Classifier-Centric Adaptive Framework for Open-Vocabulary Camouflaged Object Segmentation

Hanyu Zhang[1,*], Yiming Zhou[1,*], Jinxia Zhang[1,2,†]
[1]School of Automation, Southeast University, Nanjing, China
[2]Advanced Ocean Institute, Southeast University, Nantong, China

***Abstract*—Open-vocabulary camouflaged object segmentation requires models to segment camouflaged objects of arbitrary categories unseen during training, placing extremely high demands on generalization capabilities. Through analysis of existing methods, it is observed that the classification component significantly affects overall segmentation performance. Accordingly, a classifier-centric adaptive framework is proposed to enhance segmentation performance by improving the classification component via a lightweight text adapter with a novel layered asymmetric initialization. Through the classification enhancement, the proposed method achieves substantial improvements in segmentation metrics compared to the OVCoser baseline on the OVCamo benchmark: cIoU increases from 0.443 to 0.493, cSm from 0.579 to 0.658, and cMAE reduces from 0.336 to 0.239. These results demonstrate that targeted classification enhancement provides an effective approach for advancing camouflaged object segmentation performance.**

***Index Terms*—Open-vocabulary camouflaged object segmentation, text adapter, initialization strategy, vision-language model, parameter-efficient fine-tuning**

I. Introduction

Open-vocabulary camouflaged object segmentation (OVCOS) is a challenging task that requires segmenting camouflaged objects from unseen categories [1]. The difficulty arises from combining visual ambiguity in camouflaged scenes with the need for semantic generalization to novel object classes.

While existing methods focus on closed-vocabulary scenarios [12], [13], recent work like OVCoser [1] has shown the feasibility of open-vocabulary settings. Experimental analysis reveals that improving classification accuracy leads to substantial segmentation performance gains, indicating that enhanced semantic understanding is key to advancing OVCOS.

Based on this insight, this work proposes a classifier-centric framework that improves OVCOS through lightweight text adapters with a novel layered asymmetric initialization strategy. The approach achieves significant improvements on the OVCamo benchmark across all metrics.

The main contributions include: **(i)** demonstrating that classification enhancement directly improves OVCOS performance; **(ii)** proposing an effective layered asymmetric initialization for lightweight adapters; **(iii)** achieving state-of-the-art results on OVCamo benchmark.

*These authors contributed equally to this work. †Corresponding author.

II. Method

A. Method Overview

The proposed framework improves OVCOS performance by enhancing classification capabilities through iterative mask-guided refinement. The approach employs a two-stage inference process that leverages the synergy between segmentation and classification components.

The framework consists of three key components: (1) an enhanced Alpha-CLIP [3] classifier with lightweight text adapters, (2) a layered asymmetric initialization strategy for improved adapter performance, and (3) an iterative refinement process where predicted masks guide classification improvement.

As illustrated in Figure 1, the architecture consists of two main processing pathways. The visual pathway processes RGB input images and predicted masks through Alpha-CLIP's visual encoder to extract multimodal features. On the textual side, class-specific prompts("A photo of the camouflaged ⟨class⟩") are processed by the frozen CLIP textual encoder [2] enhanced with a lightweight text adapter at the final layer.

The text adapter employs a three-layer bottleneck architecture with projection matrices W_{down}, W_{mid}, and W_{up}, initialized using the proposed layered asymmetric initialization where $\sigma_{\text{down}} > \sigma_{\text{mid}} > \sigma_{\text{up}}$. The adapter output is scaled by a learnable factor s and added via residual connection.

The iterative refinement process generates initial mask predictions, which are fed back to provide spatial guidance for improved classification. This creates a feedback loop where better masks enable more accurate classification. The framework uses frozen CLIP backbone networks to preserve pre-trained knowledge while fine-tuning only the text adapter for efficient adaptation.

