Keep the Balance: A Parameter-Efficient Symmetrical Framework for RGB+X Semantic Segmentation

Supplementary Material

This supplementary material includes the following content.

- More specific implementation details;
- Additional hyper-parameter analysis;
- Robustness analysis and its visualization;
- More comparison results;

1. Implementation Details

Training. We construct a training-friendly pipeline that can be efficiently trained on a workstation equipped with a single NVIDIA RTX 3090 using PyTorch 1.9.1. Following the practice of [10], data augmentation is performed by random resize with ratio range from 0.5 to 2.0, random horizontal flipping, random color jitter, and random gaussian blur. Additionally, following many existing works (e.g., [1, 9, 10]), we adopt multi-scale testing augmentations with the scales $\{0.5, 0.75, 1, 1.25, 1.5, 1.75\}$ and horizontal flip. We select the AdamW optimizer with weight decay 1e-2. The batch size is set to 4 for all datasets. For the NYUDv2 dataset [6], the learning rate is set to 4e-4, and the fine-tuning process spans 400 epochs. Regarding the SUN RGB-D dataset [7], we adjust the learning rate to 1e-4 and randomly crop the images to 480×480 pixels, training the network for a total of 300 epochs. On the MFNet dataset [2], we employ a learning rate of 8e-5 and conduct 200 epochs of training. For the PST900 dataset [5], the learning rate is established at 1e-4, and the images are randomly cropped to 480×640, with the model being trained for 200 epochs. As for the MCubeS dataset [4], we set the learning rate to 4e-4 and randomly crop the images to 512×512, completing the training over 250 epochs. Finally, on the DeLiVER dataset [10], we adopt a learning rate of 1e-4, with images being randomly cropped to 768×768, and the training lasts for 200 epochs. In the GoPT [3] paper, the parameters of its segmentation head (i.e., SETR [11]) are not counted. For a fair comparison, we have taken these parameters into account.

Inference. During inference, we do not employ the Masked Modality Self-Teaching (MMST) strategy, meaning we input clean data without masking any modality and discard the auxiliary heads S_2 and S_3 used for single-modality segmentation. We measure the inference time, which is tested for a 480×640 image on the same hardware device, *i.e.*, a workstation with an NVIDIA RTX 3090 GPU.

$\tilde{X}_i$ rank	Params(M)	mIoU(%)
[32, 64, 128, 256]	8.2	57.4
[16, 32, 64, 128]	7.5	59.0
[8, 16, 32, 64]	7.3	58.3

Table 1. Ablation studies of $\tilde{X}_i$ rank in MAPA module. The i in $\tilde{X}_i$ denotes different modalities (RGB or X). Our version is highlighted in gray.

Adapter rank	Params(M)	mIoU(%)
[64, 64, 64, 64]	8.9	55.7
[16, 32, 64, 128]	9.2	55.8
[8, 16, 32, 64]	7.5	59.0
[4, 8, 16, 32]	6.7	58.1

Table 2. Ablation studies of Adapter rank in MAPA module.

DSCF rank	Params(M)	mIoU(%)
[32, 64, 128, 256]	9.6	58.1
[16, 32, 64, 128]	7.5	59.0
[8, 16, 32, 64]	6.9	57.1

Table 3. Ablation studies of DSCF rank.

2. Hyper-parameter Analysis

We conduct additional hyper-parameter analysis experiments on the NYUDv2 dataset using the Swin-B as the backbone, with channel dimensions of $\{128, 256, 512, 1024\}$ across four stages.

