Test-Time Multi-Prompt Adaptation for Open-Vocabulary Remote Sensing Image Segmentation

Supplementary Material

In the supplementary materials, we first present the complete example associated with Fig. 3 in Section 3.2. We further provide extended experimental analyses, including more qualitative visualizations, ablation studies on the adaptation steps during inference, the choice of large language models (LLMs) for text description generation in Cat-Prompt, and the number of low-entropy visual features per category in VGTA, as well as the evaluation of our TMPA on natural image datasets.

S6. Supplementary Example for Cat-Prompt

We present an example of generating single-sentence text descriptions for the WHU$^{\text{Aerial}}$ dataset [24], as illustrated in Fig. S7. Descriptions derived solely from class labels often lack visual specificity (upper). While incorporating general visual constraints can enhance detail, these attributes may be inconsistent with the target scene; for instance, 'circular' structures or 'white' colors do not reflect the typical appearance of buildings in Wuhan (middle). In contrast, by integrating these visual constraints with dataset-specific, scene-aware text, the resulting descriptions align well with the visual characteristics of the target scene (bottom). This example demonstrates that Cat-Prompt generates context-aware, visually informative text descriptions that are well-suited to the target dataset.

S7. Additional Experiments

S7.1. Additional Predictive Entropy Visualization

We provide more qualitative examples of predictive entropy analysis to further illustrate the effect of prompt bias. As shown in Fig. S8, we visualize the predictive entropy computed from the similarity between visual features and text embeddings. As shown in the first two rows, without prompt bias, correctly predicted regions such as cars generally exhibit low entropy (high confidence), whereas the surrounding pavement that is mistakenly classified as cars shows high entropy (low confidence). After applying prompt bias to refine the text embeddings using reliable visual features, these misclassified pavement regions around cars are effectively corrected, and the segmentation boundaries become clearer and more precise. As illustrated in the last two rows, certain pavement regions are incorrectly predicted as buildings, and these regions exhibit high entropy, reflecting prediction uncertainty. After applying prompt bias to calibrate the text embeddings, these erroneous predictions are effectively mitigated, leading to more accurate segmentation results. These examples demonstrate that TMPA, by leveraging high-confidence visual features to refine text embeddings, can effectively reduce uncertainty in ambiguous regions and improve the overall reliability of segmentation.

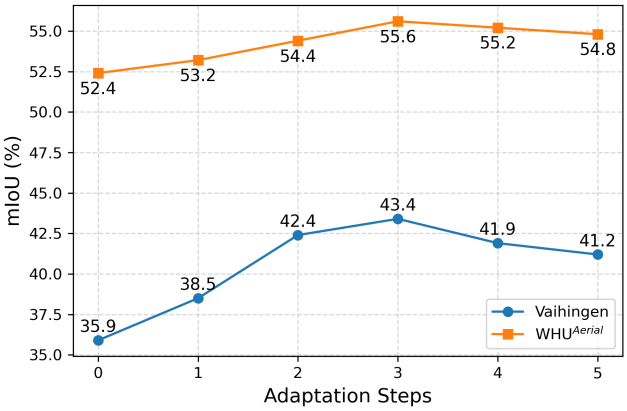

Figure S6. Results of our TMPA with various adaptation steps during inference on Vaihingen and WHU$^{\text{Aerial}}$ [24].

S7.2. Additional Qualitative Results

We present additional qualitative results on UDD5 [10], VDD [6], UAVid [37], and Vaihingen. As shown in Fig. S9, the predictions from our TMPA exhibit superior region connectivity and more accurate segmentation masks. For instance, TMPA achieves a more complete segmentation of the 'building' region (first row). Additionally, Segearth-OV misclassifies background as 'road' (third row, left) and incorrectly labels part of an impervious surface as 'car' (fourth row), whereas our TMPA delivers significantly more precise segmentation in these cases. Furthermore, TMPA successfully handles fine-grained prompts such as 'roof' and 'facade' (second row), showcasing the potential of OVRSIS to move beyond class-level segmentation towards component-level understanding of remote sensing imagery. These qualitative examples underscore the effectiveness of our method, confirming that TMPA provides an effective solution to mitigate textual ambiguity and achieve superior segmentation performance.

S7.3. Additional Ablation Study

Influence of adaptation steps during inference. As illustrated in Fig. S6, segmentation performance exhibits an initial upward trend as the number of adaptation steps in-

Example

Please generate one descriptive sentence about the category ‘Building’.

From humble shelters to towering skyscrapers, a building is a structure crafted by human hands to enclose space, provide sanctuary, and stand as a tangible marker of civilization on the landscape.

