APNet: Accurate Prompting Network with Modality Guidance and Structural Awareness for RGB-D Semantic Segmentation

Junzhe Zhao, Jintao Su, Jun Liu, *Senior Member, IEEE*, Miaohui Wang, *Senior Member, IEEE*, Ye Liu, *Member, IEEE*

***Abstract*—Parameter-efficient fine-tuning (PEFT) is promising for RGB-D semantic segmentation, as lightweight prompters enable frozen pre-trained RGB backbones to leverage massive RGB pretraining knowledge without full fine-tuning on limited paired RGB-D data. However, existing PEFT methods have two critical limitations: static modal fusion ignores the dynamic reliability of RGB and depth across scenes, leading to suboptimal performance in complex environments; conventional prompts lack structural awareness, causing the loss of edge and texture details essential for dense prediction. To solve these problems, we propose the Accurate Prompting Network (APNet) for precise prompt injection in frozen backbones with two core modules. A Modality Effectiveness Guider (MEG) conducts input-level modal reliability assessment and dynamically generates scene-adaptive modal weights by capturing scene characteristics (e.g., illumination, texture richness). A Structural Awareness Prompter (SAP) injects directional structural priors into prompts via multi-directional gating convolution, endowing prompts with explicit edge and texture information to match semantic segmentation demands. MEG and SAP collaboratively form a precise prompting mechanism that realizes dynamic modal contribution allocation and structural detail preservation, facilitating efficient and accurate cross-modal knowledge transfer to the frozen backbone. Extensive experiments on NYUDv2 and SUN RGB-D show that APNet achieves state-of-the-art mIoU of 59.6% and 52.6% with only 6.2M trainable parameters, realizing a superior trade-off between segmentation accuracy and parameter efficiency.**

***Index Terms*—Semantic segmentation, multimodal learning, prompt-based learning, attention mechanism**

I. Introduction

RGB-D semantic segmentation leverages the complementary strengths of RGB appearance and depth geometry, and plays a vital role in scene understanding for robotics, autonomous driving, and augmented reality. While RGB provides rich texture and color cues, depth offers geometric and spatial information that helps disambiguate cluttered or low-texture scenes. However, effectively integrating these two modalities and achieving high-performance segmentation remains challenging. A common solution is the dual-branch encoder-decoder framework[1], [2], [3], [4], where modality-specific features are extracted and fused through interaction modules. Despite their intuitive design, such models are largely limited by the scale of available paired RGB-D data, which directly restricts the performance ceiling of these dual-branch methods. In contrast, the RGB modality has access to massive amounts of annotated data, providing abundant pretraining knowledge that can be exploited to enhance RGB-D segmentation performance.

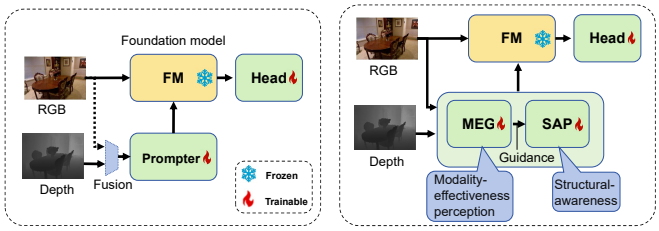

Fig. 1. Left: Previous PEFT framework for RGB-D Segmentation, Right: The proposed Accurate Prompting framework.

Thus recent research has shifted toward Parameter-Efficient Fine-Tuning (PEFT) based approaches. Inspired by advances in PEFT for large language models[5], [6], [7], [8], [9], several studies have extended this paradigm to multimodal segmentation[10], [11], [12]. The core advantage of PEFT lies in its ability to fully exploit the massive RGB pretraining knowledge: by freezing a large pre-trained RGB backbone (which is trained on massive RGB data) and introducing lightweight prompters, these methods can bypass the constraint of limited RGB-D data scale, thereby overcoming the performance bottleneck of traditional dual-branch approaches while maintaining parameter efficiency.

However, as shown in Fig. 1, existing PEFT-based RGB-D prompting methods fail to unleash the full potential of foundation models accurately, with such imprecision manifesting in two key aspects. First, their fixed RGB-depth fusion strategy lacks the ability to perceive the dynamic variation of modal reliability across different scenes, leading to inflexible weight allocation between the two modalities. Second, they overlook the structural differences in distinct regions of images during the prompting process, unable to incorporate region-specific structural cues into prompt generation.

