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

Ting Yang[1], Qilong Wang[1]*, Qibin Hou[2], Qinghua Hu[1]
[1]Tianjin University [2]Nankai University
{yting_123, qlwang, huqinghua}@tju.edu.cn, houqb@nankai.edu.cn

Abstract

The rise of vision-language models (VLMs) has driven the initial exploration of open-vocabulary remote sensing image semantic segmentation (OVRSIS), enabling recognition of unseen categories in complex Earth observation scenes. However, existing methods primarily focus on enhancing visual representations of domain-specific remote sensing images, while overlooking the effect of textual information. In this paper, we argue that there exists a crucial issue of textual ambiguity in OVRSIS task, limiting final segmentation performance. Therefore, we propose a plug-and-play yet effective Test-time Multi-Prompt Adaptation (TMPA) method to mitigate textual ambiguity in OVRSIS. Specifically, TMPA first generates diverse, context-aware descriptions for each category instead of the naive class name by executing a large language model with a task-driven prompt, which can effectively avoid some textual ambiguity, i.e., background class has different meanings in various tasks. Furthermore, TMPA develops a visual-guided test-time adaptation strategy for the generated multi-prompts, which adaptively refines the prompt representations of each category with high-confidence visual features for the uncertain predictions with high entropy, making TMPA better applicable to different scenarios. Particularly, a pixel-level loss with entropy minimization is proposed to optimize the text prompt with a bias during inference, where prompt bias is constructed based on a weighted combination of high-confidence visual features. Our TMPA can be flexibly integrated into existing methods for boosting their performance. Extensive experiments are conducted on 17 remote sensing datasets, and the results show our TMPA can significantly improve its counterparts, while achieving state-of-the-art performance.

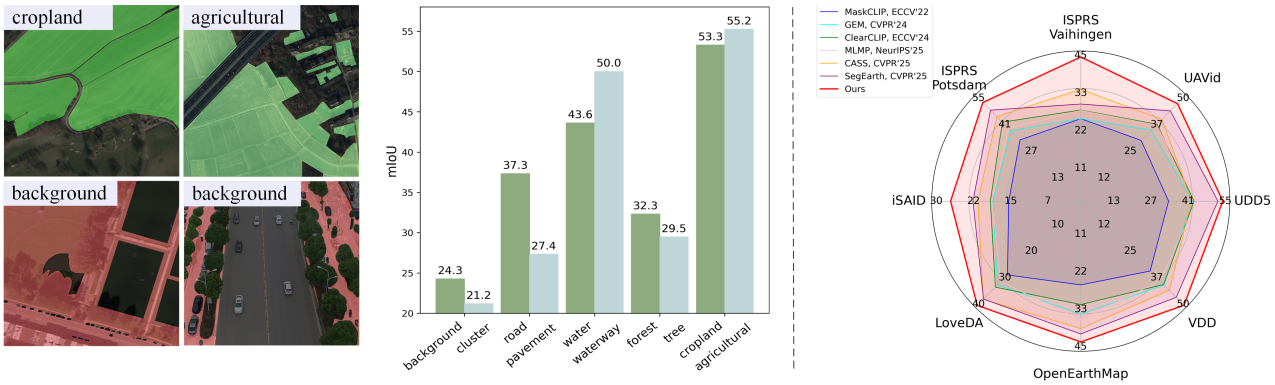
1. Introduction

Remote sensing image semantic segmentation as an important task in computer vision community benefits diverse applications, including land-use and land-cover mapping [23, 49], urban planning [18, 23], and environmental monitoring [35]. However, the current remote sensing image semantic segmentation methods [9, 64] mainly operate under a closed-set setting, restricting their recognition to a predefined set of classes and hindering their ability to generalize to new, unseen categories. This makes them ill-suited for dynamic real-world scenarios, where novel features continuously emerge, such as newly constructed infrastructures or continually evolving land-cover types.

To overcome this limitation, recent studies have begun to explore open-vocabulary remote sensing image semantic segmentation (OVRSIS) in remote sensing imagery [7, 30, 31]. By leveraging the power of vision-language models and cross-modal learning [25, 47], open-vocabulary semantic segmentation (OVSS) allows models to assign pixel-level labels based on arbitrary textual descriptions, breaking free from the restrictions of predefined training categories. As a seminal work, SegEarth-OV [31] argues that OVSS models designed for natural images are sub-optimal for remote sensing images, and introduces a feature upsampler module to refine low-resolution features while preserving semantic consistency with the image content. Meanwhile, OVRS [7] presents a rotation-aggregative similarity computation module and integrates multi-scale features to progressively generate scale-aware semantic maps, effectively addressing challenges posed by arbitrary orientations and significant scale variations in remote sensing images. Subsequently, RSKT-Seg [30] proposes an efficient framework tailored for remote sensing images by considering rotational invariance, while incorporating DINO [8] to enhance semantic representations with richer spatial information.

Although advanced efforts are made, aforementioned methods primarily focus on enhancing visual representations for domain-specific remote sensing images [7, 30, 31], but neglect the effect of textual information. In this paper, we argue that there exists the crucial issue of textual ambiguity in OVRSIS. As shown in Fig. 1 (a), similar visual information has different class names, while the same class name is corresponding to different visual information

*Corresponding author.



(a) Textual Ambiguity (b) Sensitivity to Text Prompts (c) Performance Comparison (mIoU, %)

Figure 1. (a) Examples of text ambiguity, where dark green and light green indicate cropland and agricultural areas, and red indicates background. Text ambiguity mainly arises from synonymy (top) and polysemy (bottom), where visually similar features are assigned different labels (e.g., 'cropland' vs. 'agricultural') and a single class name corresponds to visually distinct features (e.g., 'background'), respectively. (b) SegEarth-OV [31] exhibits inconsistent performance on the LoveDA [56] dataset by using various synonymous prompts (e.g., 'background' vs. 'cluster' or 'road' vs. 'pavement'), indicating OVRSIS model potentially is sensitivity to text prompts. (c) Our TMPA consistently outperforms SOTA methods (e.g., SegEarth-OV [31]), demonstrating the importance in handling textual ambiguity.

in various tasks (i.e., background). Meanwhile, OVRSIS model [31] with different synonymous text descriptions results in clearly different performance, as shown in Fig. 1 (b). They indicate that the issue of textual ambiguity potentially limits the OVRSIS performance. Furthermore, OVRS [7] and RSKT-Seg [30] perform OVRSIS in a domain adaptation paradigm, which require to train the models on some annotated benchmarks and potentially suffer from risk of overfitting and limited scalability.

