MPerS: Dynamic MLLM MixExperts Perception-Guided Remote Sensing Scene Segmentation

Supplementary Material

1. Effect of Different Text Encoder backbones

In our text encoding module, we not only utilize the CLIP text encoder [5] with a ViT-B backbone, but also investigate other encoders, including BILP [2], CLIP-L, and RemoteCLIP [4], with RemoteCLIP specifically designed for remote sensing (RS) tasks. Notably, we further explore whether text encoders trained with text alignment on the DINOv3 dataset can provide semantic textual information that more effectively guides the model's visual features. Table 1 presents the quantitative analysis of different Vision-Language (VL) text encoder backbones on the Vaihingen dataset.

Table 1. Qualitative comparison with other vision–language text-encoder methods on the Vaihingen dataset. The visual backbone uses the DINOv LVD-0.3B distilled weights.

Method	VL backbone	Architecture	IoU%		OA	mIoU	mF1
			building	car			
MPerS	-	-	80.56	68.52	87.95	74.68	85.12
MPerS	BLIP	FLAN-T5	89.34	75.12	88.33	76.42	86.3
MPerS	RemoteCLIP	ViT-B	92.29	**83.35**	88.41	76.91	86.67
MPerS	DINOv3-Text	ViT-L	94.12	70.46	88.51	76.98	86.68
MPerS	CLIP	ViT-L	94.48	70.69	88.50	76.83	86.61
MPerS(Ours)	CLIP	ViT-B	**94.49**	72.09	**88.52**	**77.10**	**86.74**

From the table, we observe that, although RemoteCLIP is trained on RS data, it does not surpass the performance of CLIP ViT-B. This may be attributed to the relatively low image resolution employed during training or to the possibility that its training data do not comprehensively cover all general RS scenarios. Furthermore, DINOv3-Text [7] does not exhibit superior performance and continues to show notable limitations in its capacity to interpret complex sentence-level captions.

2. Visualization of different Caption feature maps

We visualize the heatmaps corresponding to different categories on the Vaihingen dataset. These generated heatmaps offer an intuitive means to compare how captions of varying quality affect the guidance and fusion of the model's visual features, as detailed in Section 4.4.2 of the main paper.

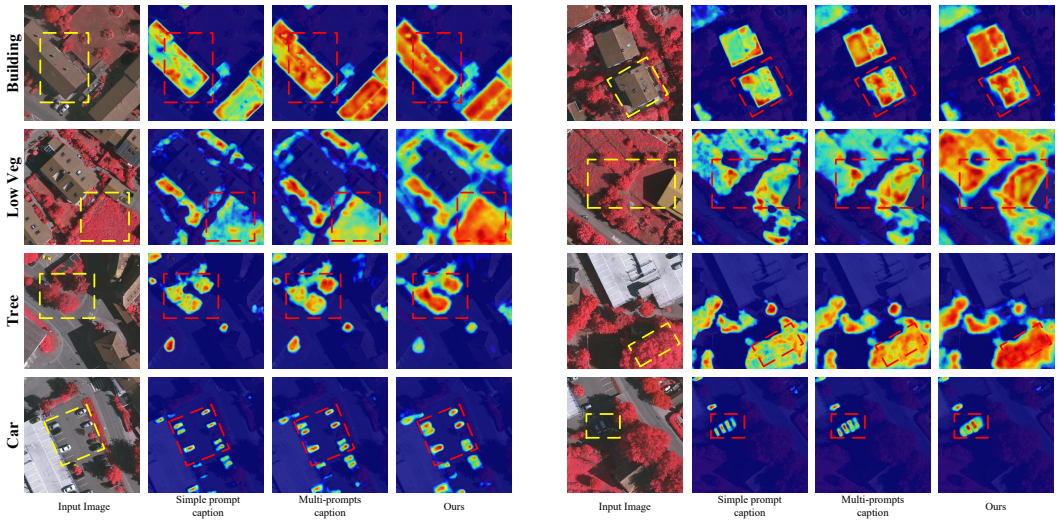

Figure 1. Qualitative visual comparison with state-of-the-art methods on the Vaihingen dataset. Dashed bounding boxes indicate regions where our model produces more precise segmentation results.

