SCIENCE CHINA
Information Sciences

• LETTER •

June 2026, Vol. 69, Iss. 6, 169101:1–169101:2
https://doi.org/10.1007/s11432-025-4747-y

Reference patch momentum distillation for open-vocabulary semantic segmentation

Yajie LIU[1,2], Jinjin ZHANG[1,2], Qingjie LIU[2] & Di HUANG[1,2*]

[1] *State Key Laboratory of Complex and Critical Software Environment, Beihang University, Beijing 100191, China*
[2] *IRIPLaboratory, School of Computer Science and Engineering, Beihang University, Beijing 100191, China*

Received 25 June 2025/Revised 16 October 2025/Accepted 18 December 2025/Published online 27 January 2026

Citation Liu Y J, Zhang J J, Liu Q J, et al. Reference patch momentum distillation for open-vocabulary semantic segmentation. Sci China Inf Sci, 2026, 69(6): 169101, https://doi.org/10.1007/s11432-025-4747-y

Open-vocabulary semantic segmentation (OVSS) aims to assign each pixel in an image to a semantic label from arbitrary categories. Given the inherent challenge of associating visual elements with unbounded textual concepts, CLIP [1] has been widely adopted in OVSS for its strong open-vocabulary understanding capabilities. However, CLIP's image-level pretraining paradigm, optimized for global understanding, poses significant challenges for dense prediction that needs fine-grained spatial understanding. Current OVSS utilizes CLIP in two distinct ways. Region-based approaches [2] apply CLIP to classify mask proposals typically generated by auxiliary models (e.g., Swin Transformer) trained on limited data, resulting in high computational costs and weak generalization to novel scenes. Alternative pixel-based frameworks [3] directly fine-tune CLIP for pixel-wise prediction, incorporating cost-aggregation modules to mitigate overfitting. Despite these advances, adaptation strategies that effectively preserve CLIP's generalization capability during its transfer to dense prediction remain underexplored.

To investigate this issue, we analyze the representations of CLIP in the segmentation context. We observe that both patch features and class-name text embeddings exhibit poor category discriminability, with high intra-class variance and low inter-class separation. Moreover, some patch features are contaminated by the global category, lacking the local perceptual details essential for segmentation. These findings raise a core dilemma: CLIP's representations require adaptation for segmentation, yet naively fine-tuning risks degrading its generalization to unseen categories. Most OVSS methods address this by retaining text embeddings and fine-tuning only the visual encoder. Others attempt to distill the knowledge from CLIP. MAFT+ [2] employs multi-scale pooling to extract region-level features for distillation. However, naive pooling may aggregate contaminated patches, yielding suboptimal semantic representations.

To address this challenge, we conduct a quantitative study on the impact of adapting different components of CLIP on its generalization. Specifically, we evaluate three configurations: adapting only the visual encoder, only the text encoder, and both jointly. Given that OVSS relies on both patch features and text embeddings, we analyze the semantic shift in each modality and its correlation with generalization. To reliably assess patch-level semantic shift, we partition the patches in CLIP into two groups: (1) a noisy group comprising globally contaminated patches, and (2) a reliable group containing patches capturing accurate local semantics, termed reference patches. This analysis uncovers a key insight: maintaining the semantics of reference patches in CLIP shows a stronger correlation with open-vocabulary generalization performance than preserving text embeddings (see Appendix B.1).

Method. Motivated by this finding, we propose reference patch momentum distillation (RPMD), a novel training framework to preserve and distill the generalizable knowledge encoded in CLIP's reference patch features, as shown in Figure 1. RPMD introduces a dual-model architecture comprising (1) a target CLIP-V model adapted for OVSS, and (2) a momentum CLIP-V model, which provides stable and generalizable reference patch features to guide the target model through two key modules: selective patch distillation (SPD) and class-prototype alignment (CPA).

Momentum CLIP-V. The limited segmentation capability of the initial model constrains the number of reliable reference patches available for distillation. To address this, RPMD maintains a momentum CLIP-V E_v^M updated via exponential moving average (EMA) of the target CLIP-V E_v: $E_v^M = \alpha E_v^M + (1-\alpha) E_v$, with both initialized from the original CLIP-V. As training proceeds, the momentum CLIP-V benefits from the progressively refined dense predictions of E_v, gradually enriching the reference patch space. Given that lightly adapted models preserve stronger generalization than heavily fine-tuned ones, the early-stage reference patches derived from the E_v^M provide generalizable supervision to guide the ongoing training of the target model.