To address above challenges, we propose a plug-and-play yet effective Test-time Multi-Prompt Adaptation (TMPA) method to dynamically adapt textual representations at test time, whose goal is to mitigate textual ambiguity and strengthen visual-language matching of the uncertain predictions for further boosting OVRSIS performance. To this end, our TMPA consists of two key components: a Context-Aware Text Prompt Generator (Cat-Prompt) and a Visual-Guided Test-Time Adaptation (VGTA) strategy. Specifically, our Cat-Prompt first constructs a task-driven prompt involving basic information of task (e.g., task scenario and class names), which further guides a large language model (e.g., Gemini [12]) to produce a group of diverse, context-aware descriptions for each category. As such, these generated prompts can mitigate the textual ambiguity inherent in naive class names and resolve semantic inconsistencies across different datasets. By considering misalignment between pre-generated textual prompts and visual features across various tasks (especially for similar visual features), our VGTA dynamically refines embeddings of the generated prompts at test time. Particularly, predictions of regions with visual-language misalignment frequently occur uncertain with high entropy. To strengthen visual-language matching for uncertain predictions, VGTA constructs text prompt bias with a weighted combination of high-confidence visual features. Then, a pixel-level loss with entropy minimization is proposed to optimize the prompt bias (i.e., weights of visual features) during inference stage. As such, VGTA improves visual-language matching with help of matching between visual features with uncertain predictions and high-confidence ones, leading better generalization across different scenarios. The overview of our TMPA is shown in Fig. 2, which can be flexibly integrated into existing methods (e.g., Segearth-OV [31] and CASS [27]), bringing clear performance gains (see Fig. 1 (c)). To evaluate our TMPA, experiments are conducted with SegEarth-OV [31] on 17 benchmarks, while four existing models are used for generalization verification. The contributions of this work are concluded as follows.

- This paper, to our best knowledge, makes the first attempt to mitigate the issue of textual ambiguity in OVRSIS. Particularly, we propose a plug-and-play yet effective Test-time Multi-Prompt Adaptation (TMPA) method, which can be flexibly integrated into existing methods for boosting their performance.
- To this end, our TMPA presents a Context-Aware Text Prompt Generator (Cat-Prompt) and a Visual-Guided Test-Time Adaptation (VGTA) strategy. Cat-Prompt produces a group of diverse, context-aware descriptions for mitigating the textual ambiguity, while VGTA dynamically refines embeddings of the generated prompts at test time, guaranteeing better generalization.
- Extensive experiments on 17 remote sensing datasets demonstrate that our TMPA clearly outperforms its counterparts, while achieving state-of-the-art performance.

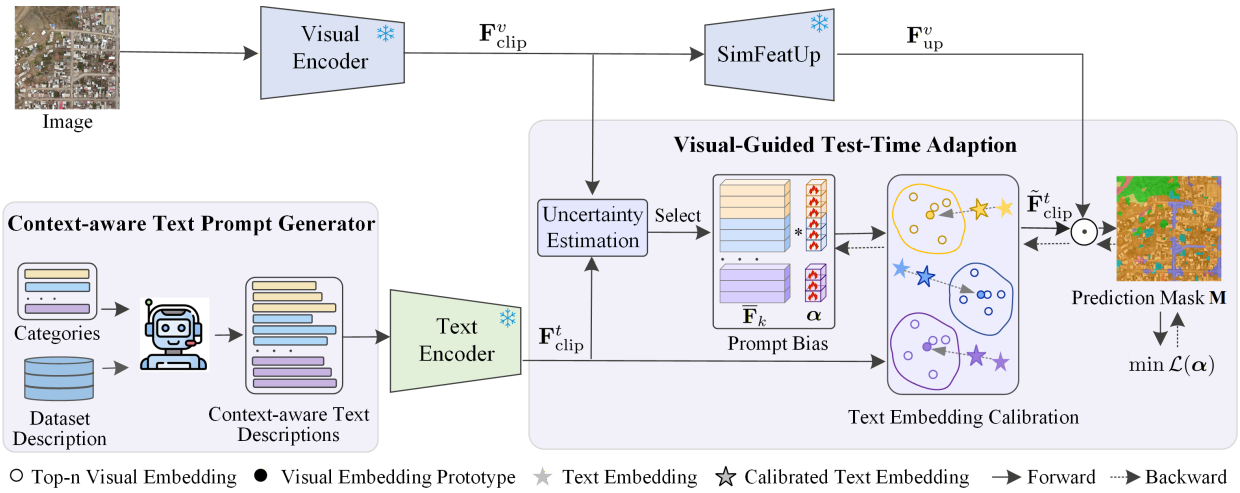


Figure 2. Overview of our proposed TMPA method for OVRSIS, whose core components involve a Context-Aware Text Prompt Generator (Cat-Prompt) and Visual-Guided Test-Time Adaptation (VGTA). Specifically, Cat-Prompt generates diverse, context-aware textual descriptions for each candidate category. Then, VGTA adaptively calibrates pre-generated textual embeddings by optimizing a prompt bias, which are built based on high-confidence visual features and optimized by minimizing a pixel-level entropy loss. Finally, the calibrated textual features are combined with up-sampling visual features to produce the final segmentation results.

2. Related Work

In this section, we briefly review some related works, including open-vocabulary semantic segmentation (OVSS), open-vocabulary remote sensing image segmentation (OVRSIS), and test-time adaptation (TTA).

Open-Vocabulary Semantic Segmentation. The advent of vision-language models (VLMs), particularly CLIP [47], has significantly propelled the progress of OVSS. CLIP-based OVSS methods can be broadly categorized by training-based and training-free ones. For training-based methods, existing studies [33, 36, 43, 48] focus on training a localization-aware CLIP model for direct dense prediction, while the remaining ones [11, 15, 34, 59, 60] adapt the CLIP model for segmentation tasks by fine-tuning a subset of parameters or introducing lightweight, trainable modules. For training-free OVSS methods, some works [4, 16, 29, 55] aim to enhance spatial precision by refining self-attention within the visual encoder, and others [3, 26, 50, 52] adopt a two-stage strategy that generates category-agnostic mask proposals before classifying them. Although OVSS task has been well studied, transferring of OVSS methods to remote sensing domain is still an open and challenging problem.

Open-Vocabulary Remote Sensing Image Segmentation. Recently, OVRSIS has attracted increasing research interests. Specifically, OVRS [7] considers the arbitrary orientations and significant scale variations inherent in remote sensing image and enhances the visual encoder of CatSeg [11] to improve robustness. GSNet [62] introduces a specialist-generalist model to incorporate domain-specific knowledge. RSKT-Seg [30] obtains rotation-invariant features while additionally introducing DINO [8] as the visual encoder to enhance fine-grained representations. Notably, these methods require training on labeled datasets, limiting the scalability and generalization. In contrast, training-free SegEarth-OV [31] introduces a general feature upsampler, effectively enhancing low-resolution features while maintaining semantic consistency with the image content. However, these methods primarily focus on enhancing visual representations, and overlook the effect of text information.

Test-Time Adaptation. TTA aims to adapt a pre-trained model online to a target domain without labeled data. Early studies [54] adapt models by minimizing prediction entropy at test time through updating normalization parameters. Recently, the scope of TTA has been extended to vision-language models. Specifically, TPT [51] and DiffTPT [17] optimize text prompts to encourage consistent predictions across augmented views by minimizing marginal entropy. CLIPArTT [21] enhances text supervision by merging multiple class descriptions into a unified prompt that serves as a pseudo-label, while CLIP-OT [40] formulates pseudo-labeling as an Optimal Transport [14] problem, iteratively distilling knowledge from multiple text prototypes. However, existing VLM-based TTA methods are primarily designed for classification and are ill-suited for pixel-level segmentation. Recently, MLMP [44] extends TTA to OVSS by dynamically fusing multi-layer visual features and fine-tuning the LN layer of the visual encoder based on prediction uncertainty. Different from it, our TMPA aims to adaptively refine text prompt embeddings with high-confidence visual features for uncertain predictions in OVRSIS.