As illustrated in Fig. 1, compared to simple prompt captions, multi-prompts captions provide superior guidance and facilitate more effective fusion of the model's visual features. Furthermore, it is evident that when a single MLLM is employed to generate perceptual captions, the model exhibits certain limitations in extracting comprehensive scene information. In contrast, leveraging multiple MLLMs enables a more holistic capture of the elements within a RS scene from diverse expert visual perspectives, thereby allowing more precise guidance for segmentation.

3. Additional Experiments Analysis

3.1. Vision Dilateformer Adapter Detial

The DilateFormer [1] incorporated in our framework serves, to a certain extent, as an adaptation module for DINOv3, improving its suitability and effectiveness for RS tasks. To reduce model complexity and enhance computational efficiency, the DilateFormer is instantiated with a single dilation configuration. As illustrated in Fig. 2, we present the design details of the single-dilation DilateFormer and provide a comparison with the Multi-Head Self-Attention (MHSA).

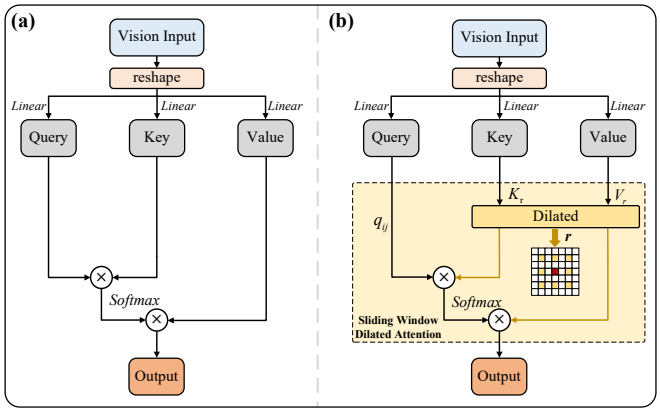

Figure 2. Comparison of different Transformer adapter structures: (a) multi-head self-attention, (b) single-scale dilated Vision Transformer.

The DilateFormer utilizes a sliding window dilated attention mechanism, in which the query q_{ij} at position (i, j) in the original feature map sparsely samples its corresponding keys K_{ij}^r and values V_{ij}^r o perform self-attention within the defined window. The corresponding mathematical formulation is given as follows:

$$y_{ij}^r = \text{Softmax}(\frac{q_{ij}(K_{ij}^r)^T}{\sqrt{d_k}})V_{ij}^r \tag{1}$$

where r denotes the dilation coefficient. We further conduct a comprehensive ablation study, as presented in Table. 2, to analyze the impact of different dilation r and identify the optimal receptive field range. Additionally, we compare with conventional multi-head self-attention adapters, which further validates the effectiveness of the proposed approach.

Table 2. Qualitative comparison with Transformer-based methods on the Vaihingen dataset. The variable r denotes the dilation parameter.

Method	Vision Transformer	Per-class IoU(%)/F1(%)					OA(%)	mIoU(%)	F1(%)
		Impervious surfaces	Buildings	Low vegetation	Trees	Cars			
MPerS	MHSA	80.62/89.27	89.40/94.40	62.91/77.23	78.58/88.00	70.92/82.98	88.26	76.49	86.49
MPerS	dialteformer r=1	**81.02/89.52**	**89.55/94.49**	**63.76/77.87**	**79.08/88.32**	**72.09/83.78**	**88.54**	**77.10**	**86.79**
MPerS	dialteformer r=2	80.65/89.29	89.28/94.34	63.66/77.80	78.77/88.13	70.86/82.95	88.37	76.65	86.50
MPerS	dialteformer r=3	80.88/89.43	89.54/94.48	63.44/77.63	78.72/88.10	71.06/83.08	88.43	76.73	86.54

