Label-Free Text Prototype Adaptation for Open Vocabulary Segmentation

Keli Wang[1], Meixuan Li[1], Tianyu Li[1]✉, and Guoqing Wang[1]

University of Electronic Science and Technology of China (UESTC), Chengdu, China
{tylisky, gqwang0420}@uestc.edu.cn,

Abstract. Vision-language models (VLMs) have recently been adopted for open-vocabulary segmentation by aligning visual features with text prototypes. However, this alignment often degrades under domain shift, as text prototypes encode language priors learned from natural images that may not hold in the target domain. To address this challenge, we propose VPTA, a framework that adapts text prototypes to the target domain to mitigate language prior bias using only unlabeled target-domain images. VPTA gathers reliable visual evidence from confident pixels, estimates per-class evidence reliability to drive class-dependent prototype updates, and explicitly preserves prototype separation to prevent collapse between similar categories. This lightweight adaptation proceeds through a small number of iterations with negligible computational overhead. We conduct extensive evaluations under diverse domain shifts across three representative domains, including remote sensing, autonomous driving, and natural scene segmentation. VPTA achieves state-of-the-art performance on ten remote sensing datasets and consistently improves segmentation performance on autonomous driving and natural scene benchmarks. These results demonstrate the strong generalization ability of VPTA and highlight its effectiveness as a practical solution for label-free prototype adaptation in vision-language models.

Keywords: vision language models · open vocabulary semantic segmentation · label-free adaptation

1 Introduction

CLIP-based vision-language models learn to align images and text in a shared embedding space [20]. This alignment enables open-vocabulary recognition, where a model can recognize unseen categories by matching image features to text embeddings without retraining [20]. Recently, this capability has been extended from image-level recognition to dense prediction [10, 15, 26]. In open-vocabulary semantic segmentation, each pixel is classified by comparing its visual feature with a set of text prototypes [10, 15, 26]. This formulation is naturally scalable: new categories can be introduced by changing only the text side, without retraining a closed-set classifier head [15].

✉ Corresponding author.

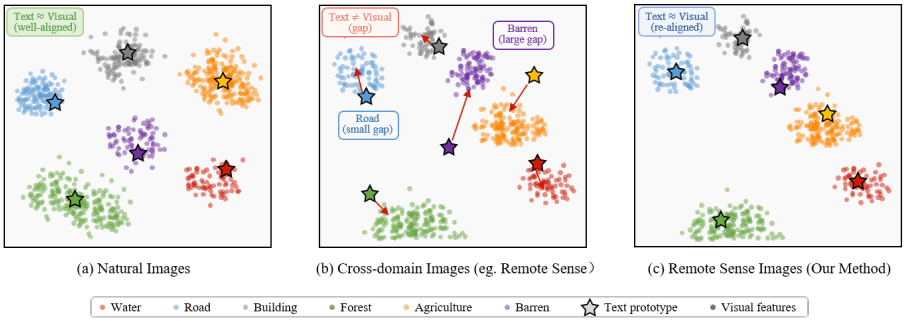

Fig. 1: t-SNE visualization of text prototypes (stars) and visual features (dots). In natural images (a), prototypes align with their visual clusters. In the target domain (b), a class-dependent gap opens between text and visual features. After VPTA (c), each prototype moves toward its visual cluster.

Despite its simplicity, CLIP-based open-vocabulary segmentation is highly sensitive to the quality of text prototypes. A text encoder maps each category into a prototype, while a visual encoder maps each pixel into the same space; prediction reduces to a nearest-prototype match [10, 15, 26]. When text prototypes are well aligned with the target visual patterns, this mechanism works remarkably well. However, once prototypes become misaligned, errors can arise even if pixel features are strong, since the prototypes serve as the reference points that ultimately determine the decision boundary.

This sensitivity becomes particularly severe under domain shift. CLIP is primarily trained on large-scale data dominated by natural images and their accompanying text [20], so its text embeddings are implicitly calibrated to align with natural-image visual embeddings. When target images follow different visual statistics (e.g., remote sensing imagery or autonomous driving scenes), pixel embeddings from the target domain may form clusters that no longer align with the corresponding text prototypes. We argue that a key reason is *language prior bias* induced by the frozen text encoder. Text prototypes reflect frequent contexts in the pretraining data [20], which introduces a prior over how each category "should" look. Under domain shift, these frozen prototypes become biased anchors. Pixels can be pulled toward categories favored by the prior rather than those correct for the target domain. Importantly, even with strong pixel features, biased prototypes can still dominate decisions because they define the matching targets.

Fig. 1 illustrates this phenomenon. In the target domain, a class-dependent gap appears between text prototypes and the corresponding visual clusters. This observation suggests a direct remedy: instead of forcing the target-domain visuals to fit frozen text anchors, we should *adapt the prototypes* to better reflect target-domain evidence.

A natural direction is to adapt the model to the target domain. Existing approaches typically rely on labeled target data. Test-time adaptation methods

update model parameters using only target images [22]. However, these strategies sit uneasily with open-vocabulary segmentation, which requires an open vocabulary, frozen model weights, and the plug-and-play flexibility of text-driven prediction. Prompt learning shows that manipulating the text side can strongly affect vision-language behavior [30], but most methods still require supervision.

This paper therefore asks a simple question: *Can we mitigate language prior bias without labels, using only target images available at test-time?* We introduce Visual Prototype-guided Text Adaptation (VPTA), a label-free framework that adapts CLIP text prototypes to a target domain using only unlabeled images. VPTA keeps all model weights frozen and updates only the text prototypes. It uses the model's current predictions to collect reliable target-domain evidence and then shifts each prototype toward this evidence in a controlled manner.

VPTA is built on three key ideas. First, it mines visual evidence from confident pixels to reduce the influence of noisy pseudo labels. Second, it estimates a per-class reliability score to modulate class-dependent update strength. Third, it preserves prototype separation to prevent collapse among semantically similar categories. Applying these steps for a few rounds yields stable adaptation without any supervision.