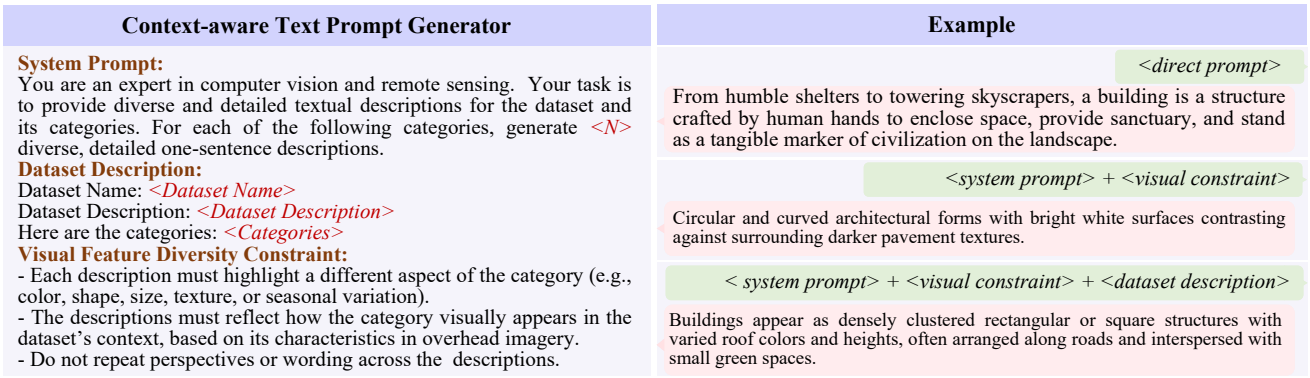


Figure 3. **Left:** Style of our Context-Aware Text Prompt Generator (Cat-Prompt), which consists of a system prompt to define task-specific instructions, a dataset description to provide contextual guidance, and a visual feature diversity constraint to enforce the generation of varied and visually grounded category descriptions. **Right:** Examples of generated textual descriptions for the "building" category in $\text{WHU}^{\text{Aerial}}$ [24] under different prompts. *Direct prompt* indicates textual descriptions generation with LLM without any constraints.

3. Method

In this section, we first introduce the overview of our proposed Test-time Multi-Prompt Adaptation (TMPA) method, aiming to mitigate the issue of textual ambiguity lying in OVRSIS by involving two core components: Context-Aware Text Prompt Generator (CAT-Prompt) and Visual-Guided Test-Time Adaptation (VGTA). Then, we describe our CAT-Prompt and VGTA in details, respectively.

3.1. Overview of TMPA

Given an input remote sensing image $\mathbf{X} \in \mathbb{R}^{H\times W\times 3}$ and a set of concepts $\mathbf{C} = \{C_1, C_2, \ldots, C_K\}$ defined in natural language, open-vocabulary remote sensing image segmentation (OVRSIS) aims to predict a semantic mask $\mathbf{M} \in \{1, \ldots, K\}^{H\times W}$ that assigns a concept label to each pixel. In OVRSIS, the concept set $\mathbf{C}$ can be of arbitrary length, which plays a key role in the performance of such text-guided segmentation. However, textual ambiguity inherent in the concept label heavily limits the segmentation performance and attracts less attention.

Therefore, we propose a plug-and-play yet effective TMPA method, which can be flexibly integrated into existing methods (e.g., Segearth-OV [31], CASS [27] and ClearCLIP [29]). By taking Segearth-OV [31] as an example, the overview of our TMPA is illustrated in Fig. 2. Specifically, following the existing works [27, 29, 31], we leverage pre-trained CLIP model [47] to extract textual features $\mathbf{F}^t_{\text{clip}} \in \mathbb{R}^{KN\times D}$ and visual features $\mathbf{F}^v_{\text{clip}} \in \mathbb{R}^{HW\times D}$ from the context-aware textual descriptions and the input image, respectively. Here, K denotes the number of categories, N is the number of context-aware textual descriptions per category, H and W represent the height and width of the visual features, and D is the dimension of both textual and visual embeddings. Particularly, for visual branch we also adopt a SimFeatUp [31] module to obtain the up-sampled visual features $\mathbf{F}^v_{\text{up}} \in \mathbb{R}^{H'\times W'\times D}$ based on $\mathbf{F}^v_{\text{clip}}$.

To mitigate the textual ambiguity inherent in naive class names and resolve semantic inconsistencies across different datasets, our TMPA first designs a Context-Aware Text Prompt Generator (CAT-Prompt). Given basic task information (e.g., task scenario and class names), CAT-Prompt constructs task-driven prompts for guiding a large language model (LLM), such as Gemini [12], to generate a set of N diverse, context-aware textual descriptions for each candidate class C_k, where $k \in \{1, \ldots, K\}$. Furthermore, a Visual-Guided Test-Time Adaptation (VGTA) is proposed to optimize the initial textual embeddings at test time, where a prompt bias is first generated by a weighted combination of high-confidence visual features. Then, the calibrated textual embeddings $\tilde{\mathbf{F}}^t_{\text{clip}}$ are optimized by minimizing a pixel-level entropy loss during inference. Finally, similarity between the calibrated textual features $\tilde{\mathbf{F}}^t_{\text{clip}}$ and the visual features $\mathbf{F}^v_{\text{up}}$ is computed to generate segmentation mask $\mathbf{M}$:

$$\mathbf{M} = \arg\max_k \text{softmax}\left(\text{sim}\left(\mathbf{F}^v_{\text{up}}, \tilde{\mathbf{F}}^t_{\text{clip}}\right)\right), \quad (1)$$

where sim denotes cosine similarity. In the followings, we will introduce details of our CAT-Prompt and VGTA.

3.2. Context-Aware Text Prompt Generator

In OVRSIS, VLMs enable open-vocabulary inference by projecting test images and candidate class descriptions into a shared embedding space. Therefore, the quality of text prompts directly shapes the understanding of target concepts and influences generalization to new scenarios. However, raw class names as prompts introduce text ambiguity caused by inconsistent annotation standards and lexical variations (e.g., polysemy and synonymy), impairing the alignment between text and visual features (see Fig. 1).