3.2. MLLMs' Perception of Caption Details

As required in Section 3.3.2 of the main text, multi-prompts need to satisfy three criteria: (i) clearly indicate the presence of a wide range of land-cover objects, (ii) describe the proportional relationships among different objects, and (iii) explicitly instruct the MLLM to describe the relationships between the objects. The multi-prompts for different MLLMs were slightly adjusted according to the input requirements. The specific multi-prompt instructions are provided as follows:

- Prompt1: This image contains the following land cover types: $< object{-}class_1 >$,$< object{-}class_2 >$,...,$< object{-}class_n >$.
- Prompt2: This image contains: $< object{-}class_1 > (\omega_1\%)$, $< object{-}class_2 > (\omega_2\%)$,..., $< object{-}class_n > (\omega_n\%)$
- Prompt3: Please describe the RS image in natural language in one concise paragraph, including its approximate location and spatial relationships.

where n denotes the number of land-cover categories present in the current detected RS image, and ω_n represents the proportion of each category. We use captions generated by the MLLM to compute the similarity between CLIP image features and those in this work, and normalize the values to the range $[0, 1]$. This allows for a more intuitive comparison of effectiveness. The specific formulation is as follows:

$$\text{Similarity}(f^v, \Phi^t) = \frac{f^v \cdot \Phi^t}{2 \left\| f^v \right\| \left\| \Phi^t \right\|} + \frac{1}{2} \tag{2}$$

where f^v corresponds to the visual feature vector produced by the CLIP image encoder, while Φ^t denotes the feature vector derived from the CLIP text encoder. As show in the check flowchart in the main text, Vision-language similarity value below the threshold τ, this suggests that the MLLM has limited perception of the current scene. To mitigate this limitation, we employ a prompt fine-tuning strategy to regenerate higher-quality captions that are more accurately aligned with the scene.

3.3. Decoder Analysis

Table 3 presents the model's evaluation under different decoding methods. Specifically, pixelUp reconstructs features using only deep semantic convolutions and upsampling, without leveraging skip connections or intermediate features from the visual encoder. In contrast, FPN [3] and U-Net [6] employ the standard decoding strategies commonly used in prior work.

Table 3. Qualitative comparison with different decoder methods on the Vaihingen dataset.

Decoder Method	Per-class IoU(%)/F1(%)					OA(%)	mIoU(%)	F1(%)
	Impervious surfaces	Buildings	Low vegetation	Trees	Cars			
MPerS+pixelUp	80.00/88.89	88.24/93.75	60.94/75.73	78.50/87.95	66.75/80.06	87.86	74.88	85.28
MPerS+FPN	80.29/89.06	89.32/94.36	**64.57/78.21**	76.81/86.65	70.61/82.77	88.23	76.32	86.21
MPerS+Unet (Ours)	**81.02/89.52**	**89.55/94.49**	63.76/77.87	**79.08/88.32**	**72.09/83.78**	**88.54**	**77.10**	**86.79**

It can be observed that the U-Net decoding structure adopted in this work achieves strong performance, further highlighting the importance of multi-scale information for remote sensing tasks. However, the decoder in this work is relatively simple. In future work, a more refined decoder design is expected to achieve even better performance.

3.4. Analysis of Language-Guided Positions

We design an interesting investigation to examine how textual semantic information from captions perceived by the MLLM guides visual features across different stages, highlighting potential variations in its effectiveness. Table 4 demonstrates that, overall, textual semantic information provides guidance for the fusion of visual features across different stages. However, its influence exhibits notable variation depending on the stage. The guidance effect of the Dynamic MixExperts Text Encoder (DMTE) on shallow visual features is limited, likely because salient land-cover features have not yet been extracted at the early stages. In contrast, the deepest stage exhibits the most effective guidance.

Table 4. Qualitative comparison of DMTE in guiding visual features at different stages on the Vaihingen dataset.

module	Stage 1	Stage 2	Stage 3	Stage 4	OA	mIoU	mF1
DMTE	✓				88.29	76.22	86.21
DMTE		✓			88.49	76.83	86.61
DMTE			✓		88.41	76.49	86.40
DMTE (Ours)				✓	**88.54**	**77.10**	**86.75**
DMTE			✓	✓	88.51	76.98	86.68

4. Additional Qualitative Visualizations

In this section, we provide qualitative comparisons with several state-of-the-art (SOTA) models on the Vaihingen, Potsdam, and SynDrone datasets, as illustrated in Figures 3, 4, and 5, respectively.