We evaluate VPTA under diverse domain shifts and observe consistent improvements with fixed hyperparameters. We summarize our contributions as follows:

- We propose VPTA, a *label-free* framework that refines CLIP text prototypes using only unlabeled target images, mitigating language prior bias under domain shift.
- We develop a reliability-controlled refinement scheme that mines confident pixels via margin filtering, estimates per-class evidence reliability, and performs class-dependent prototype updates with separation preservation, enabling stable label-free adaptation.
- We demonstrate strong performance across diverse shifts, achieving state-of-the-art results on ten remote sensing datasets and consistent gains on autonomous driving and natural scene benchmarks.

2 Related Work

Open vocabulary semantic segmentation. Open vocabulary segmentation classifies pixels by matching dense visual features to language embeddings, allowing novel categories to be introduced through text alone. Early methods require explicit supervision. LSeg trains a pixel encoder with language-guided mask annotations [15], and OpenSeg scales to region-text pairs from grounding data [10]. Both demonstrate strong transfer but depend on costly labeling pipelines.

More recent work builds on a frozen CLIP backbone and introduces task-specific components for dense prediction. SAN adds a side adapter trained on mask labels [26], CAT-Seg aggregates richer text-to-pixel interactions [6], and

FC-CLIP attaches a panoptic-style head [27]. Training-free variants instead improve feature quality directly. ClearCLIP removes noisy self-attention residuals [13], while CLIP-DINOiser fuses CLIP with DINO features for better spatial coherence [24]. All these efforts enhance the visual representation while keeping text prototypes unchanged. When the target domain departs from pretraining, however, even improved visual features can be matched to misaligned prototypes. VPTA addresses this complementary failure mode by adapting the text prototypes after visual features have been extracted, without changing the visual encoder.

Adapting vision-language models. A large body of work adapts VLMs by learning continuous prompts from labeled examples. CoOp optimises task-specific text prompts [31], CoCoOp conditions prompts on each input image [30], and MaPLe extends prompt learning to both text and visual branches [12]. Feature adapter methods such as CLIP-Adapter [9] and Tip-Adapter [28] insert lightweight layers but still require a small labeled set and modify model parameters.

Test-time methods reduce label dependency further. TPT minimises prediction entropy over augmentations of a single image [21], and PromptAlign aligns token-level statistics at test-time [1]. These methods share the label-free motivation with our work but differ in three important respects. First, they operate per image and reset afterwards, so no class-level knowledge accumulates across the target set. Second, they were designed for image classification and do not naturally extend to dense prediction, where thousands of pixel-level decisions must be made per image. Third, they modify learnable prompt tokens while the resulting text embeddings remain implicit, whereas VPTA directly adapts the prototypes that anchor every pixel-level decision.

Open vocabulary segmentation in shifted domains. Shifted domains pose an additional challenge because the same category name may correspond to fundamentally different pixel-level appearances, making prototype alignment critical. Remote sensing is a representative case. RemoteCLIP and GeoRSCLIP pretrain CLIP-style models on domain-specific data to improve image-level alignment [18, 29]. OVRS introduces spatially aware modules for remote sensing segmentation [4], RSKT-Seg transfers knowledge from CLIP to strengthen pixel features [14], and SegEarth-OV constructs a training-free pipeline around frozen CLIP with carefully designed components [17]. These methods mainly improve the visual encoder or the segmentation head while text prototypes are still taken from the frozen text encoder.

VPTA is complementary. It treats frozen prototypes as the key bottleneck under domain shift and mitigates language prior bias by updating prototypes from unlabeled target images with class-dependent step sizes and separation preservation. Because only the prototypes are modified, VPTA does not update the visual backbone. In addition to SegEarth-OV, we verify the plug-in use of VPTA on ClearCLIP and GEM, and also test RemoteCLIP as an alternative backbone in our experiments.

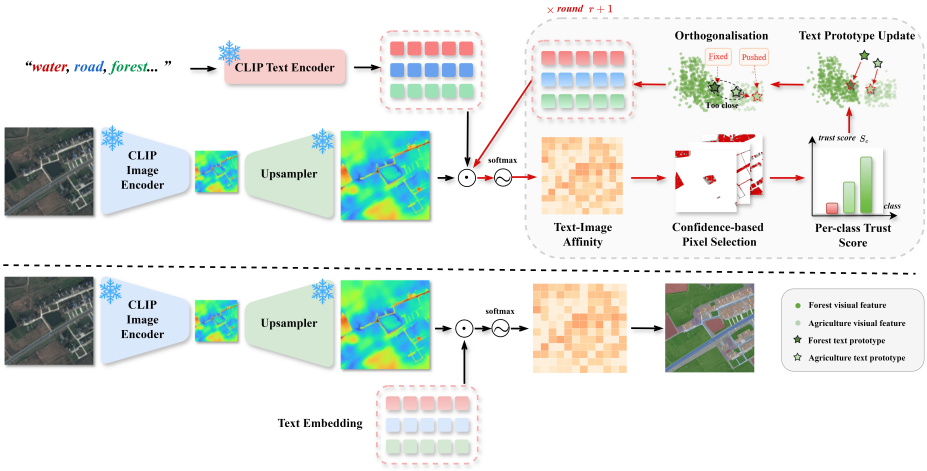

Fig. 2: Overview of VPTA. The top half illustrates the refinement loop, which iterates over unlabeled target-domain images to update text prototypes via confidence-based pixel selection, per-class reliability score estimation, text prototype update, and orthogonalization. The bottom half shows inference using the final adapted prototypes.

3 Method

3.1 Overview

A frozen CLIP model provides a text encoder E_t and a visual encoder E_v [20]. Both map inputs to a shared d-dimensional feature space. Given K categories $\{y_1, y_2, \ldots, y_K\}$, we pass each name through the text encoder to obtain an initial text prototype $\mathbf{t}_c^{(0)} = E_t(\texttt{prompt}(y_c))$ for class $c \in \{1, \ldots, K\}$. We also have N unlabeled images from the target domain. Our goal is to adapt the frozen text prototypes so that they better match target-domain visual features, improving open-vocabulary segmentation without any labeled data and without updating model weights.

