb and thermal data fusion," *IEEE TASE*, vol. 18, no. 3, pp. 1000–1011, 2020.

[40] S. Dong, W. Zhou, X. Qian, and L. Yu, "Gebnet: graph-enhancement branch network for rgb-t scene parsing," *IEEE SPL*, vol. 29, pp. 2273–2277, 2022.

[41] W. Zhou, S. Dong, M. Fang, and L. Yu, "Cacfnet: Cross-modal attention cascaded fusion network for rgb-t urban scene parsing," *IEEE TIV*, 2023.

[42] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*, 2014.

[43] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: A benchmark and algorithms," in *ECCV*, 2014.

[44] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *IMCS*, 2014.

[45] 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 TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2020.

[46] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, "Saliency detection on light field," in *CVPR*, 2014.

[47] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling, "Cross-modal weighting network for rgb-d salient object detection," in *ECCV*, 2020.

[48] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy, "Rgb-d salient object detection with cross-modality modulation and selection," in *ECCV*, 2020.

[49] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu, "Select, supplement and focus for rgb-d saliency detection," in *CVPR*, 2020.

[50] 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 *ECCV*, 2020.

[51] 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 TIP*, vol. 32, pp. 1329–1340, 2023.

[52] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo, "Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach," in *IGTA*. Springer, 2018.

[53] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang, "Rgb-t image saliency detection via collaborative graph learning," *IEEE TMM*, vol. 22, no. 1, pp. 160–173, 2019.

[54] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu, "Rgbt salient object detection: A large-scale dataset and benchmark," *IEEE TMM*, 2022.

[55] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *CVPR*, 2019.

[56] N. Liu, N. Zhang, and J. Han, "Learning selective self-mutual attention for rgb-d saliency detection," in *CVPR*, 2020.

[57] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, "Rgb-t salient object detection via fusing multi-level cnn features," *IEEE TIP*, vol. 29, pp. 3321–3335, 2019.

[58] M. Paul, M. Danelljan, L. Van Gool, and R. Timofte, "Local memory attention for fast video semantic segmentation," in *IROS*, 2021.

[59] 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 *CVPR*, 2023.

[60] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *ICCV*, 2017.

[61] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012.

[62] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014.

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