To tackle this issue, we propose a Context-Aware Text Prompt Generator (Cat-Prompt) to generate diverse,

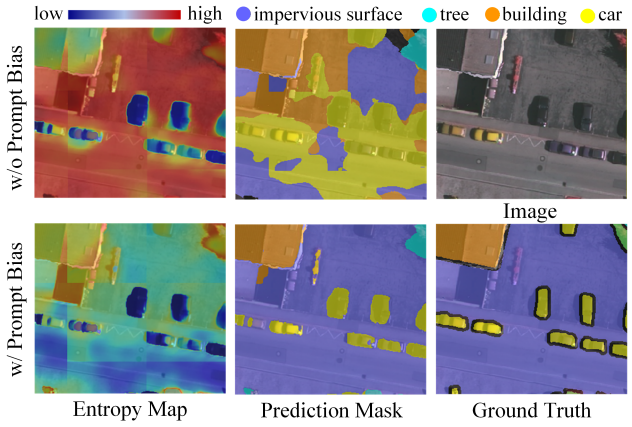


Figure 4. Visualization of segment results with/without prompt bias. In first row, incorrect regions generally correspond to higher entropy (e.g., impervious surfaces are wrongly predicted as cars). After refining text embeddings by our prompt bias (second row), uncertain regions are correctly predicted with high confidence.

context-aware descriptions for each category. As shown in Fig. 3 (left), instead of directly prompting LLMs for synonyms or descriptions [1, 46], our Cat-Prompt comprises three core components: a system prompt, a dataset description, and a visual feature diversity constraint. In practice, we generate multiple distinct descriptions per category (e.g., five per category) as textual prototypes for open-vocabulary segmentation. The system prompt defines the overall objective of the task and specifies the expected output format, guiding the LLM (e.g., Gemini [12]) to generate well-structured and task-driven responses. The dataset description provides domain- and scene-level context, guiding the LLM to produce textual descriptions aligned with the visual characteristics of the dataset. Finally, the visual feature diversity constraint encourages varied and visually grounded descriptions, improving robustness to textual ambiguity during segmentation. Fig. 3 (right) illustrates an example of generated descriptions on WHU$^{\text{Aerial}}$ [24], where direct prompt provides a general description but fails to match the visual features of remote sensing imagery. The visual constraint introduces details such as 'circular' or 'curved' shapes and 'white' color; however, these features are inconsistent with the typical appearance of Wuhan buildings. Combining visual constraints with dataset description (scene-aware text) can accurately capture visual patterns and spatial layout of buildings. As such, Cat-Prompt provides effective textual descriptions with more visual characteristics on the target dataset. The complete details of the example are provided in the supplementary.

3.3. Visual-guided Test-Time Adaptation

Although the pre-generated multi-prompt can alleviate the issue of textual ambiguity, visual-text misalignment still exist along scenarios and tasks changing. Particularly, as illustrated in Fig. 4, we visualize the predictive entropy derived from the similarity distribution between visual features and text embeddings, where we can observe that regions with correct predictions consistently exhibit low entropy (i.e., high confidence), while incorrect regions correspond to high entropy (i.e., high uncertainty). Since visual-text misalignment leads to incorrect predictions, we aim to exploit auxiliary visual features matching to improve visual-text alignment for regions with high uncertain predictions. Therefore, we propose a Visual-guided Test-Time Adaptation (VGTA) strategy, whose core idea is integration of matching between visual features with uncertain predictions and high-confidence ones into multi-prompt adaptation at test time. Specifically, VGTA first constructs a prompt bias by using a weighted combination of high-confidence visual features, which is added to pre-generated prompt embeddings and optimized by minimizing a pixel-level entropy loss during inference, aiming to adaptively refine prompt embeddings to enhance visual-text alignment.

For constructing prompt bias based on high-confidence visual features, we first compute the similarity between the visual features $\mathbf{F}^v_{\text{clip}}$ and textual embeddings $\mathbf{F}^t_{\text{clip}}$ to obtain the predicted probability $\hat{\mathbf{P}} \in \mathbb{R}^{H\times W\times K}$ over all classes for each pixel:

$$\hat{\mathbf{P}} = \text{softmax}\left(\mathbf{F}^v_{\text{clip}}\left(\mathbf{F}^t_{\text{clip}}\right)^\top\right), \tag{2}$$

where $\top$ indicates matrix transposition, and softmax is softmax function. Then, we estimate the pixel-wise prediction uncertainty $\mathbf{U} \in \mathbb{R}^{H\times W}$ by computing the entropy of the predicted probability distribution $\hat{\mathbf{P}}$ at each pixel:

$$U_{x,y} = -\sum_{k=1}^{K} \hat{P}_{x,y,k} \log \hat{P}_{x,y,k}, \tag{3}$$

where $\hat{P}_{x,y,k}$ represents the predicted probability of assigning the pixel at position (x, y) to class k, $U_{x,y}$ denotes the prediction uncertainty at position (x, y).

For the k-th category, we select the Top-n pixel positions with the lowest entropy those predicted as class k, and extract their visual features to form the feature set $\mathcal{F}_k$:

$$\mathcal{F}_k = \{\mathbf{f}_{x,y} \mid (x, y) \in \text{Top}_n(\mathbf{U}, \mathcal{Z}_k)\}, \tag{4}$$

$$\mathcal{Z}_k = \{(x, y) \mid \hat{\mathbf{M}}_{x,y} = k \text{ and } U_{x,y} \leq \bar{U}\}, \tag{5}$$

where $\hat{\mathbf{M}} = \arg\max_k \hat{\mathbf{P}}_{x,y,k}$ is a mask prediction, and $\bar{U}$ indicates the average value of $\mathbf{U}$. $\mathbf{f}_{x,y} \in \mathbb{R}^D$ means the visual feature at position (x, y) of $\mathbf{F}^v_{\text{clip}}$, and $\mathcal{Z}_k$ is the position set of high-confidence visual features in k-th class. As such, we pick the highest confident visual features for all predicted categories and compute their mean:

$$\bar{\mathbf{f}}_k = \frac{1}{|\mathcal{F}_k|} \sum_{\mathbf{f}\in\mathcal{F}_k} \mathbf{f}, \tag{6}$$

which are used to refine the pre-generated prompts, and the final calibrated textual embeddings can be computed as:

$$\tilde{\mathbf{F}}^t_{\text{clip}}(\boldsymbol{\alpha}) = (\mathbf{1} - \boldsymbol{\alpha}) \odot \mathbf{F}^t_{\text{clip}} + \boldsymbol{\alpha} \odot \overline{\mathbf{F}}, \tag{7}$$

where $\boldsymbol{\alpha} \odot \overline{\mathbf{F}}$ acts as a visual-guided prompt bias, $\overline{\mathbf{F}} \in \mathbb{R}^{KN \times D}$ is a matrix by repeating the category-wise mean visual features $\bar{\mathbf{f}}_k$, matching the size of $\mathbf{F}^t_{\text{clip}}$, $\boldsymbol{\alpha} \in [0, 1]^{KN}$ is a learnable, zero-initialized matrix that determines the strength of visual feature injection for each category description, and $\odot$ denotes element-wise multiplication.