Under domain shift, CLIP text prototypes can encode a visual prior that is mismatched with the target domain [20]. This language prior bias makes pixel-to-text matching unreliable. A simple approach is to segment the target images with frozen prototypes, average the assigned pixel features per class, and shift each prototype toward this visual estimate. However, pseudo assignments produced by biased prototypes contain noise. Uniform refinement then creates a feedback loop: early misassigned pixels are averaged into the class prototype, the shifted prototype makes similar pixels more likely to be assigned to the same wrong class in the next round, and the error keeps reinforcing itself. Some classes provide strong target evidence, while others provide few confident pixels and are dominated by confusion with similar categories.

We introduce VPTA, a label-free framework that mitigates language prior bias using only target images. Each refinement round has two stages. First, we

estimate how reliable the target-domain evidence is for each class using confidence filtered pixels (Sec. 3.2). Second, we update each text prototype with a class-dependent step size and preserve separation between prototypes to prevent collapse (Sec. 3.3). The reliability score acts as a control signal throughout. Well supported classes are adapted more, while poorly supported classes remain close to their CLIP starting point(See Fig. 2).

3.2 Evidence Reliability Estimation

Uniform prototype refinement. At each round we segment the unlabeled target images using the current prototypes. Every pixel is assigned to the class with the highest cosine similarity, as commonly done in CLIP-based open-vocabulary segmentation [10, 15, 26]. We average the visual features of pixels assigned to class c to form a visual prototype $\mathbf{v}_c$ and apply ℓ_2 normalization. This baseline can move prototypes toward the target domain, but it is unstable under domain shift because it trusts every assigned pixel equally.

The main errors come from ambiguous regions. Pixels near class boundaries often have similar cosine scores for multiple classes and are easily misassigned. Small or rare objects face a related issue. They occupy few pixels and their features are influenced by surrounding context, which makes the top class score less decisive. Including these unreliable pixels in $\mathbf{v}_c$ shifts the estimate away from the true class center. Over rounds, such errors accumulate and can distort prototypes, especially for hard classes.

We therefore treat only reliable pseudo assignments equally. Instead, we keep only reliable evidence and quantify its strength per class.

Confidence based pixel selection. We retain only the pixels for which the current prototypes are confident and discard the rest. Confidence is measured by the prediction margin m_i, defined as the cosine similarity gap between the top class and the runner up for pixel i,

$$m_i = \cos(\mathbf{f}_i, \mathbf{t}_{c_1}) - \cos(\mathbf{f}_i, \mathbf{t}_{c_2}), \tag{1}$$

where $\mathbf{f}_i$ is the visual feature of pixel i and c_1, c_2 are the first- and second-ranked classes. A large margin means the pixel aligns clearly with one prototype. A pixel enters the visual prototype only when $m_i > \tau_m$ for a fixed threshold τ_m.

Let $\mathcal{P}_c = \{i : \hat{c}_i = c \text{ and } m_i > \tau_m\}$ be the surviving pixel set for class c. After filtering we recompute the visual prototype from these pixels alone,

$$\mathbf{v}_c = \frac{\bar{\mathbf{v}}_c}{\|\bar{\mathbf{v}}_c\|}, \qquad \bar{\mathbf{v}}_c = \frac{1}{|\mathcal{P}_c|} \sum_{i \in \mathcal{P}_c} \mathbf{f}_i. \tag{2}$$

This filtered $\mathbf{v}_c$ replaces the unfiltered estimate in all later steps.

Filtering reduces noise but exposes a second challenge. Evidence quality is highly uneven across classes. Some classes keep many confident pixels, while others keep only a few. We therefore require a class-wise measure of evidence reliability.

Per-class reliability score. We compute a reliability score s_c for each class using three statistics of $\mathcal{P}_c$. The first is the number of surviving pixels $n_c = |\mathcal{P}_c|$. The second is their mean margin $\bar{m}_c$. The third is the standard deviation of the margin σ_c. The score is

$$s_c = \min\Big(1,\ \frac{n_c}{n_{\min}}\Big) \cdot \frac{\bar{m}_c}{\bar{m}_c + \sigma_c + \epsilon}, \tag{3}$$

where $n_{\min}$ saturates the sample sufficiency term and ϵ is a small constant. The first factor penalizes classes with too few pixels. The second factor favors evidence that is both strong and consistent. Classes with many confident pixels and stable margins have s_c close to 1, while classes with weak or noisy evidence have s_c close to 0. We use s_c as the control signal for prototype adaptation in the next stage.

3.3 Prototype adaptation with Reliability Control

With reliability scores in hand we adapt each text prototype by interpolating it toward the target visual prototype and reprojecting onto the unit sphere,

$$\mathbf{t}_c^{(r)} = \frac{(1-\beta_c)\,\mathbf{t}_c^{(r-1)} + \beta_c\,\mathbf{v}_c}{\|(1-\beta_c)\,\mathbf{t}_c^{(r-1)} + \beta_c\,\mathbf{v}_c\|}, \tag{4}$$

where $\beta_c = \beta_{\max} \cdot s_c$ and $\beta_{\max}$ controls the maximum step size. Since cosine similarity is used for matching, the direction of the prototype is what matters. Hence, the update modifies only the direction of the prototype on the unit hypersphere. High reliability classes take a larger step toward target evidence. Low reliability classes move conservatively, which limits error amplification.

Preventing collapse of similar prototypes. The update in Eq. 4 adapts each class independently. When two classes have similar target appearances, their visual prototypes $\mathbf{v}_c$ can become close and the adapted text prototypes may drift toward the same region. This reduces class separation and increases confusion, which is especially harmful for fine grained categories.

We address this by preserving separation in a reliability-aware manner. We adopt an orthogonalization step based on a Gram–Schmidt style process to remove overlapping components. However, the result depends on the processing order. If an unreliable class is processed early, it can anchor the space and force reliable classes to yield.