B. Lightweight Text Adapter

To achieve efficient domain-specific semantic enhancement for camouflaged object recognition, a lightweight text adapter is designed and inserted at the final layer of the text encoder. Let $\mathbf{z}$ denote the input text feature at this layer. The adapter transformation is defined as:

$$A^t(\mathbf{z}) = W_{\text{up}}^t \cdot \delta\big(W_{\text{mid}}^t \cdot \delta\big(W_{\text{down}}^t \cdot \mathbf{z}\big)\big), \tag{1}$$

where $W_{\text{down}}^t \in \mathbb{R}^{r\times d}$ projects features to a bottleneck with $r \ll d$, $W_{\text{mid}}^t \in \mathbb{R}^{r\times r}$ applies nonlinear transformation, $\delta(\cdot)$

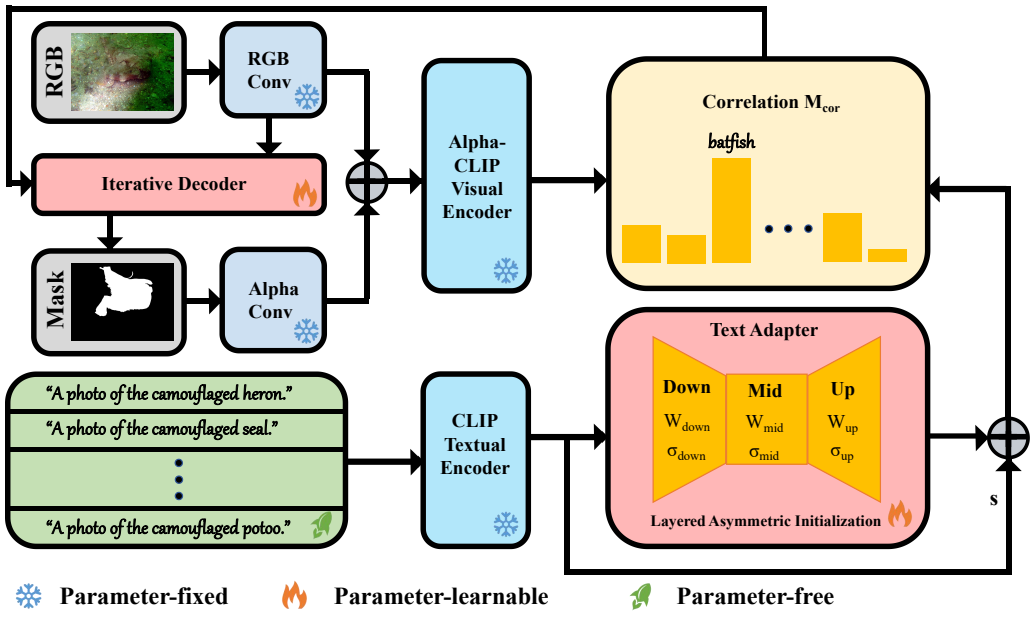

Fig. 1. Overall architecture of the proposed framework. RGB images and masks are processed through separate pathways in Alpha-CLIP. A lightweight text adapter with layered asymmetric initialization enhances the textual encoder for improved classification.

denotes ReLU activation [4], and $W_{\text{up}}^t \in \mathbb{R}^{d\times r}$ restores the original dimension.

The residual connection with learnable scaling factor s is applied as:

$$\mathbf{y} = \mathbf{z} + s \cdot A^t(\mathbf{z}). \tag{2}$$

This bottleneck design enables nonlinear refinement of camouflage-related semantics while maintaining low parameter overhead.

The system processes RGB images and masks through separate pathways in Alpha-CLIP's visual encoder. Class labels are embedded into prompt templates ("A photo of the camouflaged ⟨class⟩") and processed by the CLIP textual encoder with our adapter at the final layer. The [EOS] token is used for text-image similarity computation following standard protocols.