To optimize the prompt bias with parameters of $\boldsymbol{\alpha}$ during inference, we introduce a pixel-level loss with entropy minimization for segmentation task. Specifically, given a test image and its context-aware textual descriptions, we can obtain the corresponding up-sampling visual features $\mathbf{F}^v_{\text{up}} \in \mathbb{R}^{H' \times W' \times D}$ and the calibrated textual embeddings $\tilde{\mathbf{F}}^t_{\text{clip}} \in \mathbb{R}^{KN \times D}$ by using Eq. (7). Then, $\boldsymbol{\alpha}$ is learned by optimizing the following objective function:

$$\min \mathcal{L}(\boldsymbol{\alpha}) = \arg\min_{\boldsymbol{\alpha}^\star} \frac{1}{H'W'} \sum_{x,y} \mathbf{P}^\top_{x,y}(\boldsymbol{\alpha}) \log \mathbf{P}_{x,y}(\boldsymbol{\alpha}),$$

$$\mathbf{P}(\boldsymbol{\alpha}) = \text{softmax}\left(\mathbf{F}^v_{\text{up}} \left(\tilde{\mathbf{F}}^t_{\text{clip}}(\boldsymbol{\alpha})\right)^\top\right), \tag{8}$$

which aims to increase separability of per-pixel class probability distributions, while alleviating the issue of prediction uncertainty. Given the optimized $\boldsymbol{\alpha}^\star$, the final segmentation results can be obtained by Eq. (1). As shown in Fig. 4, prediction uncertainty of ambiguous regions can be clearly alleviated by calibrated textual embeddings, thereby yielding more robust and reliable segmentation results.

4. Experiment

In this section, we evaluate our TMPA method on 17 diverse datasets, including 8 multi-class semantic segmentation datasets and 9 single-class land-cover extraction datasets. Specifically, we first describe the datasets and implementation details, then compare with state-of-the-art (SOTA) methods across all datasets, and finally conduct ablation studies to assess the effectiveness of the key components.

4.1. Dataset

Multi-Class Semantic Segmentation. Following [31], 8 widely used remote sensing semantic segmentation datasets are used to conduct experiments, including OpenEarthMap [58], LoveDA [56], iSAID [57], Potsdam, Vaihingen[1], UAVid [37], UDD5 [10], and VDD [6]. The first five datasets primarily consist of satellite imagery, while the remaining three focus on UAV imagery. All datasets provide pixel-level annotations across multiple foreground categories along with a background class.

[1]https://www.isprs.org/education/benchmarks/UrbanSemLab

Single-Class Extraction. Besides, 9 single-class land-cover extraction datasets are used for evaluation, spanning 3 object categories: building extraction ($\text{WHU}^{\text{Aerial}}$ [24], $\text{WHU}^{\text{Sat.II}}$ [24], Inria [38], xBD^{pre} [20]), road extraction (CHN6-CUG [66], DeepGlobe[2], Massachusetts [41], SpaceNet [53]), and flood detection (WBS-SI[3]). All datasets adopt a binary annotation.

4.2. Implementation Details

For comparing with SOTA, we use SegEarth-OV [31] as our base model, where visual and text encoders are initialized from CLIP [47] of ViT-B/16. The input images are resized to a long side of 448 pixels, and inference is conducted using a 224 × 224 sliding window with a 112 pixel stride. During test-time adaptation, all parameters of the original SegEarth-OV [31] are frozen, and only prompt bias is optimized. The adaptation process is conducted with 3 steps using the Adam optimizer [28]. We evaluate semantic segmentation performance using the mean Intersection over Union (mIoU) metric. For single-class extraction, IoU of foreground class is reported. All experiments are conducted on a server equipped with 8 NVIDIA A6000 GPUs. Code is available at https://github.com/TiY68/TMPA.

4.3. Comparison with State-of-the-art

To validate the effectiveness of our TMPA, we compare with nine SOTA methods on 17 remote sensing datasets.

Results on Semantic Segmentation. As shown in Table 1, our proposed TMPA consistently surpasses all compared methods by a clear margin, including MaskCLIP [65], ClearCLIP [29], SCLIP [55], GEM [4] and recently proposed CASS [27] and MLMP [44]. Note that our TMPA outperforms SegEarth-OV [31] by 4.6% on average, while achieving new state-of-the-art results. Particularly, our TMPA achieves a notable performance gain on the Vaihingen dataset, where it surpasses SegEarth-OV [31] by an impressive 14.3% and exceeds CASS [27] by 9.9%. Besides, TMPA outperforms SegEarth-OV on iSAID (4.5%), Potsdam (4.0%), and VDD (3.7%), highlighting the strong generalization ability in diverse remote sensing scenarios. These results highlight that text ambiguity is a crucial challenge in OVRSIS task and our TMPA has great ability to handle text ambiguity, bringing clear performance gains.

Furthermore, our TMPA achieves 10.6% improvement gains over MLMP [44], which performs TTA on semantic segmentation of natural images, underscoring the significant domain gap between natural and remote sensing imagery. Nevertheless, our TMPA also shows strong generalization on natural image datasets, whose results are given in supplementary. Additionally, TMPA outperforms training-

[2]http://deepglobe.org

[3]https://www.kaggle.com/datasets/shirshmall/water-body-segmentation-in-satellite-images

Table 1. Comparison (mIoU, %) with state-of-the-art methods on multi-class remote sensing segmentation datasets. $*$ indicates a training-based OVRSIS method, where supervised training is performed on iSAID [57] dataset and evaluation is conducted on other datasets. **Best** and second best performances are highlighted.

Methods	OpenEarthMap	LoveDA	iSAID	Potsdam	Vaihingen	UAVid	UDD5	VDD	Avg
MaskCLIP [65]	25.1	27.8	14.5	31.7	24.7	28.6	32.4	32.9	27.2
SCLIP [55]	29.3	30.4	16.1	36.6	28.4	31.4	38.7	37.9	31.1
GEM [4]	33.9	31.6	17.7	36.5	24.7	33.4	41.2	39.5	32.3
ClearCLIP [29]	31.0	32.4	18.2	40.9	27.3	36.2	41.8	39.3	33.4
CASS [27]	38.2	37.0	20.7	43.8	33.5	38.5	40.9	42.0	37.4
MLMP [44]	35.5	30.4	17.9	37.6	27.3	35.9	42.9	37.56	33.1
SegEarth-OV [31]	39.8	36.9	21.7	47.1	29.1	42.5	50.6	45.3	39.1
OVRS* [7]	-	-	93.3	19.9	20.8	-	-	-	-
RSKT-Seg* [30]	-	28.1	93.2	20.3	17.5	17.2	13.9	25.3	-
TMPA (Ours)	**42.2** ↑2.4	**39.7** ↑2.8	**26.2** ↑4.5	**51.1** ↑4.0	**43.4** ↑14.3	**45.9** ↑3.4	**52.4** ↑1.8	**49.0** ↑3.7	**43.7** ↑4.6

Table 2. Comparison (IoU, %) with state-of-the-art methods on single-class remote sensing segmentation datasets. **Best** and second best performances are highlighted.