We therefore process classes in descending order of reliability. Reliable prototypes are kept as anchors, while less reliable ones are adjusted around them. Let $\tilde{\mathbf{t}}_c = \mathbf{t}_c^{(r)}$ denote the updated prototype from Eq. 4. For class c processed after all higher reliability classes in the ordered set $\mathcal{O}(c)$ we compute

$$\hat{\mathbf{t}}_c = \tilde{\mathbf{t}}_c - \sum_{j \in \mathcal{O}(c)} w_{jc}\, \frac{\langle \tilde{\mathbf{t}}_c,\, \hat{\mathbf{t}}_j \rangle}{\|\hat{\mathbf{t}}_j\|^2}\, \hat{\mathbf{t}}_j, \tag{5}$$

where the weight

$$w_{jc} = \frac{s_j}{s_j + s_c + \epsilon} \tag{6}$$

controls how much class c yields to class j. When j has much higher reliability, w_{jc} approaches 1 and c yields strongly. When reliabilities are similar, w_{jc} is near 0.5, so neither class has a dominant influence. After orthogonalization, we renormalize each prototype to the unit hypersphere,

$$\mathbf{t}_c^{(r)} \leftarrow \frac{\hat{\mathbf{t}}_c}{\|\hat{\mathbf{t}}_c\|}. \tag{7}$$

This step maintains separation while respecting evidence quality.

Iterative refinement. The two stages above form one refinement round. At the start of round $r+1$ we re-segment the unlabeled images with the adapted prototypes $\mathbf{t}_c^{(r)}$, recompute filtered visual prototypes and reliability scores, and repeat the update and separation steps. Iterative refinement is important because the earliest pseudo assignments are generated by the most biased prototypes and are therefore the noisiest. Using small controlled steps allows prototypes to improve gradually. As prototypes improve, the pixel evidence becomes cleaner and the reliability scores are refreshed, which further stabilizes adaptation. We run the loop for R rounds. The final prototypes $\mathbf{t}_c^* = \mathbf{t}_c^{(R)}$ are used directly for inference.

4 Experiments

4.1 Setup

Datasets. We evaluate on twelve datasets spanning remote sensing and autonomous driving to show that our method generalizes across imaging conditions. For satellite imagery, we use ISPRS Vaihingen[1], ISPRS Potsdam[1], LoveDA [23], and OpenEarthMap [25]. For UAV imagery, we use UDD5 [5], UAVid [19], and VDD [3]. For single-class extraction, we include WHU Aerial [11] (building), DeepGlobe[2] (road), and WBS-SI[3] (water body). For all ten RS datasets, the training split is used without labels during adaptation and ground-truth masks are accessed only for evaluation. We further test on PASCAL VOC 2012 [8] and Cityscapes [7] to verify that gains are not specific to aerial imagery.

Following SegEarth-OV, Vaihingen and Potsdam are loaded as IR-R-G composites, so vegetation appears red rather than green, as visible in Fig. 3. All dataset preparation follows the SegEarth-OV protocol. The only exception is DeepGlobe(road), where only a validation split is publicly available; we divide

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

[2] http://deepglobe.org

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

it into 85% for unlabeled adaptation and 15% for evaluation. All other datasets are processed identically to SegEarth-OV, ensuring a fair comparison under the same experimental setting.

Metrics. Semantic segmentation performance is measured by mIoU (%) over all categories. For single-class extraction tasks (WHU Aerial, DeepGlobe, WBS-SI), we report the foreground IoU directly.

Implementation. We build on the SegEarth-OV pipeline with a frozen CLIP ViT-B/16 backbone and a frozen universal upsampler. For text prompts, we follow SegEarth-OV and average 80 OpenAI ImageNet templates per class. For classes whose definitions vary across datasets, we apply a renaming strategy shared by all methods; *clutter* is renamed to *background* and *impervious surface* is expanded to a synonym set of {*impervious surface*, *parking lot*, *road*}, taking the highest-scoring sub-class as the class prediction. Our method only modifies text prototypes and leaves all network weights untouched. Hyperparameters are fixed across all twelve datasets: R=5 refinement rounds, $\beta_{\max}$=0.10, margin threshold τ_m=0.10, and coverage threshold $n_{\min}$=300.These defaults were chosen from experiments without using validation masks or per-dataset tuning. Performance stabilizes after round 5 and does not regress with additional rounds. Adaptation runs on two NVIDIA RTX 4090 GPUs and takes roughly 15 minutes on UAVid, with negligible inference overhead (6,650 vs. 6,660 GFLOPs).

Table 1: Open-vocabulary semantic segmentation (OVSS) results (mIoU %) comparing different methods on multi-class remote sensing datasets. For Vaihingen, A denotes the original-label protocol, and C follows the stronger naming protocol used in SegEarth-OV2 [16]. The detailed difference between A and C is discussed in Sec. 4.4 and Tab. 4.

Method	Venue	Remote Sensing (Multi-class)							
		Vaihingen		Potsdam	LoveDA	UDD5	UAVid	VDD	OpenEarthMap
		A	C						
CLIP [20]	ICML'21	10.3	10.8	14.5	12.4	9.5	10.9	14.2	12.0
ClearCLIP [13]	ECCV'24	27.3	36.2	40.9	32.4	41.8	36.2	39.3	31.0
OVRS [4]	TGRS'25	20.8	–	19.9	30.8	31.9	16.2	31.0	–
RSKT-Seg [14]	AAAI'26	25.6	–	28.6	29.6	34.0	17.4	33.4	–
SegEarth-OV [17]	CVPR'25	29.1	40.1	47.1	36.8	50.4	42.5	45.0	40.3
VPTA (Ours)	—	**37.2**(+8.1)	**41.5**(+1.4)	**48.5**(+1.4)	**40.5**(+3.7)	**52.0**(+1.6)	**44.3**(+1.8)	**46.4**(+1.4)	**41.2**(+0.9)

Table 2: Open-vocabulary semantic segmentation (OVSS) results (mIoU %) on single-class remote sensing extraction, autonomous driving, and natural scene datasets.

Method	Venue	RS (Single-class extraction)			Autonomous Driving	Natural Scene
		WHU Aerial	DeepGlobe	WBS-SI	Cityscapes	VOC 2012
CLIP [20]	ICML'21	17.7	3.9	18.6	6.5	49.4
ClearCLIP [13]	ECCV'24	36.6	5.7	44.9	16.2	53.1
OVRS [4]	TGRS'25	–	–	–	–	–
RSKT-Seg [14]	AAAI'26	–	–	–	–	–
SegEarth-OV [17]	CVPR'25	49.1	23.4	60.2	20.3	50.2
VPTA (Ours)	—	**56.3**(↑7.2)	**28.2**(↑4.8)	**61.9**(↑1.7)	**23.5**(↑3.2)	**56.3**(↑3.2)

Table 3: Component ablation (mIoU %). **P**: prototype refinement. **A**: pixel selection. **B**: per-class reliability score s_c. **C**: reliability-scaled step size $\beta_{\max} \cdot s_c$. **D**: orthogonalisation.