Low-Rank: To achieve an effective and efficient model, low-rank latent space projection plays a vital role in our model, influencing our Modality-Aware Prompting and Adaptation (MAPA) module and Dynamic Sparse Cross-modality Fusion (DSCF) module. Therefore, the choice of rank is critical for both the efficiency and effectiveness of our approach. In Tables 1, 2 and 3, we systematically explore the impact of different low-rank choices within each component. Table 1 presents the ablation study on the low-rank selection of $\tilde{X}_i$ during the generation of unifying dual-modality features. Lower ranks lead to poorer performance due to insufficient information captured by sparse representations. Conversely, higher ranks, which capture more modality-specific details, complicate the identification of

λ_2	Params(M)	mIoU(%)
1	7.5	57.6
0.1	7.5	58.8
0.01	7.5	59.0
0.001	7.5	57.8
0	7.5	57.3

Table 4. Ablation studies of λ_2 in overall loss.

Model	RGB + Depth	RGB	Depth	RGB + Depth†
Ours	59.0	55.9	41.0	57.8

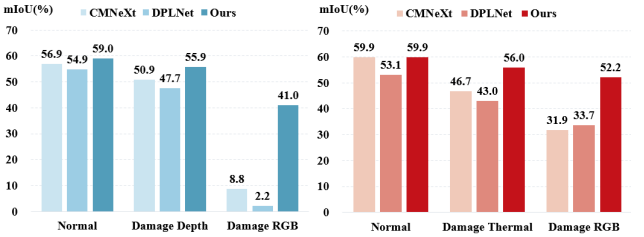
Table 5. Overall performance with different modalities. †: Depth generated from RGB using Depth Anything V2 [8].

unifying features. As shown in Tables 2 and 3, similar trends are observed when investigating the low-rank choices for adapters and DSCF, as low ranks struggle to capture essential information, while higher ranks tend to result in overfitting. These experiments emphasize the importance of selecting the best low-rank representations.

Balance Weight: Balance weight λ_2 determines the weight applied to the loss $\mathcal{L}_{S_2}$ and $\mathcal{L}_{S_3}$. As shown in Table 4, setting λ_2 too high causes single-modal features to overfit cross-modal information, losing modality-specific details and resulting in minimal gains. Conversely, setting λ_2 too low leads to underfitting, with limited performance improvement. We set $\lambda_2 = 0.01$ for optimal performance.

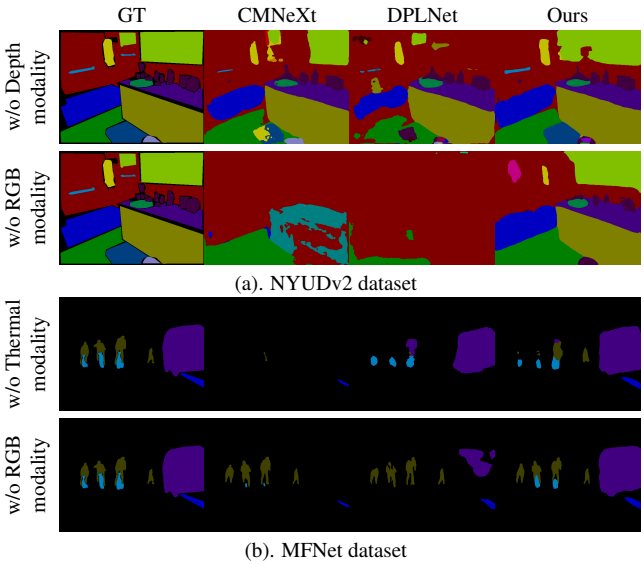
3. Robustness Analysis

Our proposed Masked Modality Self-Teaching (MMST) strategy mitigates the model's over-reliance on a single modality. Consequently, we find that this strategy enhances the robustness of our model in challenging scenarios, such as sensor damage. We conduct experiments under various challenging conditions, e.g., the absence of X modality images or the absence of RGB images. As shown in Fig. 1, our model demonstrates superior robustness compared to CMNeXt [10] and DPLNet [1], especially when the RGB modality is absent. To more intuitively illustrate the robustness of our model in the above challenging scenarios, As illustrated in Fig. 2, we observe that, when the RGB sensor is damaged, competing methods struggle to recognize objects within the scene. In contrast, our model can still identify most objects. This confirms the robustness of our approach and highlights its potential for applications where one camera in a dual-modality imaging device fails. As shown in Table 5, we further test the performance of our method using accurately generated depth maps from Depth Anything V2 [8]. This demonstrates that the proposed method can



(a) NYUDv2 dataset (b) MFNet dataset

Figure 1. The performance comparison in challenging conditions on the NYUDv2 [6] and MFNet [2] datasets.