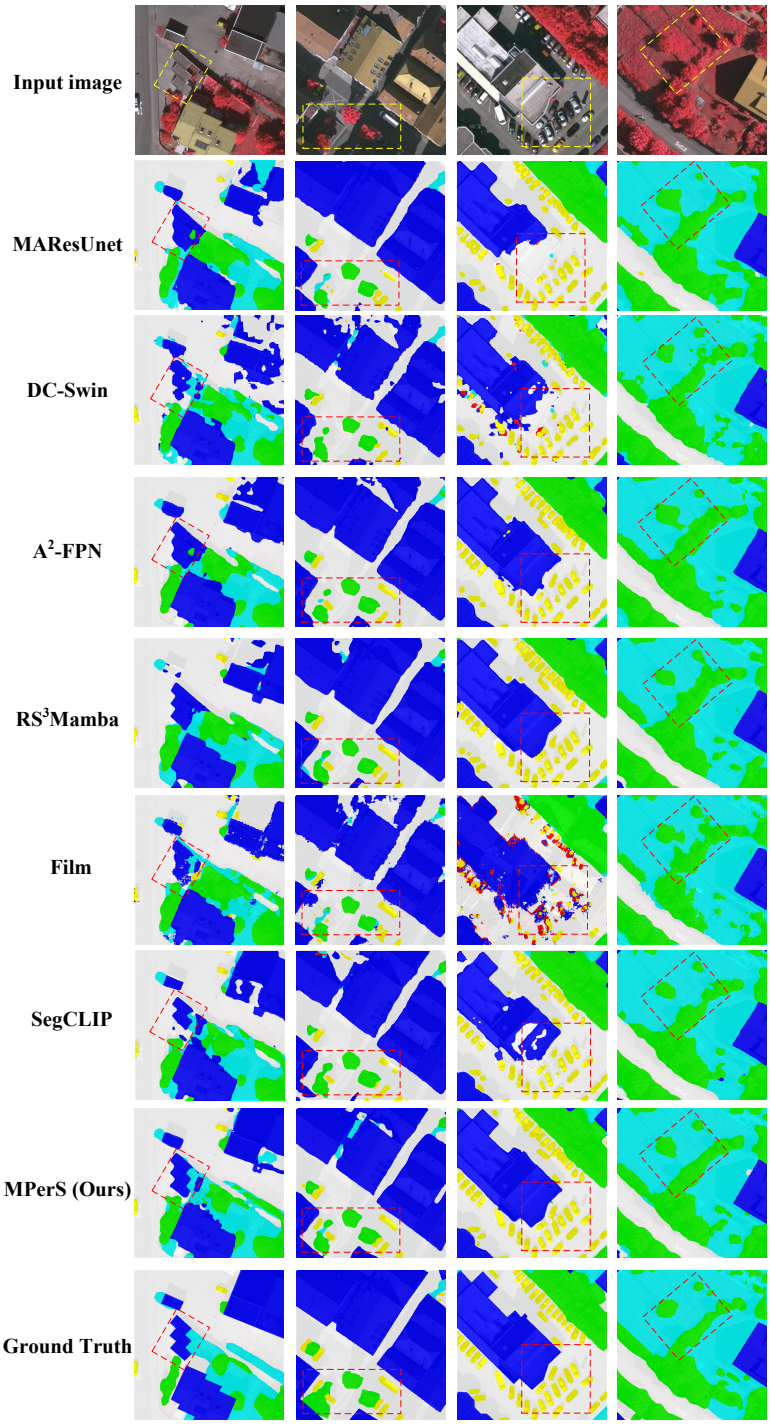

Figure 3. Qualitative visual comparison with state-of-the-art (SOTA) methods on the Vaihingen dataset. Dashed bounding boxes indicate regions where our model produces more precise segmentation results.