	Component					mIoU (%)			
	P	A	B	C	D	Vaihingen	LoveDA	WHU Aerial	Cityscapes
Baseline						29.1	36.8	49.1	20.3
+P						31.5(+2.4)	38.4(+1.6)	51.2(+2.1)	21.2(+0.9)
+P +A						32.9(+3.8)	39.0(+2.2)	52.4(+3.3)	21.7(+1.4)
+P +A +B +C						35.3(+6.2)	39.8(+3.0)	54.8(+5.7)	22.9(+2.6)
+P +A +B +D						34.9(+5.8)	39.6(+2.8)	54.4(+5.3)	22.6(+2.3)
+P +A +B +C +D (Ours)						**37.2**(+8.1)	**40.5**(+3.7)	**56.3**(+7.2)	**23.5**(+3.2)
Effect of pixel selection on reliability estimation									
+P +B +C +D (w/o A)						35.6(+6.5)	39.7(+2.9)	55.1(+6.0)	22.8(+2.5)
Effect of ordering									
+P +A +B +C +D†					†	36.4(+7.3)	40.1(+3.3)	55.8(+6.7)	23.2(+2.9)

† Random class ordering instead of reliability-descending order.

4.2 Main Results

Tab. 1 and Tab. 2 summarize the results across all twelve datasets.

Multi-class RS segmentation. VPTA ranks first on all seven datasets and outperforms SegEarth-OV under the reported protocols. On Vaihingen, the original-label setting A gives a large gain of +8.1 mIoU, while the stronger naming setting C gives a smaller but still positive gain of +1.4 mIoU. This confirms that VPTA benefits from better initial class names but still improves over the corresponding baseline under the same naming setting. On UAV datasets the margins are smaller (+1.4 to +1.8) but consistent across different altitudes and resolutions.

Single-class extraction. VPTA improves over SegEarth-OV on all three extraction tasks. The gain on WHU Aerial (+7.2) is the largest because building footprints have clear visual patterns that text prototypes can align to. DeepGlobe road extraction remains difficult for all methods due to thin and elongated geometry, yet VPTA still improves by +4.8.

Generalization beyond remote sensing. We evaluate on Cityscapes and PASCAL VOC 2012 to test whether VPTA generalizes to other domains. VPTA improves over the baseline on both datasets (+3.2 each) with the same hyperparameters used for remote sensing, confirming that the method is not tied to a specific domain.

4.3 Ablation Study

We ablate each component on Vaihingen, LoveDA, WHU Aerial, and Cityscapes in Tab. 3.

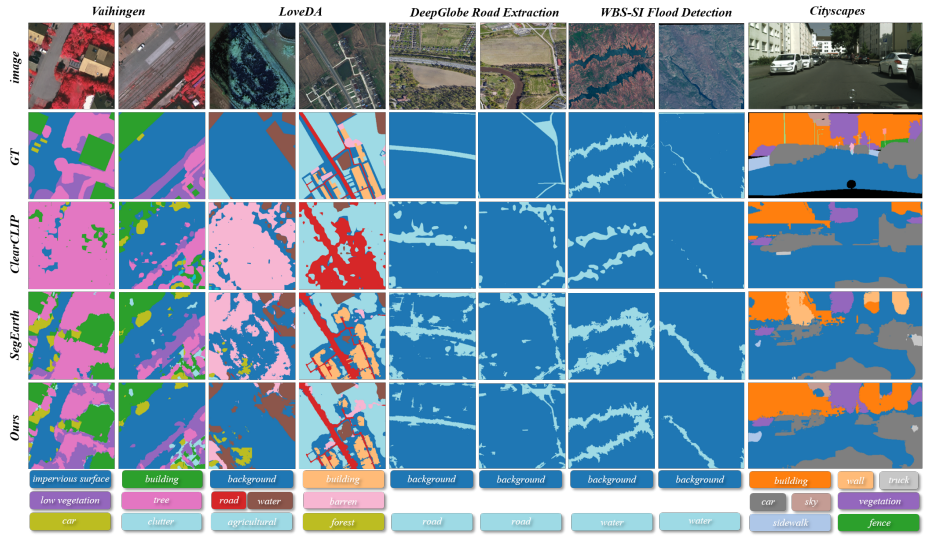

Fig. 3: Qualitative results on Vaihingen, LoveDA, DeepGlobe, WBS-SI, and Cityscapes (left to right). Rows show the input image, ground truth, SegEarth-OV, and VPTA. VPTA recovers fine-grained categories that SegEarth-OV collapses into a dominant class, and produces cleaner boundaries on thin structures such as roads and waterways.

Build-up analysis. Prototype refinement (+P) moves text prototypes toward target-domain visual prototypes with a fixed-step update, showing that even noisy visual evidence is useful. Pixel selection (+A) adds a further gain by removing unreliable boundary pixels. Adding the reliability score with scaled updates (+B +C) brings the largest single improvement, especially on Vaihingen (+2.4) where inter-class difficulty varies widely. Reliability-aware orthogonalization (+B +D) addresses a different problem. It prevents visually similar classes such as forest and agriculture from collapsing onto each other. Combining C and D outperforms either alone, confirming that they fix complementary failure modes.

Pixel selection and reliability quality. Removing pixel selection while keeping the reliability score (w/o A) drops performance below the full model. Without filtering, noisy boundary pixels inflate margin variance and degrade reliability score quality. Pixel selection is a prerequisite for accurate reliability estimation.

reliability-descending order. Replacing reliability-descending ordering with random ordering drops 0.4 to 0.8 mIoU across all four datasets. A poorly estimated prototype placed first becomes the anchor and forces better-estimated classes to bend around it.