(a). NYUDv2 dataset

(b). MFNet dataset

Figure 2. Qualitative visual comparison with CMNeXt [10], and DPLNet [1] in modality-incomplete scenarios.

effectively work even with less accurate modality information, such as depth data generated by monocular depth estimation methods, thereby enhancing the practical relevance of the approach.

4. More Comparison Results

In Figs. 3, 4, 5, and 6, we showcase more comparison results from NYUDv2 [6], MFNet [2], MCubeS [4], DeLiVER [10], against the CMNext [10]. As observed, our approach outperforms the competing method. Notably, our model excels in segmenting small objects in nighttime environments (e.g., cars on the first three rows of Fig. 4) and demonstrates superior accuracy in handling complex scenes and textures (e.g., the first row of Fig. 3 and the last two rows of Fig. 5), indicating the superior segmentation ability of our model. As shown in Table 4, we test the performance of our method using accurately generated depth maps from DepthAnything. We will incorporate this in the revision.

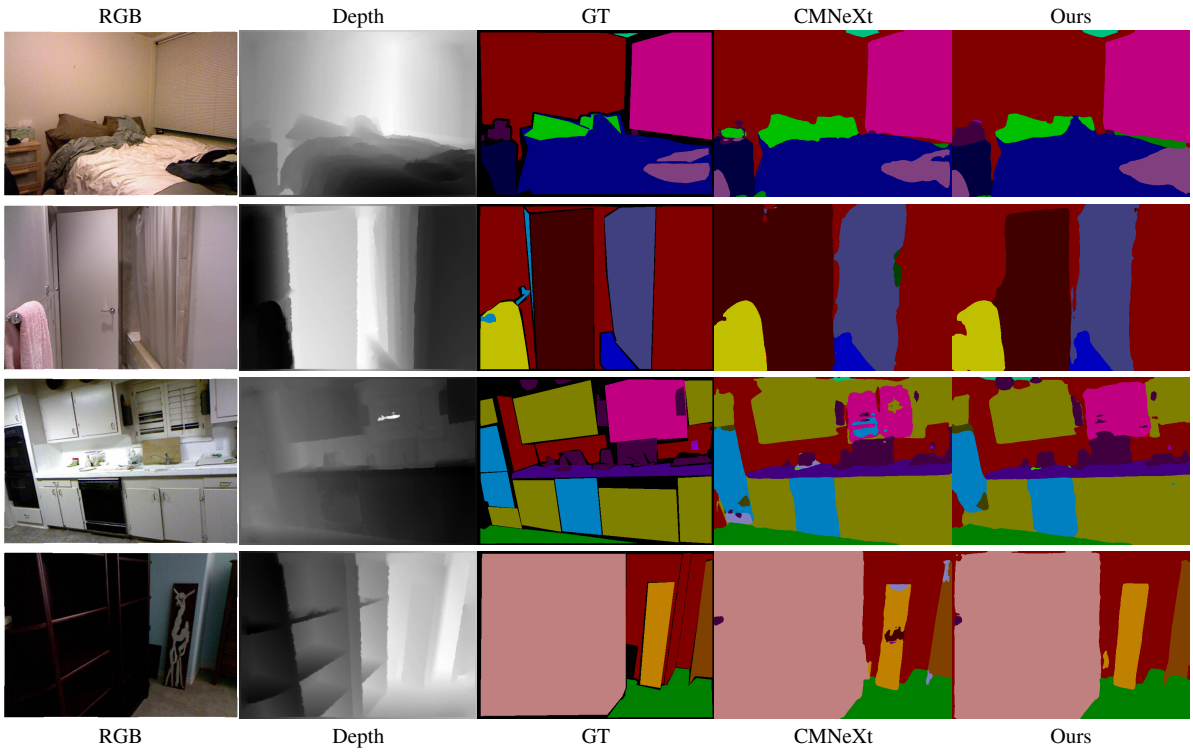


Figure 3. We illustrate several examples from NYUDv2 [6], comparing our method against CMNeXt [10].

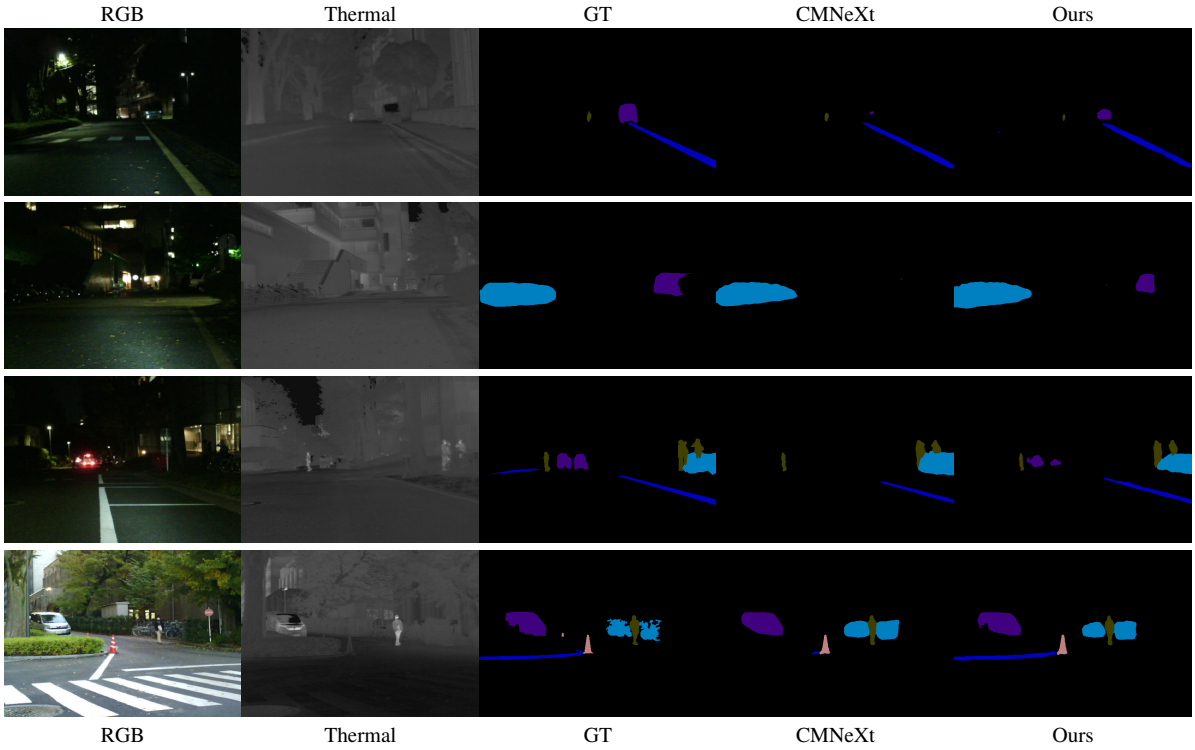


Figure 4. We illustrate several examples from MFNet [2], comparing our method against CMNeXt [10].