Method	Building Extraction				Road Extraction				Flood Detection
	$\text{WHU}^{\text{Aerial}}$	$\text{WHU}^{\text{Sat.II}}$	Inria	xBD^{pre}	CHN6-CUG	DeepGlobe	Massachusetts	SpaceNet	WBS-SI
448 × 448:									
MaskCLIP [65]	29.8	14.0	33.4	29.2	28.1	13.2	10.6	20.8	39.8
SCLIP [55]	33.4	21.0	34.9	25.9	21.1	7.0	7.4	14.9	32.1
GEM [4]	24.4	13.6	28.5	20.8	13.4	4.7	5.1	11.9	39.5
ClearCLIP [29]	36.6	20.8	39.0	30.1	25.5	5.7	6.4	16.3	44.9
CASS [27]	38.9	20.0	39.1	33.3	13.0	5.1	5.8	13.8	52.0
MLMP [44]	37.0	24.6	38.7	33.8	13.7	5.0	8.6	14.2	38.7
SegEarth-OV [31]	49.2	28.4	44.6	37.0	35.4	17.8	11.5	23.8	60.2
TMPA (Ours)	**55.6** ↑6.4	**31.1** ↑2.7	**47.5** ↑2.9	**40.6** ↑3.6	**36.8** ↑1.4	**21.5** ↑3.7	**13.6** ↑2.1	**26.0** ↑2.2	**66.0** ↑5.8
896 × 896:									
SegEarth-OV [31]	49.9	-	48.9	43.1	32.8	20.1	17.2	29.1	57.9
TMPA (Ours)	**57.8** ↑7.9	-	**51.9** ↑3.0	**48.6** ↑5.5	**34.2** ↑1.4	**28.5** ↑8.4	**19.6** ↑2.4	**31.8** ↑2.7	**66.5** ↑8.6

based methods (e.g., OVRS [7] and RSKT-Seg [30]) by a significant margin. Particularly, training-based methods tend to overfit their source data (i.e., iSAID [57]) and struggle to generalize in open-vocabulary scenarios. Remarkably, our TMPA achieves 10%∼38% gains over these two methods on open-vocabulary benchmarks without relying on labeled data. Therefore, our TMPA provide a promising solution for effective OVRSIS, which eliminates the requirement of expensive annotation while exhibiting superior generalization across various scenarios.

Results on Single-class Extraction. Following the settings in [31], we experiment on single-class extraction with two input resolutions. As compared in Table 2, our TMPA consistently achieves state-of-the-art performance. With an input resolution of 448×448, TMPA boosts the IoU of SegEarth-OV by 3.9% for building extraction and 5.8% for flood detection. Notably, it also secures a 2.4% gain on the more challenging road extraction task. By up-scaling the input resolution to 896×896, our TMPA further improves performance, yielding gains of 5.5%, 3.7%, and 8.6% for building, road, and flood detection, respectively. These results demonstrate the effectiveness of TMPA in mitigating textual ambiguity for boosting OVRSIS performance again.

Qualitative Results. We further visualize the segmentation results of our proposed TMPA, and compare with the strong counterpart SegEarth-OV [31]. As shown in Fig. 5, our TMPA method can generate more precise prediction masks, especially for small and challenging objects, such as cars (first row). Furthermore, the masks predicted (second row) by TMPA show superior regional coherence compared with SegEarth-OV. These visualizations highlight that TMPA provides an effective method to mitigate textual ambiguity, so leading to better segmentation results. More visualization results can be found in the supplementary.

4.4. Ablation Study

Impact of Key Components. To evaluate effect of key components (i.e., Cat-Prompt and VGTA), we conduct experiments on Vaihingen and $\text{WHU}^{\text{Aerial}}$ datasets. As shown in Table 3, our Cat-Prompt respectively brings 6.8% and 3.2% performance gains over strong baseline (SegEarth-OV [31]) on Vaihingen and $\text{WHU}^{\text{Aerial}}$, where dataset de-

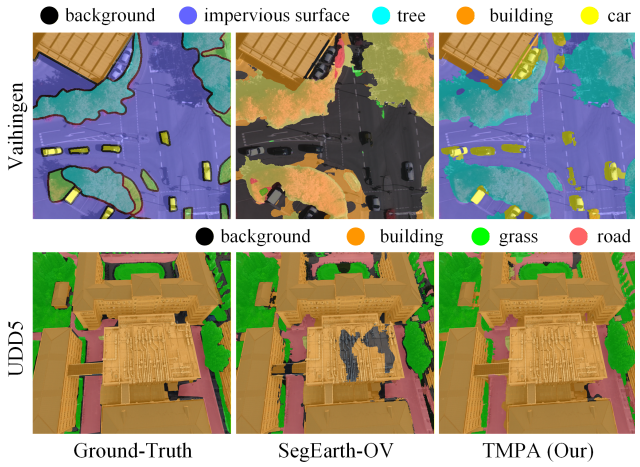


Figure 5. Visualization comparison of our TMPA with SegEarth-OV [31] on the Vaihingen and UDD5 [10] datasets.

scriptions (DD) obtain 3.8% and 2.0% gains. They clearly verify the effectiveness of our Cat-Prompt while demonstrating scene-specific context can help to generate better textual representations aligned with visual features. Built on Cat-Prompt, our VGTA further boosts performance by 7.5% and 3.2% on Vaihingen and WHU$^{\text{Aerial}}$, respectively. Particularly, prompt bias based on visual features (VF) leads to 2.1% and 1.1% improvement over direct learning of a zero-initialized vector (i.e., w/o VF). These results indicate that calibrated textual embeddings with high-confidence visual features can effectively enhance visual-text alignment and thereby yields superior segmentation results.

Effect of Number of Text Descriptions. We further investigate effect of number of generated text descriptions in Cat-Prompt on Potsdam, Vaihingen, and WHU$^{\text{Aerial}}$ datasets. As shown in Table 4, more text descriptions generally bring larger performance gains and peak performance is achieved with five descriptions, which achieves 1.7%, 6.8%, and 3.2% gains over naive category names on Potsdam, Vaihingen, and WHU$^{\text{Aerial}}$, respectively. They highlight the effectiveness of enriched semantic context. The overmuch text descriptions potentially introduce noisy or redundant information, resulting in a slight performance drop and more computational cost. Therefore, we use five text descriptions per category as the default setting.

Generalization to Various Methods. To verify generalization of our plug-and-play TMPA, we apply it to three existing OVSS methods, including CASS [27], ClearCLIP [29], and SCLIP [55]. As shown in Table 5, TMPA consistently yields significant performance gains across different methods. Notably, on WHU$^{\text{Aerial}}$, TMPA achieves significant gains of 18.5%, 10.5%, and 9.5% over these three methods, respectively. The consistent improvements are also observed on other datasets, such as Potsdam (4.3%∼4.7% gains) and Vaihingen (4.7%∼8.9% gains). These results clearly show our TMPA can be well generalized to existing methods for enhancing OVRSIS performance.

Table 3. Results (%) of different configurations for TMPA.

Cat-Prompt		VGTA		Vaihingen	WHU$^{\text{Aerial}}$
w/o DD	w/ DD	w/o VF	w/ VF		
				29.1	49.2
✓				32.1	50.4
	✓			35.9	52.4
	✓	✓		41.3	54.5
	✓		✓	**43.4**	**55.6**

Table 4. Results (%) of various number of text descriptions.