4.4 Analysis

Sensitivity to class name choice. Open-vocabulary methods take class names as input, so naming directly affects performance. We test three naming conven-

Table 4: Performance Across Different Settings (A/B/C)

Method	Vaihingen(A)	Vaihingen(B)	Vaihingen(C)	Potsdam(A)	Potsdam(B)	Potsdam(C)
SegEarth-OV	29.1	38.8	40.1	35.5	47.1	48.5
Ours	**37.2**	**40.8**	**41.5**	**38.4**	**48.5**	**50.0**

tions on Vaihingen and Potsdam (Tab. 4). Set A uses the original ISPRS labels (impervious surface, building, low vegetation, tree, car, clutter). Set B splits impervious surface into {road, parking lot} and clutter into {clutter, background}, keeping the highest-scoring sub-class per pixel. Set C replaces impervious surface with road, low vegetation with grass, and clutter with {clutter, background}, following SegEarth-OV2 [16].

Since class names affect text-driven OVSS, we report both the original-label setting and stronger naming settings rather than using setting A alone as the headline. Simply changing names improves the baseline from 29.1 (A) to 40.1 (C) on Vaihingen without any model change. VPTA improves over the baseline under all three settings by +8.1 (A), +2.0 (B), and +1.4 (C). The gain is largest when initial prototypes are poorly aligned and smaller but consistent when names are already well chosen.This sensitivity supports our premise that text anchors matter.

Cross-domain and mixed-domain adaptation. Tab. 5 tests VPTA under two additional settings. In setting (a), we use LoveDA images to adapt prototypes and evaluate on Vaihingen. Performance drops from 40.1 to 21.9 mIoU. The main cause is spectral mismatch. Vaihingen uses IR-R-G composites where vegetation appears red, while LoveDA uses standard RGB where vegetation is green. Prototypes built from LoveDA encode green tree features that do not match red trees in Vaihingen, causing Tree IoU to collapse from 58.9 to 1.6. Building IoU also drops from 58.8 to 32.4, likely due to the appearance difference between Chinese urban buildings in LoveDA and European buildings in Vaihingen. This shows that VPTA adapts to the visual statistics of its input. When source and target differ in spectral bands, the adapted prototypes hurt rather than help.

In setting (b), we pool Vaihingen and Potsdam images for joint adaptation. Both datasets share the same IR-R-G format from the ISPRS benchmark. VPTA improves Vaihingen from 40.1 to 42.0 and Potsdam from 48.5 to 49.3. More same-domain images provide better prototype estimates without introducing any domain gap.

Text-visual misalignment varies across classes. Fig. 4a shows the cosine gap between each frozen text prototype and its visual cluster on LoveDA. Agriculture has the smallest gap (0.47) while Barren has the largest (0.98). Fig. 4b plots this gap against per-class IoU under frozen CLIP (Spearman ρ=−0.96). Classes with smaller gaps segment better, confirming that a uniform adaptation cannot fit all classes.

Table 5: Class-wise IoU (%) under cross-domain and mixed-domain settings on Vaihingen and Potsdam.

Evaluation Set	Imp. Surf.	Building	Low Veg.	Tree	Car	Clutter	**mIoU**
(a) Cross-Domain (LoveDA → Vaihingen)							
Vaihingen (SegEarth-OV)	56.9	58.8	25.6	58.9	32.3	7.2	40.1
Vaihingen (**Ours**)	**46.3**	**32.4**	**19.6**	**1.6**	**29.2**	**2.4**	**21.9**
(b) Mixed-Domain (Vaihingen + Potsdam)							
Vaihingen (SegEarth-OV)	56.9	58.8	25.6	58.9	32.3	7.2	40.1
Vaihingen (**Ours**)	**60.1**	**61.7**	**32.3**	**58.1**	**32.9**	**6.9**	**42.0**
Potsdam (SegEarth-OV)	63.1	59.9	55.9	46.9	48.8	16.3	48.5
Potsdam (**Ours**)	**64.2**	**63.9**	**56.3**	**46.4**	**48.1**	**16.9**	**49.3**

Table 6: Plug-in results on different OVSS pipelines and backbones. "Base" denotes the result before applying VPTA.

Method / pipeline	**Backbone / setting**	**Dataset**	**Metric**	**Base**	**+VPTA (Ours)**
SegEarth-OV [17]	ViT-B/16	Vaihingen	mIoU	29.1	**37.2** (+8.1)
SegEarth-OV [17]	ViT-B/16	LoveDA	mIoU	36.8	**40.5** (+3.7)
SegEarth-OV [17]	ViT-B/16	WHU Aerial	mIoU	49.1	**56.3** (+7.2)
SegEarth-OV [17]	RemoteCLIP [18]	Vaihingen	mIoU	21.9	**26.3** (+4.4)
SegEarth-OV [17]	RemoteCLIP [18]	LoveDA	mIoU	37.8	**39.6** (+1.8)
SegEarth-OV [17]	RemoteCLIP [18]	WHU Aerial	mIoU	43.7	**46.9** (+3.2)
ClearCLIP [13]	ViT-B/16	Vaihingen	mIoU	30.3	**37.5** (+7.2)
ClearCLIP [13]	ViT-B/16	LoveDA	mIoU	36.5	**39.4** (+2.9)
ClearCLIP [13]	ViT-B/16	WHU Aerial	mIoU	51.1	**54.0** (+2.9)
GEM [2]	ViT-B/16	Vaihingen	mIoU	29.3	**36.0** (+6.7)
GEM [2]	ViT-B/16	LoveDA	mIoU	35.1	**37.2** (+2.1)
GEM [2]	ViT-B/16	WHU Aerial	mIoU	49.3	**52.7** (+3.4)

Refinement diverges. Fig. 4c compares VPTA with three simplified variants over ten rounds. A large uniform step (β=0.30) locks in noise and drops below baseline after round 3. A small uniform step (β=0.10) avoids collapse but plateaus 2 mIoU below VPTA. Removing pixel selection also diverges. Only VPTA converges stably by letting the reliability score set each class's step size individually.

Hyperparameter sensitivity. Fig. 4d sweeps the three hyperparameters on LoveDA. τ_m varies less than 1.1 mIoU between 0.06 and 0.14. $n_{\min}$ stays within 1.0 mIoU across two orders of magnitude. $\beta_{\max}$ has a narrower plateau around 0.08 to 0.12. All default values fall inside their respective plateaus and are fixed across all twelve datasets.

Per-class reliability dynamics. Fig. 4e tracks reliability scores and IoU gains for four classes. Water, Road, and Agriculture start with moderate to high reliability and rise further as prototypes align, gaining +8.7, +4.6, and +2.9 IoU.