To address these issues, we propose APNet (Accurate

Junzhe Zhao and Jintao Su are with the College of Automation, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: 1224056328@njupt.edu.cn; 1022051501@njupt.edu.cn).

Jun Liu is with the School of Computing and Communications, Lancaster University, LA1 4YW Lancaster, U.K. (e-mail: j.liu81@lancaster.ac.uk).

Miaohui Wang is with Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen Institute of Artificial Intelligence and Robotics for Society, Shenzhen University, Shenzhen 518060, China (email: wang.miaohui@gmail.com).

Ye Liu is with the College of Artificial Intelligence, Nanjing University of Posts and Telecommunications, Nanjing 210023, China (e-mail: yeliu@njupt.edu.cn).

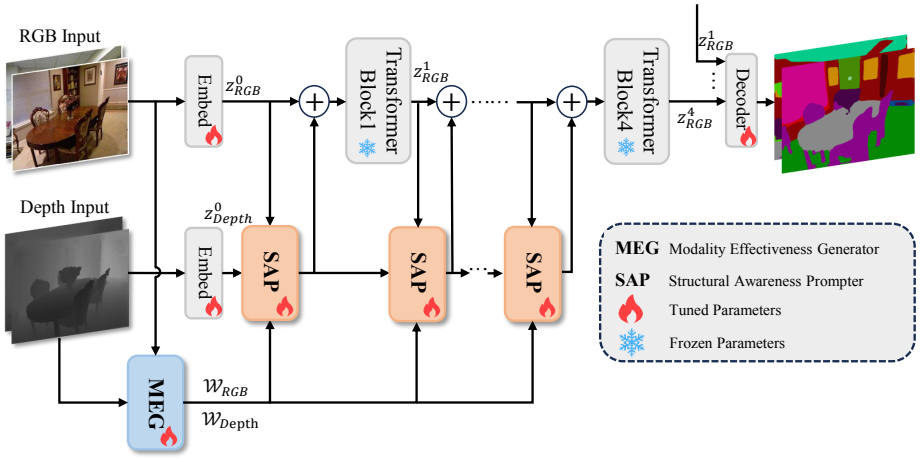

Fig. 2. The APNet structure: MEG produces modality weights to guide SAP, which injects prompts into the Transformer encoder for RGB-depth semantic segmentation.

Prompting Network), a parameter-efficient RGB-D segmentation framework enabling precise, adaptive prompt injection into frozen RGB backbones. We design MEG (Modality Effectiveness Guider) to dynamically generate scene-adaptive RGB-depth weights by perceiving modal reliability variations. We also introduce SAP (Structural Awareness Prompter) to capture regional structural differences and inject directional priors via multi-directional gating convolution. MEG and SAP synergize in APNet: MEG enables flexible cross-modal fusion, while SAP endows prompts with region-aware structural details, unlocking the foundation model's potential for accurate RGB-D segmentation. Experiments on NYUDv2 and SUN RGB-D validate APNet's superiority. Our key contributions are summarized as follows:

- We propose APNet, a novel parameter-efficient RGB-D semantic segmentation framework that enables precise and adaptive prompt injection for frozen pre-trained backbones, solving the problem that existing PEFT-based methods cannot perceive dynamic variations in modal reliability across scenes and lack structural awareness for dense prediction.
- We design the Modality Effectiveness Guider (MEG) to assess input-level modal reliability and dynamically generate scene-adaptive modal weights by capturing scene characteristics, realizing flexible cross-modal contribution allocation.
- We introduce the Structural Awareness Prompter (SAP) that injects directional structural priors into prompts via multi-directional gating convolution, providing explicit edge and texture information to strengthen the structural expressiveness of prompt features.

II. METHOD

A. APNet for RGB-D Semantic Segmentation

APNet is a PEFT-based parameter-efficient RGB-D semantic segmentation framework with a frozen pre-trained RGB backbone, leveraging RGB pretraining knowledge while avoiding limited RGB-D data constraints. As illustrated in Fig. 2, its core is the Structural Awareness Prompter (SAP), guided by the Modality Effectiveness Generator (MEG). MEG produces modality weights to enable scene-adaptive modal priority judgment, and SAP injects structural-aware prompts into the frozen Transformer blocks to dynamically fuse cross-modal information.