Text-to-patch pre-alignment. To fully exploit the generalization of CLIP, we enrich the initial reference patch space without altering the visual encoder. We introduce a text-to-patch pre-alignment (TPA) stage, where the visual encoder remains frozen and only the text encoder is lightly adapted through a few warm-up steps. This adaptation enhances the discriminative power of text embeddings, enabling more reliable reference patch selection from the frozen CLIP-V. The frozen visual encoder serves as a stable anchor, guiding the text encoder to align with generalizable patch semantics. The adapted text encoder then initializes the textual branch for subsequent joint training.

Selective patch distillation. The SPD module provides

[*] Corresponding author (email: dhuang@buaa.edu.cn)

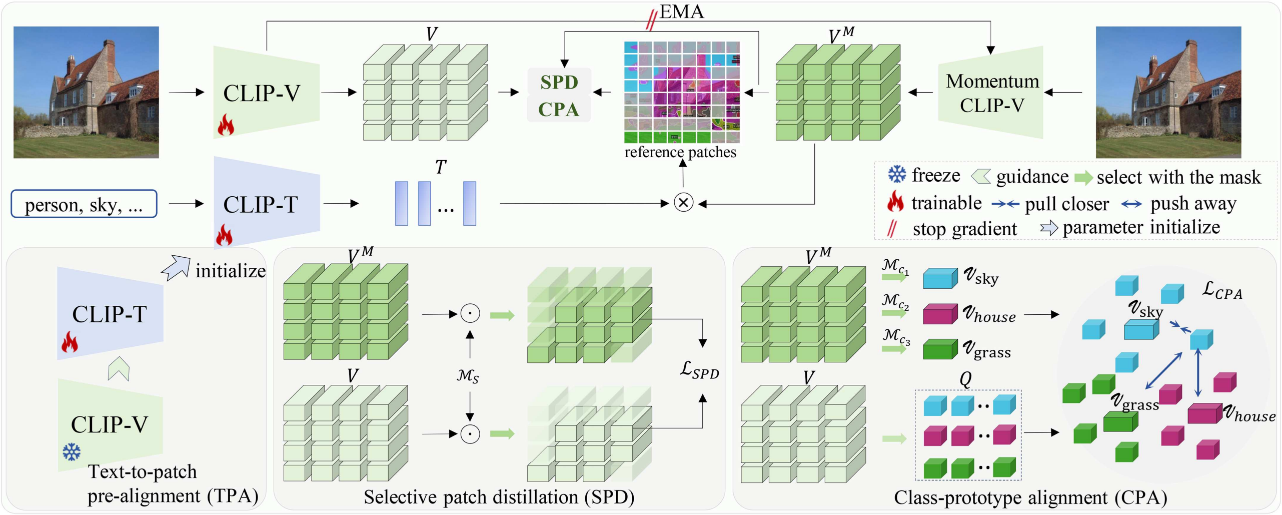

Figure 1 (Color online) Overview of the RPMD architecture. It introduces a dual-path design, where a momentum CLIP-V model constructs a generalizable reference patch space using semantically reliable patches. The TPA is employed to enrich the initial reference patch space. The reference patch semantics are distilled into the target CLIP-V model through two complementary modules: SPD for patch-level alignment and CPA for prototype-level regularization.

generalizable patch-level guidance by selectively distilling reliable patch semantics from the momentum model E_v^M to the target model E_v. Specifically, it identifies semantically accurate reference patches from E_v^M and guides E_v to align with these features at corresponding spatial locations. Let $V^M = \{V_p^M \in \mathbb{R}^D\}_{p=1}^L$ denote the patch embeddings extracted from E_v^M, where L is the number of patches and D is the feature dimension. For simplicity, we adopt a hard division strategy to derive the mask of reference patches $M_s \in \mathbb{R}^L$, where the patch is assigned 1 if it captures accurate semantics as the ground-truth label $G \in \mathbb{R}^L$, and 0 otherwise. We summarize the simple process as follows:

$$M_s = \mathbb{1}\{\mathrm{argmax}(V^M T^{\mathrm{T}}) == G\}. \tag{1}$$

The SPD loss is then computed as a masked mean squared error between the target and momentum patch features:

$$\mathcal{L}_{\mathrm{SPD}} = \frac{1}{\sum_{p=1}^L M_{s,p}} \sum\nolimits_{p=1}^L M_{s,p} \|V_p - V_p^M\|_2^2. \tag{2}$$