Descriptions Number	Potsdam	Vaihingen	WHU$^{\text{Aerial}}$
0	47.1	29.1	49.2
1	47.9	32.5	50.9
3	48.3	32.1	51.5
5	48.9	**35.9**	**52.4**
7	**49.2**	33.6	52.0

Table 5. Results (%) of different methods with our TMPA.

Methods	Potsdam	Vaihingen	WHU$^{\text{Aerial}}$
SCLIP [55]	36.6	28.4	33.4
+ TMPA	**41.2** ↑ **4.6**	**35.9** ↑**7.5**	**42.9** ↑**9.5**
ClearCLIP [29]	40.9	27.3	36.6
+ TMPA	**45.2** ↑ **4.3**	**32.0** ↑**4.7**	**47.1** ↑**10.5**
CASS [27]	43.8	33.5	38.9
+ TMPA	**48.5** ↑ **4.7**	**42.4** ↑**8.9**	**57.4** ↑**18.5**

5. Conclusion

In this paper, we proposed an effective TMPA method to alleviate issue of textual ambiguity in open-vocabulary remote sensing image semantic segmentation (OVRSIS) and boost its performance. To this end, a Context-Aware Text Prompt Generator (Cat-Prompt) is introduced to generate diverse task-specific descriptions, while a Visual-Guided Test-Time Adaptation (VGTA) strategy is proposed to refine prompt embeddings based on high-confidence visual features. Particularly, a pixel-level entropy minimization loss is formulated to dynamically optimize prompt embeddings during inference, accounting for better vision-language alignment for the regions with high uncertainty predictions. Extensive experiments on 17 remote sensing datasets by combining with various existing methods various clearly demonstrate the effectiveness and generalization of our TMPA. We hope that our TMPA can encourage more studies on effective text prompt in OVRSIS, while we will apply our TMPA to other open-vocabulary image understanding tasks (e.g., open-vocabulary object detection [39, 61] and open-vocabulary scene parsing [32, 63]).

Acknowledgments

The work was sponsored by the National Natural Science Foundation of China under Grants U23B2049, 62276186, National Natural Science Foundation of China (Grant No. 22527901) through the National Major Research Instrumentation Program, Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (JYB2025XDXM503).

References

[1] Roberto Alcover-Couso, Marcos Escudero-Viñolo, Juan C SanMiguel, and Jesus Bescos. Vlms meet uda: Boosting transferability of open vocabulary segmentation with unsupervised domain adaptation. *arXiv preprint arXiv:2412.09240*, 2024. 5

[2] Anthropic. Introducing claude 4. https://www.anthropic.com/news/claude-4. 3

[3] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *CVPR*, 2024. 3

[4] Walid Bousselham, Felix Petersen, Vittorio Ferrari, and Hilde Kuehne. Grounding everything: Emerging localization properties in vision-language transformers. In *CVPR*, 2024. 3, 6, 7

[5] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Cocostuff: Thing and stuff classes in context. In *CVPR*, 2018. 4

[6] Wenxiao Cai, Ke Jin, Jinyan Hou, Cong Guo, Letian Wu, and Wankou Yang. Vdd: Varied drone dataset for semantic segmentation. *J. Vis. Commun. Image Represent.*, 2025. 6, 1, 3

[7] Qinglong Cao, Yuntian Chen, Chao Ma, and Xiaokang Yang. Open-vocabulary high-resolution remote sensing image semantic segmentation. *IEEE TGRS*, 2025. 1, 2, 3, 7

[8] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 1, 3

[9] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017. 1

[10] Yu Chen, Yao Wang, Peng Lu, Yisong Chen, and Guoping Wang. Large-scale structure from motion with semantic constraints of aerial images. In *PRCV*, 2018. 6, 8, 1, 3

[11] Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *CVPR*, 2024. 3

[12] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 2, 4, 5, 3

[13] Marius Cordts, Mohamed Omran, Sebastian Ramos, Timo Rehfeld, Markus Enzweiler, Rodrigo Benenson, Uwe Franke, Stefan Roth, and Bernt Schiele. The cityscapes dataset for semantic urban scene understanding. In *CVPR*, 2016. 4

[14] Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. *NeurIPS*, 2013. 3

[15] Jian Ding, Nan Xue, Gui-Song Xia, and Dengxin Dai. Decoupling zero-shot semantic segmentation. In *CVPR*, 2022. 3

[16] Xiaoyi Dong, Jianmin Bao, Yinglin Zheng, Ting Zhang, Dongdong Chen, Hao Yang, Ming Zeng, Weiming Zhang, Lu Yuan, Dong Chen, et al. Maskclip: Masked self-distillation advances contrastive language-image pretraining. In *CVPR*, 2023. 3

[17] Chun-Mei Feng, Yuanyang He, Jian Zou, Salman Khan, Huan Xiong, Zhen Li, Wangmeng Zuo, Rick Siow Mong Goh, and Yong Liu. Diffusion-enhanced test-time adaptation with text and image augmentation. *IJCV*, 2025. 3

[18] B. C. Forster. An examination of some problems and solutions in monitoring urban areas from satellite platforms. *IJRS*, 1985. 1

[19] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3

[20] Ritwik Gupta, Richard Hosfelt, Sandra Sajeev, Nirav Patel, Bryce Goodman, Jigar Doshi, Eric Heim, Howie Choset, and Matthew Gaston. xbd: A dataset for assessing building damage from satellite imagery. *arXiv preprint arXiv:1911.09296*, 2019. 6

[21] Gustavo A Vargas Hakim, David Osowiechi, Mehrdad Noori, Milad Cheraghalikhani, Ali Bahri, Moslem Yazdanpanah, Ismail Ben Ayed, and Christian Desrosiers. Clipartt: Adaptation of clip to new domains at test time. In *WACV*, 2025. 3

[22] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3

[23] Mahesh Kumar Jat, Pradeep Kumar Garg, and Deepak Khare. Monitoring and modelling of urban sprawl using remote sensing and gis techniques. *International journal of Applied earth Observation and Geoinformation*, 2008. 1

[24] Shunping Ji, Shiqing Wei, and Meng Lu. Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. *IEEE TGRS*, 2018. 4, 5, 6, 1, 2, 3

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. 1

[26] Dahyun Kang and Minsu Cho. In defense of lazy visual grounding for open-vocabulary semantic segmentation. In *ECCV*, 2024. 3

[27] Chanyoung Kim, Dayun Ju, Woojung Han, Ming-Hsuan Yang, and Seong Jae Hwang. Distilling spectral graph for

object-context aware open-vocabulary semantic segmentation. In *CVPR*, 2025. 2, 4, 6, 7, 8, 3

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 6

[29] Mengcheng Lan, Chaofeng Chen, Yiping Ke, Xinjiang Wang, Litong Feng, and Wayne Zhang. Clearclip: Decomposing clip representations for dense vision-language inference. In *ECCV*, 2024. 3, 4, 6, 7, 8

[30] Bingyu Li, Haocheng Dong, Da Zhang, Zhiyuan Zhao, Junyu Gao, and Xuelong Li. Exploring efficient open-vocabulary segmentation in the remote sensing. *arXiv preprint arXiv:2509.12040*, 2025. 1, 2, 3, 7