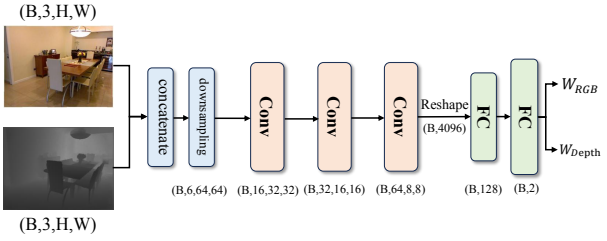

Fig. 3. The structure of MEG takes RGB and depth images as input and generates two effectiveness weights.

B. Modality Effectiveness Guider (MEG)

The contribution of RGB and depth varies across scenes, for instance, depth becomes critical in low-light conditions, while RGB dominates texture-rich scenes. This is often ignored in existing methods which limits performance. We propose the Modality Effectiveness Generator (MEG) which enables scene-adaptive modal priority judgment: it models cross-modal relationships from RGB and depth inputs to dynamically adjust modality weights, fully exploiting appearance and geometric complementarity under varying conditions.

As shown in Fig. 3, RGB and depth images are first concatenated along the channel dimension and downsampled to reduce computation. The input then passes through three blocks, each comprising a 1×1 convolution, ReLU, and max-pooling, to extract hierarchical features. These features are subsequently fed into two fully connected layers, producing a 2D vector (o_1, o_2), which is normalized as:

$$W_{RGB} = o_1/(o_1 + o_2), W_{Depth} \quad = o_2/(o_1 + o_2), \qquad (1)$$

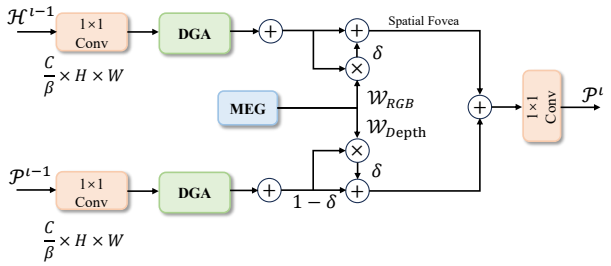

Fig. 4. The structure of SAP, which integrates DGA and MEG to fuse the two modalities to generate dynamic modal prompts.

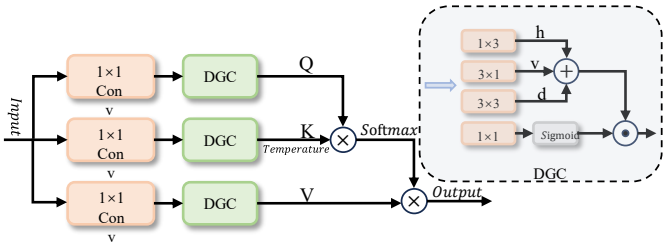

Fig. 5. The structure of DGA, which enhances structural modeling by introducing directional gate and multi-head attention.

ensuring $W_{RGB} + W_{Depth} = 1$ and providing valid weights for adaptive fusion.

C. Structural Awareness Prompter (SAP)

In PEFT-based RGB-D semantic segmentation, multimodal information is essential for guiding frozen RGB backbones. However, existing methods overlook both modality-specific effectiveness and structural cues. To this end, we propose the Structural Awareness Prompter (SAP), which establishes dynamic collaboration between MEG-generated modal weights and Directional Gate Attention (DGA) enhanced structural features. MEG adaptively adjusts modal contributions based on scene properties, DGA enriches structural details, and the Spatial Fovea operation focuses modal weights on structurally critical regions. This design suppresses irrelevant information, facilitates efficient semantic fusion, and better balances accuracy and parameter efficiency.

The network structure of SAP is illustrated in Fig. 4. The RGB and depth tokens from the first stage, denoted as $\mathcal{H}^{l-1}$ and $\mathcal{P}^{l-1}$, are first projected to a lower dimension via 1×1 convolutions:

$$\mathcal{M}_{RGB} = g_1(\mathcal{H}^{l-1}), \quad \mathcal{M}_{Depth} = g_2(\mathcal{P}^{l-1}), \tag{2}$$

where $\mathcal{M}_{RGB}$ and $\mathcal{M}_{Depth}$ have dimensions $[\frac{C}{\beta} \times H \times W]$. In APNet, the input channels C for different stages are 64, 128, 320, and 512, with a scaling factor $\beta = 4$. The projected RGB features are further processed by DGA to enhance structural representations, denoted as $\mathcal{M}^d_{RGB}$.