You are an expert in computer vision and remote sensing. For the category ‘building’ , generate a detailed and realistic one-sentence description (fewer than 20 words) that reflects how the category visually appears in the dataset’s context, based on its characteristics in overhead imagery.

Circular and curved architectural forms with bright white surfaces contrasting against surrounding darker pavement textures.

You are an expert in computer vision and remote sensing. The WHU_Aerial dataset is a large-scale, high-resolution remote sensing dataset captured over Wuhan, China, containing detailed annotations for building extraction tasks with diverse urban and suburban architectural styles. For the category ‘building’ in the WHU_Aerial dataset, generate a detailed and realistic one-sentence description (fewer than 20 words) that reflects how the category visually appears in the dataset’s context, based on its characteristics in overhead imagery.

Buildings appear as densely clustered rectangular or square structures with varied roof colors and heights, often arranged along roads and interspersed with small green spaces.

Figure S7. Examples of generated textual descriptions for the ‘building’ category in WHU$^{\text{Aerial}}$ [24] under different prompts.

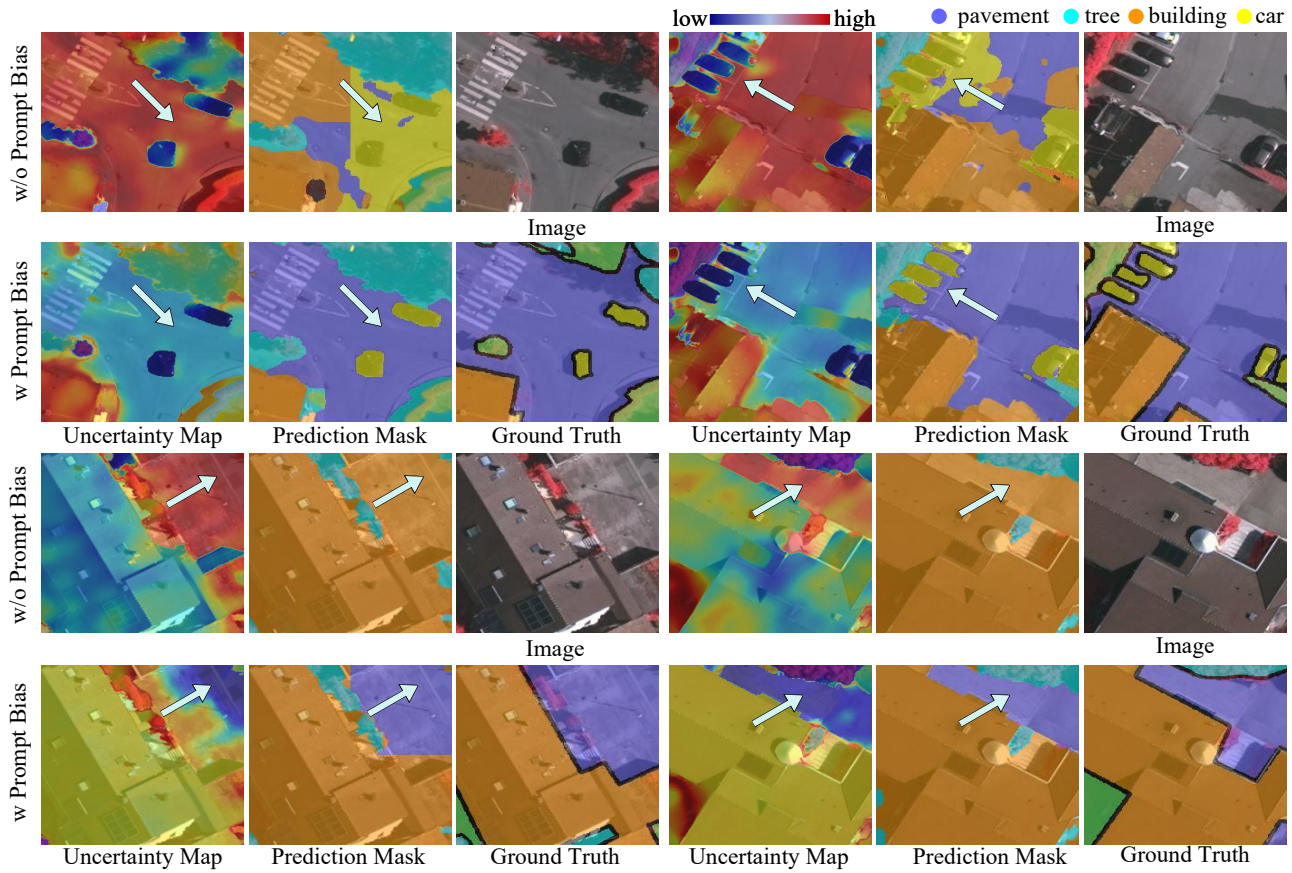

Figure S8. Additional visualizations of predictive entropy and segmentation results with and without prompt bias.