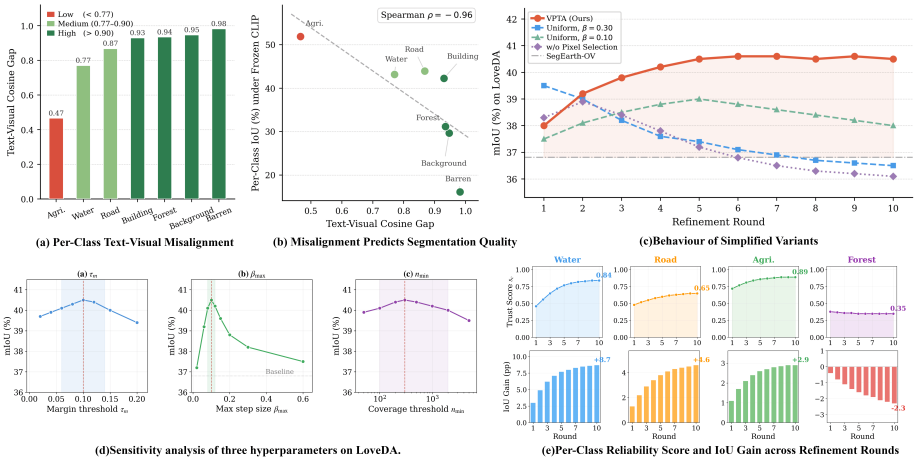

Fig. 4: Analysis on LoveDA. (a) Per-class text-visual cosine gap. (b) Gap vs. per-class IoU under frozen CLIP. (c) Convergence of VPTA and simplified variants. (d) Sensitivity to hyperparameters. (e) Per-class reliability score and IoU gain across rounds.

Forest is the exception. As Agriculture adapts, it claims Forest boundary pixels. Forest reliability stays low at 0.35 and IoU drops by 2.3, but the loss is contained because low reliability limits the step size.

Plug-in evaluation. We evaluate VPTA with SegEarth-OV [17], ClearCLIP [13], GEM [2], and RemoteCLIP [18]. As shown in Tab. 6, VPTA improves SegEarth-OV, ClearCLIP, and GEM in all tested settings, and remains effective with RemoteCLIP as the visual backbone. These results show that VPTA is not tied to a single SegEarth-OV implementation; this experiment aims to validate plug-in compatibility rather than re-rank all OVSS methods.

Table 7: Bootstrap CIs for small gains with 1000 resamples.

Dataset	N	Baseline	+VPTA (Ours)	Gain [95% CI]
Potsdam	2016	47.1	**48.5**	+1.4 [+1.2, +1.9]
UDD5	40	50.4	**52.0**	+1.6 [+1.0, +2.3]
UAVid	480	42.5	**44.3**	+1.8 [+1.5, +2.1]
VDD	80	45.0	**46.4**	+1.4 [+1.1, +2.0]
OpenEarthMap	384	40.3	**41.2**	+0.9 [+0.5, +1.3]

Statistical validation. Tab. 7 reports 95% bootstrap confidence intervals with 1000 resamples. All positive intervals confirm the reliability of small gains.

Sensitivity to $n_{\min}$. Tab. 8 compares fixed $n_{\min}$ values with a relative ratio rule $n_c < \rho HW$ at $\rho = 0.01$. VPTA is stable, and $n_{\min} = 300$ works well without per-dataset tuning.

Failure cases. VPTA degrades in three settings. On VDD, *Vehicle* drops by −6.5 pp because vehicles occupy only a few pixels at UAV altitude, leaving too

Table 8: Sensitivity to $n_{\min}$ across datasets.

Dataset	$n_{\min}$=30	**100**	**300**	**500**	**1000**	**2000**	**Rel. ratio=0.01**
Vaihingen	36.5	36.9	**37.2**	37.1	36.8	36.4	36.8
LoveDA	39.7	40.0	**40.5**	40.3	40.4	39.6	39.9
UAVid	43.5	44.0	44.3	44.2	**44.4**	43.7	43.7
OpenEarthMap	40.4	**41.3**	41.2	41.1	40.9	40.6	40.7

few high-margin pixels for a reliable visual prototype. Cross-domain adaptation also fails when spectral bands differ; adapting on LoveDA RGB and evaluating on Vaihingen IR-R-G drops mIoU by −18.2 pp because the adapted prototypes encode the wrong colour space. All three cases stem from the same root cause: VPTA depends on pseudo-label quality, and when margin filtering cannot recover clean signals, the method provides no adaptation.

5 Conclusion

We presented VPTA, a label-free framework that mitigates language prior bias in CLIP-based open-vocabulary segmentation by adapting text prototypes using only unlabeled target images. VPTA keeps all model weights frozen and refines prototypes through confidence-based evidence mining, per-class reliability estimation, and separation-preserving updates. With fixed hyperparameters across all twelve benchmarks, VPTA achieves state-of-the-art results on ten remote sensing datasets and consistent improvements on autonomous driving and natural scene benchmarks, demonstrating broad applicability across diverse domain shifts.

VPTA has clear limitations. When a class occupies very few pixels, margin filtering cannot recover enough evidence to form a reliable visual prototype, as observed for small vehicles in UAV imagery. Cross-domain adaptation also fails when the source and target differ in spectral characteristics, because prototypes built from one colour space encode visual cues that are invalid in another. Both cases trace back to the same root cause: the method relies on pseudo-label quality, and when confident pixels are scarce or misleading, the adaptation provides no correction. More broadly, VPTA adapts prototypes globally across the target set and does not account for intra-domain variation, which may limit its effectiveness when target images span a wide range of conditions within a single dataset. Extending VPTA with multi-scale evidence aggregation, or instance-level adaptation to handle rare classes and heterogeneous target domains is a promising direction for future work.

6 Acknowledgement

This work was supported in part by the National Natural Science Foundation of China under grant U2541205, U25A20535 and U23B2011, National Key Research and Development Program of China under grant 2025YFF0522500, the Sichuan Science and Technology Program under Grant 2025YFNZH0034.