Modal effectiveness weighting is then applied using weights W_{RGB} and W_{Depth} from MEG. The Spatial Fovea operation is applied on the RGB branch, expressed as:

$$\mathcal{M}^e_{RGB} = \left(\mathcal{M}^d_{RGB} + \delta \cdot W_{RGB} \cdot \mathcal{M}^d_{RGB}\right) \odot \mathcal{M}_{fovea}, \tag{3}$$

$$\mathcal{M}_{fovea} = \sigma\left(\hat{\mathcal{M}}_{RGB} \cdot \lambda\right) \odot \hat{\mathcal{M}}_{RGB}, \tag{4}$$

where σ denotes the softmax function, λ is a learnable smoothness parameter, $\hat{\mathcal{M}}_{RGB}$ is the RGB feature before fovea, and $\delta = 0.05$ controls the intensity of modal weighting. For the depth branch:

$$\mathcal{M}^e_{Depth} = (1-\delta) \cdot \mathcal{M}_{Depth} + \delta \cdot W_{Depth} \cdot \mathcal{M}_{Depth}, \tag{5}$$

Finally, the fused features are restored to the input dimension via a 1×1 convolution:

$$\mathcal{P}^l = g_3\left(\mathcal{M}^e_{RGB} + \mathcal{M}^e_{Depth}\right), \tag{6}$$

where g_3 denotes a 1×1 convolution layer.

D. Directional Gate Attention (DGA)

The core of SAP is the proposed DGA module, which uses a directional gating path with multi-head attention to address traditional attention's limitation in modeling directional information. Applied to both RGB and Depth branches, DGA extracts modality-specific edge and texture cues, equipping prompts with structural priors for subsequent cross-modal fusion.

As illustrated in Fig. 5, given an input $\mathbf{X} \in \mathbb{R}^{B \times C \times H \times W}$, directional responses are extracted via depth-wise convolutions with horizontal, vertical, and diagonal kernels:

$$\mathbf{h} = \text{Conv}_h(\mathbf{X}), \quad \mathbf{v} = \text{Conv}_v(\mathbf{X}), \quad \mathbf{d} = \text{Conv}_d(\mathbf{X}), \tag{7}$$

A learnable gate adaptively aggregates these directional responses:

$$\mathbf{X}' = \sigma(\text{Conv}_{1\times1}(\mathbf{X})) \odot (\mathbf{h} + \mathbf{v} + \mathbf{d}), \tag{8}$$

where $\sigma(\cdot)$ denotes the sigmoid function.

The enhanced features $\mathbf{X}'$ are then fed into an attention mechanism, with both query and key embeddings processed by the directional gated convolution. Attention weights are computed as

$$\mathbf{A} = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} \cdot \tau\right), \tag{9}$$

where τ is a learnable temperature. The final output aggregates values $\mathbf{V}$ with $\mathbf{A}$ and projects them back to the original space.

III. EXPERIMENTS AND ANALYSIS

A. Datasets

To assess the effectiveness of our proposed method, we perform experiments on two widely used RGB-D datasets, namely NYUDv2 [13] and SUN RGB-D [14].

The NYUDv2 [13] dataset contains 1,449 RGB-D images, with 795 samples designated for training and the remaining ones for testing, covering 41 semantic categories including background. The SUN RGB-D [14] dataset is considerably larger, consisting of 10,335 annotated RGB-D images, which are split into 5,285 training pairs and 5,050 testing pairs.

TABLE I
RGB-D SEMANTIC SEGMENTATION ON NYUDv2.

Method	Publication	Backbone	mIoU(%)	Trainable Params(M)
RTLNet[16]	2022-SPL	ResNet-50	53.1	135.6
CMX[3]	2023-TITS	MiT-B4	56.3	139.9
CMX[3]	2023-TITS	MiT-B5	56.9	181.1
CMNext[17]	2023-CVPR	MiT-B5	56.9	119.6
GoPT[11]	2024-AAAI	ViT-Large	54.3	1.0
DPLNet[10]	2024-IROS	MiT-B5	59.0	7.15
DFormer-L[15]	2024-ICLR	DFormer-L	57.2	39.0
GeminiFusion[1]	2024-ICML	MiT-B5	57.7	137.2
SwinMTL[18]	2024-IROS	SwinV2-Base	58.1	87.4
Sigma[19]	2025-WACV	VMamba-S	57.0	69.8
DFormerv2-L[20]	2025-CVPR	DFormerv2-L	58.4	59.5
HDBFormer[21]	2025-SPL	Swin-B	59.3	-
Ours	-	MiT-B5	**59.6**	6.2

TABLE II
RGB-D SEMANTIC SEGMENTATION ON SUN RGB-D.