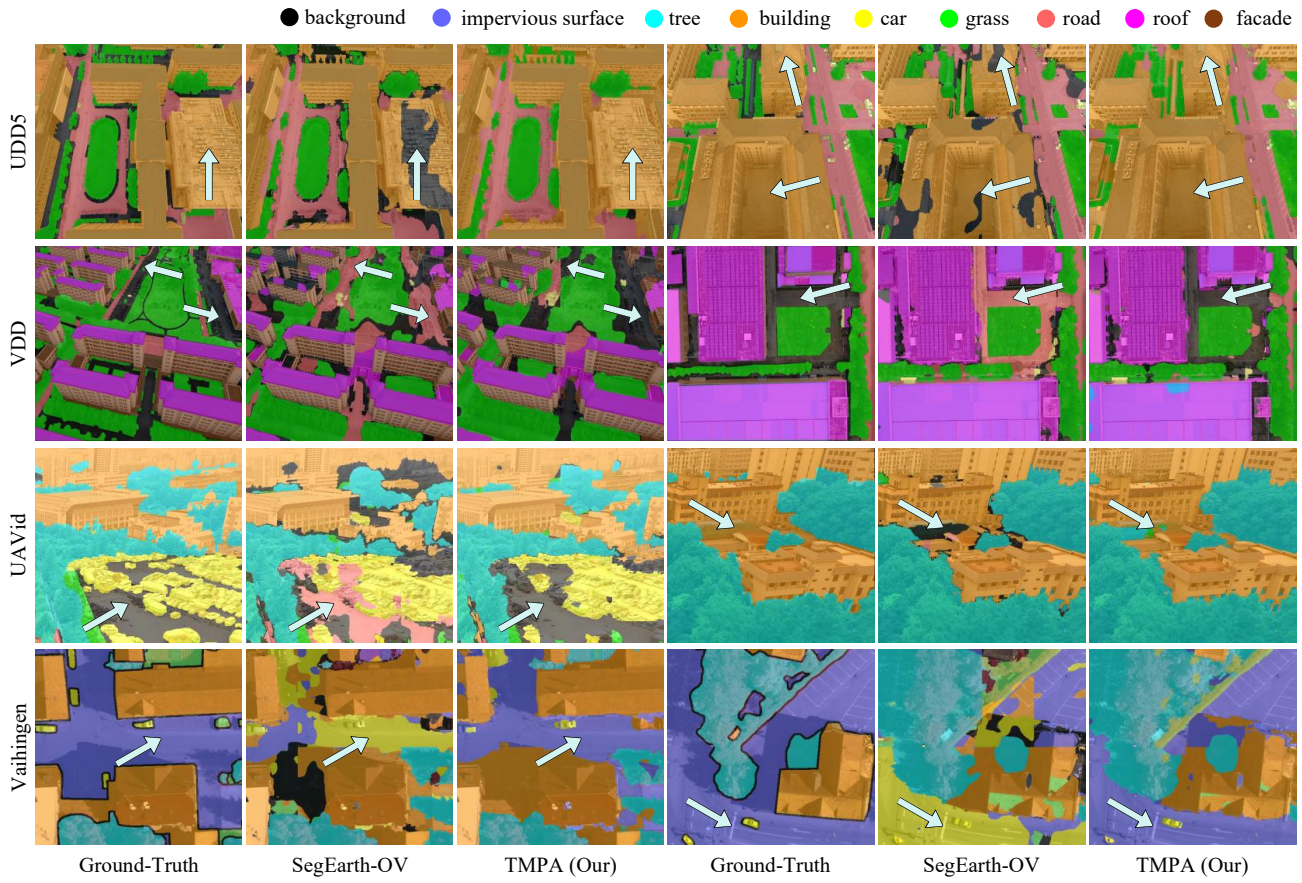

Figure S9. More visualization comparisons of our TMPA with SegEarth-OV [31] on UDD5 [10], VDD [6], UAVid [37], and Vaihingen datasets. The arrows indicate specific areas where predictions of our TMPA show notable improvement.

Table S6. Results (mIoU, %) with text descriptions generated by different large language models.