Class-prototype alignment. The CPA module provides prototype-level guidance by regularizing patch features to align with class prototypes derived from the reference patch space. For each category c in the image, we compute a prototype $\mathcal{V}_c$ by averaging the momentum patch embeddings associated with that class. The reference mask $M_c \in \mathbb{R}^L$ for category c is defined as

$$M_c = \mathbb{1}\left\{(\mathrm{argmax}(V^M T^{\mathrm{T}}) == G) \wedge (G == c)\right\}. \tag{3}$$

The class prototype $\mathcal{V}_c \in \mathbb{R}^D$ is then computed as

$$\mathcal{V}_c = \frac{1}{\sum_p M_{c,p}} \sum\nolimits_{p=1}^L M_{c,p} V_p^M. \tag{4}$$

For the target CLIP-V, we gather the set of correctly predicted patches Q for each category, and encourage their alignment with the corresponding class prototype $\mathcal{V}_{G_c}$, while pushing them away from other class prototypes. This is implemented via a contrastive loss:

$$\mathcal{L}_{\mathrm{CPA}} = -\frac{1}{|Q|} \sum_{V_p \in Q} \log \frac{\exp(\mathrm{sim}(V_p, \mathcal{V}_{G_c})/\tau)}{\sum_{c \in \mathcal{C}} \exp(\mathrm{sim}(V_p, \mathcal{V}_c)/\tau)}, \tag{5}$$

where $\mathrm{sim}(\cdot,\cdot)$ denotes the cosine similarity. τ is the temperature parameter and $\mathcal{C}$ is the set of class labels in the image. The $\mathcal{L}_{\mathrm{CPA}}$ not only aligns target patch features with generalizable reference semantics, but also enhances inter-class feature discriminability.

Overall objective. The model is trained by jointly optimizing the segmentation loss $\mathcal{L}_{\mathrm{SEG}}$, the patch-level SPD loss $\mathcal{L}_{\mathrm{SPD}}$ and the prototype-level CPA loss $\mathcal{L}_{\mathrm{CPA}}$. The total training loss is given by

$$\mathcal{L}_{\mathrm{total}} = \mathcal{L}_{\mathrm{SEG}} + \lambda_{\mathrm{SPD}} \cdot \mathcal{L}_{\mathrm{SPD}} + \lambda_{\mathrm{CPA}} \cdot \mathcal{L}_{\mathrm{CPA}}, \tag{6}$$

where λ_{SPD} and λ_{CPA} are balancing weights for the distillation objectives.

Conclusion. We uncover a strong correlation between preserving reference patch semantics and improved open-vocabulary generalization. Motivated by this, we propose RPMD, a dual-path framework that constructs a generalizable reference space and guides the target model via selective patch distillation and class-prototype alignment. RPMD effectively retains CLIP's generalization while boosting segmentation performance for the pixel-based OVSS.

Acknowledgements This work was partly supported by National Key Research and Development Plan (Grant No. 2024YFB3309302), National Natural Science Foundation of China (Grant No. 82441024), Beijing Natural Science Foundation (Grant No. L251073), Research Program of State Key Laboratory of Complex and Critical Software Environment, and Fundamental Research Funds for the Central Universities.

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision. In: Proceedings of ICML, 2021. 8748–8763
2 Jiao S, Zhu H, Huang J, et al. Collaborative vision-text representation optimizing for open-vocabulary segmentation. In: Proceedings of ECCV, 2024. 399–416
3 Cho S, Shin H, Hong S, et al. Cat-seg: cost aggregation for open-vocabulary semantic segmentation. In: Proceedings of CVPR, 2024. 4113–4123