Method	Publication	Backbone	mIoU(%)	Trainable Params(M)
RTLNet[16]	2022-SPL	ResNet-50	50.6	135.6
PGDENet[22]	2022-TMM	ResNet-34	51.0	100.7
MultiMAE[23]	2022-ECCV	ViT-Base	51.1	95.2
PDCNet[24]	2023-TCSVT	ResNet101	49.6	-
CMX[3]	2023-TITS	MiT-B4	52.1	139.9
CMX[3]	2023-TITS	MiT-B5	52.4	181.1
CMNext[17]	2023-CVPR	MiT-B5	51.9	119.6
GoPT[11]	2024-AAAI	ViT-Large	52.3	1.0
DPLNet[10]	2024-IROS	MiT-B5	52.0	7.15
DFormer-L[15]	2024-ICLR	DFormer-L	52.5	39.0
Sigma[19]	2025-WACV	VMamba-S	52.4	69.8
Ours	-	MiT-B5	**52.6**	6.2

B. Evaluation Metric and Implementation.

Following existing methods, we report the Intersection over Union (IoU) for comparison. All experiments run on a single NVIDIA 3090 GPU with PyTorch 1.12.0. For NYUDv2 [13], the learning rate is 5e-3, weight decay 5e-4. APNet trains for 500 epochs and batch size is 4. For SUN RGB-D [14], the learning rate is 4e-3, weight decay 5e-4; we train for 300 epochs and batch size is 8. As in prior RGB-D semantic segmentation works [15], we use multi-scale flip inference with scales {0.5, 0.75, 1, 1.25, 1.5}.

C. Comparison with State-of-the-art

For RGB-D semantic segmentation, we evaluate the proposed APNet on two widely adopted benchmark datasets: NYUDv2[13] and SUN RGB-D[14], and compare its performance against a comprehensive set of state-of-the-art RGB-D semantic segmentation methods.

As presented in Table I, APNet achieves 59.6% mIoU on the NYUDv2 dataset, significantly outperforming all competing models. Similarly, on the SUN RGB-D dataset, results detailed in Table II show that our APNet delivers a leading performance with an mIoU of 52.6% while keeping only 6.2M trainable parameters. The consistent gains validate the effectiveness of APNet's dynamic modality prompting. Fig. 6 shows qualitative visual comparisons between our method and DPLNet [10] on NYUDv2.

TABLE III
ABLATION ANALYSIS ON NYUDv2.

SAP	DGA	cAcc(%)	mIoU(%)
-	-	74.1	59.0
✓	-	74.7	59.3
-	✓	74.6	59.5
✓	✓	**74.9**	**59.6**

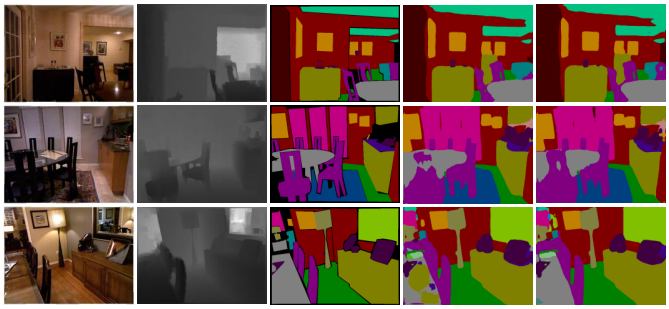

Fig. 6. Qualitative comparisons with DPLNet [10] and Ours on NYUDv2 test set. 'GT' is the ground truth.

D. Ablation Study

We conduct ablation experiments on our proposed APNet on the NYUDv2 [13], and all ablation studies are performed using multi-scale inference.

Effectiveness of DGA. To verify the effectiveness of DGA, we used DGA in the first stage. As shown in Table III, compared with baseline, the performance of mIoU is improved by 0.8% and the performance of cAcc is improved by 0.7%, which shows the necessity of DGA.

Effectiveness of SAP. To validate the effectiveness of SAP, we also conducted an experiment where SAP was used in the first stage without incorporating DGA. As indicated in Table III, this setup yielded a 0.5% improvement in mIoU and a 0.8% enhancement in cAcc compared to the baseline. These results demonstrate the essential role of SAP.

Under the synergistic effect of DGA and SAP, cAcc increased by 1.1% and mIoU increased by 1.0%, which shows the effectiveness of SAP.