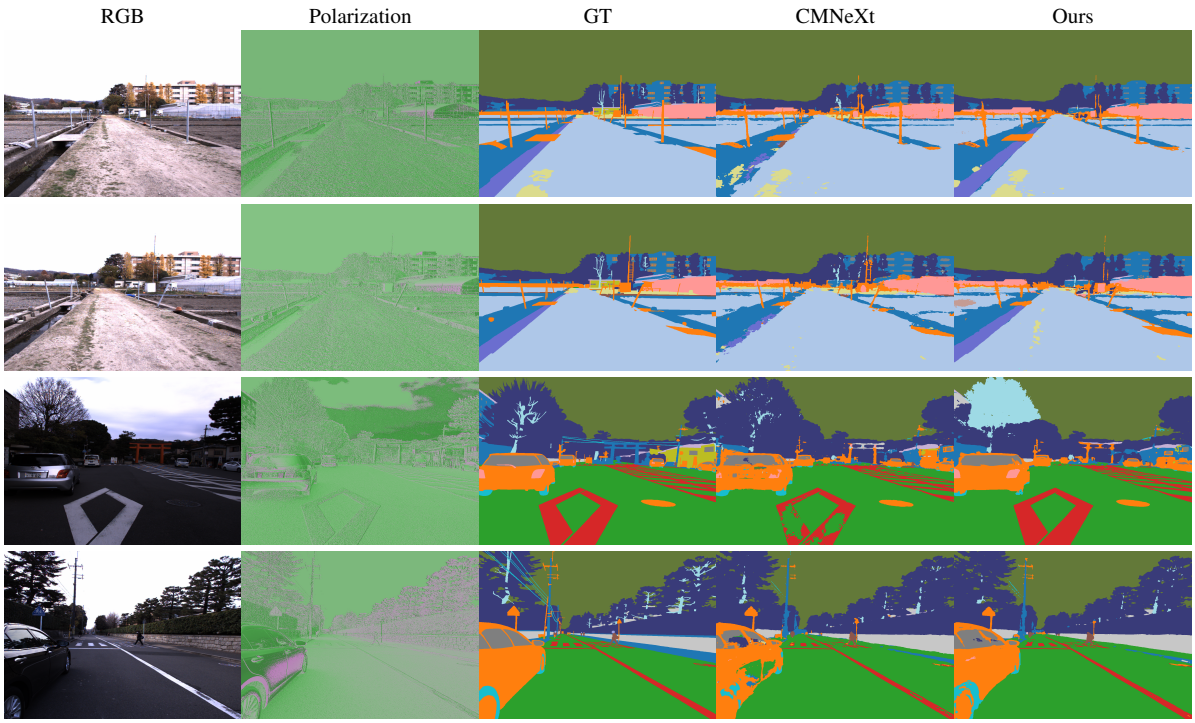


Figure 5. We illustrate several examples from MCubeS [4], comparing our method against CMNeXt [10].