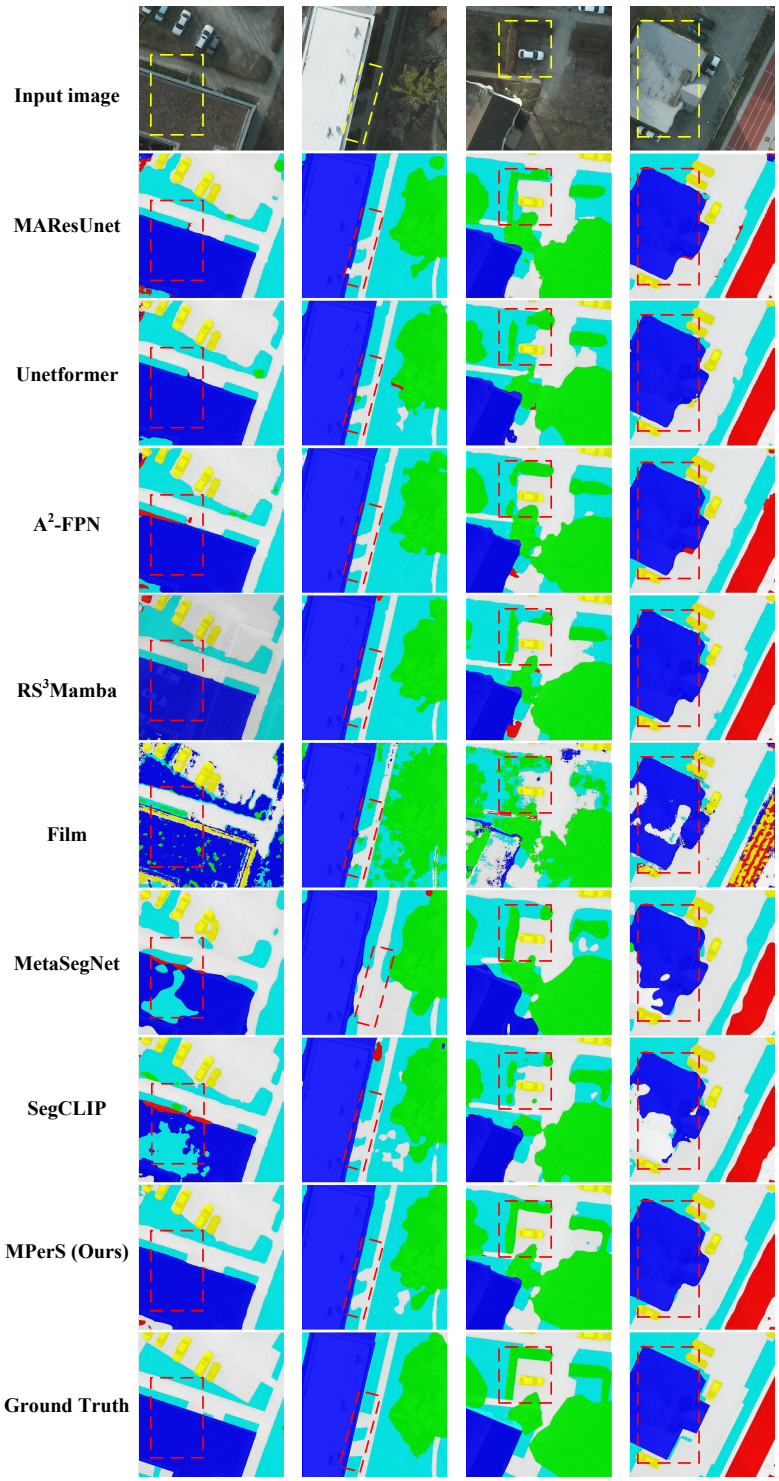

Figure 4. Qualitative visual comparison with state-of-the-art methods on the Potsdam dataset. Dashed bounding boxes indicate regions where our model produces more precise segmentation results.