[31] Kaiyu Li, Ruixun Liu, Xiangyong Cao, Xueru Bai, Feng Zhou, Deyu Meng, and Zhi Wang. Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In *CVPR*, 2025. 1, 2, 3, 4, 6, 7, 8

[32] Rongjie Li, Songyang Zhang, Dahua Lin, Kai Chen, and Xuming He. From pixels to graphs: Open-vocabulary scene graph generation with vision-language models. In *CVPR*, 2024. 8

[33] Chang Liu, Xudong Jiang, and Henghui Ding. Primitivenet: decomposing the global constraints for referring segmentation. *Visual Intelligence*, 2024. 3

[34] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *CVPR*, 2024. 3

[35] Yuanyuan Liu, Shaoze Feng, Shuyang Liu, Yibing Zhan, Dapeng Tao, Zijing Chen, and Zhe Chen. Sample-cohesive pose-aware contrastive facial representation learning. *IJCV*, 2025. 1

[36] Huaishao Luo, Junwei Bao, Youzheng Wu, Xiaodong He, and Tianrui Li. Segclip: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. In *ICML*, 2023. 3

[37] Ye Lyu, George Vosselman, Gui-Song Xia, Alper Yilmaz, and Michael Ying Yang. Uavid: A semantic segmentation dataset for uav imagery. *ISPRS*, 2020. 6, 1, 3

[38] Emmanuel Maggiori, Yuliya Tarabalka, Guillaume Charpiat, and Pierre Alliez. Can semantic labeling methods generalize to any city? the inria aerial image labeling benchmark. In *IGARSS*, 2017. 6

[39] Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. *NeurIPS*, 2023. 8

[40] Shambhavi Mishra, Julio Silva-Rodrıguez, Ismail Ben Ayed, Marco Pedersoli, and Jose Dolz. Words matter: Leveraging individual text embeddings for code generation in clip test-time adaptation. *arXiv preprint arXiv:2411.17002*, 2024. 3

[41] Volodymyr Mnih. *Machine learning for aerial image labeling*. University of Toronto, 2013. 6

[42] Roozbeh Mottaghi, Xianjie Chen, Xiaobai Liu, Nam-Gyu Cho, Seong-Whan Lee, Sanja Fidler, Raquel Urtasun, and Alan Yuille. The role of context for object detection and semantic segmentation in the wild. In *CVPR*, 2014. 4

[43] Jishnu Mukhoti, Tsung-Yu Lin, Omid Poursaeed, Rui Wang, Ashish Shah, Philip HS Torr, and Ser-Nam Lim. Open vocabulary semantic segmentation with patch aligned contrastive learning. In *CVPR*, 2023. 3

[44] Mehrdad Noori, David Osowiechi, Gustavo Adolfo Vargas Hakim, Ali Bahri, Moslem Yazdanpanah, Sahar Dastani, Farzad Beizaee, Ismail Ben Ayed, and Christian Desrosiers. Test-time adaptation of vision-language models for open-vocabulary semantic segmentation. *NeurIPS*, 2025. 3, 6, 7

[45] OpenAI. Gpt-5 system card. https://openai.com/index/gpt-5-system-card/. 3

[46] Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. What does a platypus look like? generating customized prompts for zero-shot image classification. In *ICCV*, 2023. 5

[47] 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 *ICML*, 2021. 1, 3, 4, 6

[48] Kanchana Ranasinghe, Brandon McKinzie, Sachin Ravi, Yinfei Yang, Alexander Toshev, and Jonathon Shlens. Perceptual grouping in contrastive vision-language models. In *ICCV*, 2023. 3

[49] Chen Ru, Si-Bo Duan, Xiao-Guang Jiang, Zhao-Liang Li, Yazhen Jiang, Huazhong Ren, Pei Leng, and Maofang Gao. Land surface temperature retrieval from landsat 8 thermal infrared data over urban areas considering geometry effect: Method and application. *IEEE TGRS*, 2021. 1

[50] Tong Shao, Zhuotao Tian, Hang Zhao, and Jingyong Su. Explore the potential of clip for training-free open vocabulary semantic segmentation. In *ECCV*, 2024. 3

[51] Elaine Sui, Xiaohan Wang, and Serena Yeung-Levy. Just shift it: Test-time prototype shifting for zero-shot generalization with vision-language models. In *WACV*, 2025. 3

[52] Shuyang Sun, Runjia Li, Philip Torr, Xiuye Gu, and Siyang Li. Clip as rnn: Segment countless visual concepts without training endeavor. In *CVPR*, 2024. 3

[53] Adam Van Etten, Dave Lindenbaum, and Todd M Bacastow. Spacenet: A remote sensing dataset and challenge series. *arXiv preprint arXiv:1807.01232*, 2018. 6

[54] Dequan Wang, Evan Shelhamer, Shaoteng Liu, Bruno Olshausen, and Trevor Darrell. Tent: Fully test-time adaptation by entropy minimization. *arXiv preprint arXiv:2006.10726*, 2020. 3

[55] Feng Wang, Jieru Mei, and Alan Yuille. Sclip: Rethinking self-attention for dense vision-language inference. In *ECCV*, 2024. 3, 6, 7, 8, 4

[56] Junjue Wang, Zhuo Zheng, Ailong Ma, Xiaoyan Lu, and Yanfei Zhong. Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. *arXiv preprint arXiv:2110.08733*, 2021. 2, 6

[57] Syed Waqas Zamir, Aditya Arora, Akshita Gupta, Salman Khan, Guolei Sun, Fahad Shahbaz Khan, Fan Zhu, Ling Shao, Gui-Song Xia, and Xiang Bai. isaid: A large-scale dataset for instance segmentation in aerial images. In *CVPR*, 2019. 6, 7

[58] Junshi Xia, Naoto Yokoya, Bruno Adriano, and Clifford Broni-Bediako. Openearthmap: A benchmark dataset for

global high-resolution land cover mapping. In *WACV*, 2023. 6

[59] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. San: side adapter network for open-vocabulary semantic segmentation. *IEEE TPAMI*, 2023. 3

[60] Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *CVPR*, 2023. 3

[61] Lewei Yao, Jianhua Han, Youpeng Wen, Xiaodan Liang, Dan Xu, Wei Zhang, Zhenguo Li, Chunjing Xu, and Hang Xu. Detclip: Dictionary-enriched visual-concept paralleled pretraining for open-world detection. *NeurIPS*, 2022. 8

[62] Chengyang Ye, Yunzhi Zhuge, and Pingping Zhang. Towards open-vocabulary remote sensing image semantic segmentation. In *AAAI*, 2025. 3

[63] Hang Zhao, Xavier Puig, Bolei Zhou, Sanja Fidler, and Antonio Torralba. Open vocabulary scene parsing. In *ICCV*, 2017. 8

[64] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 1

[65] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In *ECCV*, 2022. 6, 7

[66] Qiqi Zhu, Yanan Zhang, Lizeng Wang, Yanfei Zhong, Qingfeng Guan, Xiaoyan Lu, Liangpei Zhang, and Deren Li. A global context-aware and batch-independent network for road extraction from vhr satellite imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 2021. 6