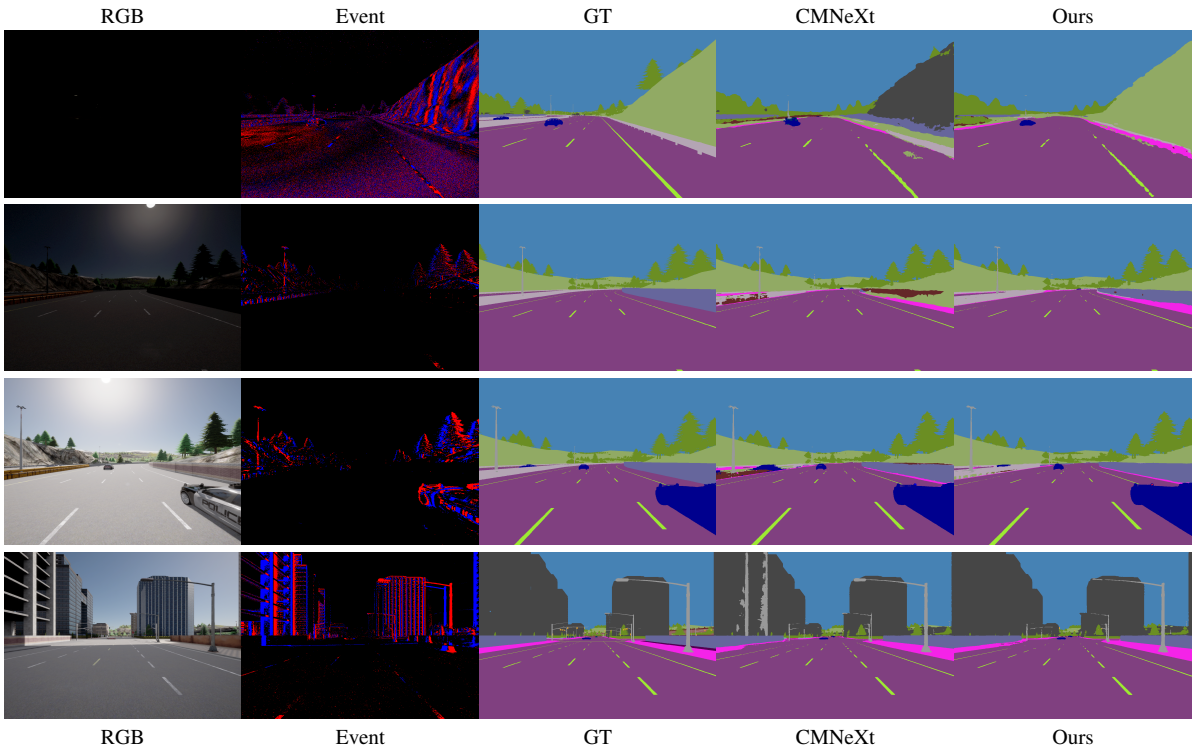


Figure 6. We illustrate several examples from DeLiVER [10], comparing our method against CMNeXt [10].

References

[1] Shaohua Dong, Yunhe Feng, Qing Yang, Yan Huang, Dongfang Liu, and Heng Fan. Efficient multimodal semantic segmentation via dual-prompt learning. *arXiv preprint arXiv:2312.00360*, 2023. 1, 2

[2] Qishen Ha, Kohei Watanabe, Takumi Karasawa, Yoshitaka Ushiku, and Tatsuya Harada. Mfnet: Towards real-time semantic segmentation for autonomous vehicles with multi-spectral scenes. In *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5108–5115. IEEE, 2017. 1, 2, 3

[3] Qibin He. Prompting multi-modal image segmentation with semantic grouping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 2094–2102, 2024. 1

[4] Yupeng Liang, Ryosuke Wakaki, Shohei Nobuhara, and Ko Nishino. Multimodal material segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19800–19808, 2022. 1, 2, 4

[5] Shreyas S Shivakumar, Neil Rodrigues, Alex Zhou, Ian D Miller, Vijay Kumar, and Camillo J Taylor. Pst900: Rgb-thermal calibration, dataset and segmentation network. In *2020 IEEE international conference on robotics and automation (ICRA)*, pages 9441–9447. IEEE, 2020. 1

[6] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part V 12*, pages 746–760. Springer, 2012. 1, 2, 3

[7] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 567–576, 2015. 1

[8] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *Advances in Neural Information Processing Systems*, 37:21875–21911, 2024. 2

[9] Bowen Yin, Xuying Zhang, Zhong-Yu Li, Li Liu, Ming-Ming Cheng, and Qibin Hou. Dformer: Rethinking rgbd representation learning for semantic segmentation. In *The Twelfth International Conference on Learning Representations*. 1

[10] Jiaming Zhang, Ruiping Liu, Hao Shi, Kailun Yang, Simon Reiß, Kunyu Peng, Haodong Fu, Kaiwei Wang, and Rainer Stiefelhagen. Delivering arbitrary-modal semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1136–1147, 2023. 1, 2, 3, 4

[11] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6881–6890, 2021. 1