References

1. Abdul Samadh, J., Gani, M.H., Hussein, N., Khattak, M.U., Naseer, M.M., Shahbaz Khan, F., Khan, S.H.: Align your prompts: Test-time prompting with distribution alignment for zero-shot generalization. Advances in Neural Information Processing Systems **36**, 80396–80413 (2023)
2. Bousselham, W., Petersen, F., Ferrari, V., Kuehne, H.: Grounding everything: Emerging localization properties in vision-language transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3828–3837 (2024)
3. Cai, W., Jin, K., Hou, J., Guo, C., Wu, L., Yang, W.: Vdd: Varied drone dataset for semantic segmentation. Journal of Visual Communication and Image Representation **109**, 104429 (2025)
4. Cao, Q., Chen, Y., Ma, C., Yang, X.: Open-vocabulary high-resolution remote sensing image semantic segmentation. IEEE Transactions on Geoscience and Remote Sensing (2025)
5. Chen, Y., Wang, Y., Lu, P., Chen, Y., Wang, G.: Large-scale structure from motion with semantic constraints of aerial images. In: Chinese Conference on Pattern Recognition and Computer Vision (PRCV). pp. 347–359. Springer (2018)
6. Cho, S., Shin, H., Hong, S., Arnab, A., Seo, P.H., Kim, S.: Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4113–4123 (2024)
7. Cordts, M., Omran, M., Ramos, S., Rehfeld, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset for semantic urban scene understanding. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3213–3223 (2016)
8. Everingham, M., Eslami, S.A., Van Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes challenge: A retrospective. International journal of computer vision **111**(1), 98–136 (2015)
9. Gao, P., Geng, S., Zhang, R., Ma, T., Fang, R., Zhang, Y., Li, H., Qiao, Y.: Clip-adapter: Better vision-language models with feature adapters. International journal of computer vision **132**(2), 581–595 (2024)
10. Ghiasi, G., Gu, X., Cui, Y., Lin, T.Y.: Scaling open-vocabulary image segmentation with image-level labels. In: European conference on computer vision. pp. 540–557. Springer (2022)
11. Ji, S., Wei, S., Lu, M.: Fully convolutional networks for multisource building extraction from an open aerial and satellite imagery data set. IEEE Transactions on geoscience and remote sensing **57**(1), 574–586 (2018)
12. Khattak, M.U., Rasheed, H., Maaz, M., Khan, S., Khan, F.S.: Maple: Multi-modal prompt learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 19113–19122 (2023)
13. Lan, M., Chen, C., Ke, Y., Wang, X., Feng, L., Zhang, W.: Clearclip: Decomposing clip representations for dense vision-language inference. In: European Conference on Computer Vision. pp. 143–160. Springer (2024)
14. Li, B., Dong, H., Zhang, D., Zhao, Z., Sun, H., Gao, J.: Exploring efficient open-vocabulary segmentation in the remote sensing. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 40, pp. 5982–5991 (2026)
15. Li, B., Weinberger, K.Q., Belongie, S., Koltun, V., Ranftl, R.: Language-driven semantic segmentation. arXiv preprint arXiv:2201.03546 (2022)

16. Li, K., Cao, X., Liu, R., Wang, S., Jiang, Z., Wang, Z., Meng, D.: Annotation-free open-vocabulary segmentation for remote-sensing images. arXiv preprint arXiv:2508.18067 (2025)
17. Li, K., Liu, R., Cao, X., Bai, X., Zhou, F., Meng, D., Wang, Z.: Segearth-ov: Towards training-free open-vocabulary segmentation for remote sensing images. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 10545–10556 (2025)
18. Liu, F., Chen, D., Guan, Z., Zhou, X., Zhu, J., Ye, Q., Fu, L., Zhou, J.: Remoteclip: A vision language foundation model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–16 (2024)
19. Lyu, Y., Vosselman, G., Xia, G.S., Yilmaz, A., Yang, M.Y.: Uavid: A semantic segmentation dataset for uav imagery. ISPRS journal of photogrammetry and remote sensing **165**, 108–119 (2020)
20. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
21. Shu, M., Nie, W., Huang, D.A., Yu, Z., Goldstein, T., Anandkumar, A., Xiao, C.: Test-time prompt tuning for zero-shot generalization in vision-language models. Advances in Neural Information Processing Systems **35**, 14274–14289 (2022)
22. Wang, D., Shelhamer, E., Liu, S., Olshausen, B., Darrell, T.: Tent: Fully test-time adaptation by entropy minimization. arXiv preprint arXiv:2006.10726 (2020)
23. Wang, J., Zheng, Z., Ma, A., Lu, X., Zhong, Y.: Loveda: A remote sensing land-cover dataset for domain adaptive semantic segmentation. arXiv preprint arXiv:2110.08733 (2021)
24. Wysoczańska, M., Siméoni, O., Ramamonjisoa, M., Bursuc, A., Trzciński, T., Pérez, P.: Clip-dinoiser: Teaching clip a few dino tricks for open-vocabulary semantic segmentation. In: European Conference on Computer Vision. pp. 320–337. Springer (2024)
25. Xia, J., Yokoya, N., Adriano, B., Broni-Bediako, C.: Openearthmap: A benchmark dataset for global high-resolution land cover mapping. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 6254–6264 (2023)
26. Xu, M., Zhang, Z., Wei, F., Hu, H., Bai, X.: Side adapter network for open-vocabulary semantic segmentation. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2945–2954 (2023)
27. Yu, Q., He, J., Deng, X., Shen, X., Chen, L.C.: Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. Advances in Neural Information Processing Systems **36**, 32215–32234 (2023)
28. Zhang, R., Zhang, W., Fang, R., Gao, P., Li, K., Dai, J., Qiao, Y., Li, H.: Tip-adapter: Training-free adaption of clip for few-shot classification. In: European conference on computer vision. pp. 493–510. Springer (2022)
29. Zhang, Z., Zhao, T., Guo, Y., Yin, J.: Rs5m and georsclip: A large-scale vision-language dataset and a large vision-language model for remote sensing. IEEE Transactions on Geoscience and Remote Sensing **62**, 1–23 (2024)
30. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Conditional prompt learning for vision-language models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 16816–16825 (2022)
31. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International journal of computer vision **130**(9), 2337–2348 (2022)