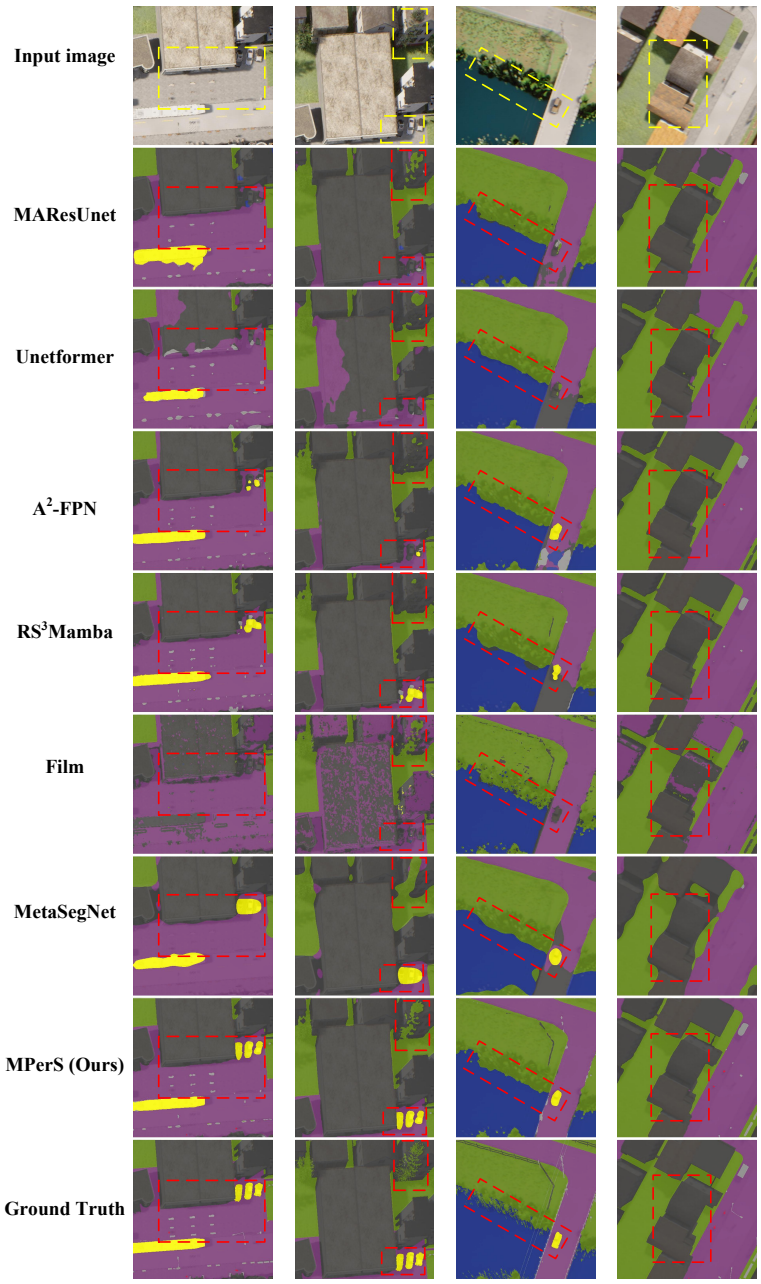

Figure 5. Qualitative visual comparison with state-of-the-art methods on the SynDrone dataset. Dashed bounding boxes indicate regions where our model produces more precise segmentation results.

References

[1] Jiayu Jiao, Yu-Ming Tang, Kun-Yu Lin, Yipeng Gao, Andy J Ma, Yaowei Wang, and Wei-Shi Zheng. Dilateformer: Multi-scale dilated transformer for visual recognition. *IEEE transactions on multimedia*, 25:8906–8919, 2023.

[2] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742, 2023.

[3] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.

[4] Fan Liu, Delong Chen, Zhangqingyun Guan, Xiaocong Zhou, Jiale Zhu, Qiaolin Ye, Liyong Fu, and Jun Zhou. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62: 1–16, 2024.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763, 2021.

[6] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241, 2015.

[7] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025.