LLM	Vaihingen	$\text{WHU}^{\text{Aerial}}$
-	29.1	49.2
ChatGPT-4o [22]	32.7	50.6
ChatGPT-5 [45]	34.3	50.1
DeepSeek-R1 [19]	33.1	51.9
Claude-4-sonnet [2]	34.5	51.6
Gemini-2.5-Pro [12]	**35.9**	**52.4**

Table S7. Performance (mIoU, %) comparison with different settings of n in the Top-n visual feature selection.

Dataset	n=0	n=1	n=3	n=5
Vaihingen	40.0	43.2	**43.4**	43.3
$\text{WHU}^{\text{Aerial}}$ [24]	53.8	55.0	**55.6**	55.4

creases, culminating in an optimal peak at step 3 with mIoU of 43.4% on Vaihingen and 55.6% on $\text{WHU}^{\text{Aerial}}$ [24]. This trajectory indicates that early updates in TMPA effectively mitigate the distribution shift between the source and test domains. However, further steps lead to overfitting, where the model begins to assimilate sample-specific noise and erroneous predictions, resulting in parameter drift and degraded generalization. Consequently, we adopt 3 adaptation steps as the default setting for our experiments.

Table S8. Results (mIoU, %) of different methods with our TMPA on nature image dataset.

Methods	PascalContext59	COCOStuff	Cityscapes
SCLIP [55]	34.2	22.4	32.2
+ TMPA	**36.5** ↑ **2.3**	**23.8** ↑**1.4**	**38.0** ↑**5.8**
ClearCLIP [29]	35.9	23.9	30.0
+ TMPA	**37.1** ↑**1.2**	**25.4** ↑**1.5**	**33.9** ↑**3.9**
CASS [27]	40.2	26.7	39.4
+ TMPA	**41.9** ↑**1.7**	**27.8** ↑**1.1**	**41.2** ↑**1.8**

Impact of LLM Selection in Cat-Prompt. We evaluate Cat-Prompt by employing different LLMs to generate

category descriptions, as shown in Table S6. Incorporating LLM-generated descriptions consistently improves segmentation performance over the label-only baseline, highlighting the value of richer linguistic cues for enhancing visual-text alignment. However, the magnitude of this improvement is model-dependent, which we attribute to variations in their capacity to generate semantically relevant attributes. Among all evaluated LLMs, Gemini-2.5-Pro [12] delivers the largest improvements, raising mIoU by 6.8% on Vaihingen and 3.2% on $\text{WHU}^{\text{Aerial}}$ [24]. These results indicate that Gemini-2.5-Pro [12] produces more discriminative and context-aware descriptions that better capture the visual characteristics of remote sensing imagery. Therefore, we adopt Gemini-2.5-Pro [12] to generate text descriptions for each category.

Furthermore, the generation of category descriptions using Gemini-2.5-Pro [12] is extremely low-cost, with the overall expense being negligible. For Cat-Prompt, input prompts ($\sim$500 tokens) cost only \$0.0006 per dataset, and the generated outputs ($\sim$40 tokens) cost \$0.0004 per description, based on rates of \$1.25 and \$10 per million tokens for input and output, respectively.

Effect of the number (n) of low-entropy visual features in VGTA. Table S7 details the ablation study on the number of selected top-n low-entropy visual features in Eq. (4). The results demonstrate that while incorporating visual features significantly boosts performance over the baseline, the model remains relatively insensitive to the specific number of features. Specifically, introducing Top-1 visual features already yields a substantial improvement of 3.2% and 1.2% on the Vaihingen and $\text{WHU}^{\text{Aerial}}$ [24], respectively. However, further increasing n results in performance saturation. For instance, increasing n from 1 to 3 yields only a marginal gain of 0.2% on Vaihingen and 0.6% on $\text{WHU}^{\text{Aerial}}$. Setting n further to 5 results in negligible performance changes across both datasets. This implies that the integration of high-confidence visual guidance is pivotal, while the model remains robust to the number of selected features. We adopt Top-3 as the balanced setting for our experiments.

Evaluation on natural image datasets. To verify the effectiveness of our proposed TMPA on general visual domains, we conducted additional experiments on natural image datasets, as shown in Table S8. The results indicate that our method delivers consistent performance gains across various baseline, thus validating the flexibility and plug-and-play of our TMPA. Specifically, on PascalContext59 [42] and COCOStuff [5], TMPA consistently enhances the performance of SCLIP [55], ClearCLIP [29], and CASS [27], yielding improvements ranging from 1.1% to 2.3%. This positive trend is even more pronounced on Cityscapes [13], where our TMPA achieves substantial gains of 5.8% (SCLIP [55]), 3.9% (ClearCLIP [29]), and 1.8% (CASS [27]). These consistent improvements across diverse natural image datasets demonstrate the generalization and effectiveness of our TMPA.