IV. CONCLUSION

In this paper, we propose APNet, a PEFT-based RGB-D semantic segmentation framework designed to overcome two limitations of existing prompt methods: static modality fusion and weak structural awareness. Its key Structural Awareness Prompter (SAP) combines MEG for adaptive modality weighting and DGA to inject structural priors through direction-aware filtering. Extensive experiments on NYUDv2 and SUN RGB-D show that APNet achieves state-of-the-art mIoU with only 6.2M trainable parameters, effectively balancing accuracy and efficiency for real-world applications.

REFERENCES

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

[2] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

[3] 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 Transactions on intelligent transportation systems*, vol. 24, no. 12, pp. 14 679–14 694, 2023.

[4] 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 Transactions on Automation Science and Engineering*, vol. 18, no. 3, pp. 1000–1011, 2020.

[5] S. Chen, C. Ge, Z. Tong, J. Wang, Y. Song, J. Wang, and P. Luo, "Adaptformer: Adapting vision transformers for scalable visual recognition," *Advances in Neural Information Processing Systems*, vol. 35, pp. 16 664–16 678, 2022.

[6] C.-H. Lin, J. Gao, L. Tang, T. Takikawa, X. Zeng, X. Huang, K. Kreis, S. Fidler, M.-Y. Liu, and T.-Y. Lin, "Magic3d: High-resolution text-to-3d content creation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 300–309.

[7] K. Zhou, J. Yang, C. C. Loy, and Z. Liu, "Conditional prompt learning for vision-language models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 16 816–16 825.

[8] ——, "Learning to prompt for vision-language models," *International Journal of Computer Vision*, vol. 130, no. 9, pp. 2337–2348, 2022.

[9] N. Kumari, B. Zhang, R. Zhang, E. Shechtman, and J.-Y. Zhu, "Multi-concept customization of text-to-image diffusion," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 1931–1941.

[10] S. Dong, Y. Feng, Q. Yang, Y. Huang, D. Liu, and H. Fan, "Efficient multimodal semantic segmentation via dual-prompt learning," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 14 196–14 203.

[11] Q. He, "Prompting multi-modal image segmentation with semantic grouping," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 38, 2024, pp. 2094–2102.

[12] Y. Liu, P. Wu, M. Wang, and J. Liu, "Cpal: Cross-prompting adapter with loras for rgb+ x semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, 2025.

[13] N. Silberman, D. Hoiem, P. Kohli, and R. Fergus, "Indoor segmentation and support inference from rgbd images," in *European conference on computer vision*. Springer, 2012, pp. 746–760.

[14] S. Song, S. P. Lichtenberg, and J. Xiao, "Sun rgb-d: A rgb-d scene understanding benchmark suite," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 567–576.

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

[16] Y. Yue, W. Zhou, J. Lei, and L. Yu, "Rtlnet: Recursive triple-path learning network for scene parsing of rgb-d images," *IEEE Signal Processing Letters*, vol. 29, pp. 429–433, 2021.

[17] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 1136–1147.

[18] P. Taghavi, R. Langari, and G. Pandey, "Swinmtl: A shared architecture for simultaneous depth estimation and semantic segmentation from monocular camera images," in *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2024, pp. 4957–4964.

[19] Z. Wan, P. Zhang, Y. Wang, S. Yong, S. Stepputtis, K. Sycara, and Y. Xie, "Sigma: Siamese mamba network for multi-modal semantic segmentation," in *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2025, pp. 1734–1744.

[20] B.-W. Yin, J.-L. Cao, M.-M. Cheng, and Q. Hou, "Dformerv2: Geometry self-attention for rgbd semantic segmentation," in *Proceedings of the Computer Vision and Pattern Recognition Conference*, 2025, pp. 19 345–19 355.

[21] S. Wei, Z. Zhou, Z. Lu, Z. Yuan, and B. Su, "Hdbformer: Efficient rgb-d semantic segmentation with a heterogeneous dual-branch framework," *IEEE Signal Processing Letters*, 2024.

[22] 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 Transactions on Multimedia*, vol. 25, pp. 3483–3494, 2022.

[23] R. Bachmann, D. Mizrahi, A. Atanov, and A. Zamir, "Multimae: Multi-modal multi-task masked autoencoders," in *European Conference on Computer Vision*. Springer, 2022, pp. 348–367.

[24] J. Yang, L. Bai, Y. Sun, C. Tian, M. Mao, and G. Wang, "Pixel difference convolutional network for rgb-d semantic segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 34, no. 3, pp. 1481–1492, 2023.