C. Layered Asymmetric Initialization

Through extensive empirical exploration, it is discovered that lightweight adapter architectures exhibit high sensitivity to initialization strategies due to their compact parameter space. Based on this observation, a layered asymmetric initialization (LAI) strategy is proposed that assigns different variances to different projection layers to achieve better convergence and performance. This initialization approach is designed to be applicable to similar lightweight adapter architectures beyond the current framework.

The proposed LAI strategy assigns different initialization variances to the three projections while keeping the residual gate small at the beginning of training. The adapter-induced residual update follows Eq. (2), and the projection matrices are initialized with zero-mean Gaussians under:

$$W_{\text{down}}^t \sim \mathcal{N}\big(0,\, \sigma_{\text{down}}^2\big), \tag{3}$$
$$W_{\text{up}}^t \sim \mathcal{N}\big(0,\, \sigma_{\text{up}}^2\big), \tag{4}$$
$$W_{\text{mid}} \sim \mathcal{N}\big(0,\, \sigma_{\text{mid}}^2\big), \tag{5}$$

with the ordering:

$$\sigma_{\text{down}} > \sigma_{\text{mid}} > \sigma_{\text{up}}. \tag{6}$$

The residual scale s is learnable and is initialized with a small value s_0 to ensure stable training at the beginning:

$$s_0 \in [\alpha_{\min}, \alpha_{\max}], \tag{7}$$

after which s is optimized jointly with the adapter.

This asymmetric ordering balances exploration and stability: a larger variance in W_{down} encourages diverse feature exploration in the bottleneck space, while a smaller variance in W_{up} helps preserve the pre-trained representation. The effect is amplified in lightweight adapters with a compact trainable parameter space (0.103M), where initialization can noticeably influence optimization trajectories and convergence.

The pronounced impact of initialization in our framework can be attributed to the lightweight nature of our adapter. With a compact trainable parameter space of approximately 0.103M parameters compared to the frozen CLIP backbone, each parameter carries relatively more influence on the final model behavior. This parameter efficiency amplifies the importance of proper initialization, as small variations in initial weights can lead to substantially different optimization trajectories and convergence points.

The initial perturbation magnitude is controlled by the product of all layer norms and the scaling factor s. The ordering $\sigma_{\text{down}} > \sigma_{\text{mid}} > \sigma_{\text{up}}$ ensures that the adapter starts with moderate perturbations to the pre-trained features while allowing sufficient parameter diversity for effective adaptation. The small initial value of s provides additional stability control during early training stages.

D. Test-Time Augmentation

To further enhance classification robustness, a test-time augmentation (TTA) strategy is employed. Multiple scale factors and horizontal flips are applied to generate diverse views of the input image. The final prediction aggregates results from all augmented views using confidence-weighted voting:

$$P_{\text{final}}(c|I) = \frac{\sum_{i=1}^{N} w_i \cdot \text{softmax}(\mathbf{z}_i/\tau)}{\sum_{i=1}^{N} w_i} \tag{8}$$

where $w_i = \max(\text{softmax}(\mathbf{z}_i))$ represents the confidence weight for each view and τ is the temperature parameter.

III. Experimental Results and Analysis

A. Experimental Setup

Implementation details. The framework is implemented using PyTorch on NVIDIA RTX 4090 GPUs . The method is built upon ViT-L/14@336px backbone with Alpha-CLIP [3] pre-trained weights. The text adapter employs bottleneck dimension $r = 64$ with learnable scaling factor initialized to 0.15. SGD optimizer is used with learning rate 0.0035, momentum 0.9, and weight decay 1×10^{-5}. Training is conducted for 10 epochs with batch size 16. Input images are resized to 336×336 pixels. Prompt template and class-text construction follow OVCOS [1]. The construction of mask follows Alpha-CLIP [3]. LAI initialization uses the specified variances with uniform sampling for the residual scale s within the range $[0.075, 0.225]$ for reproducibility.

Training details. The approach trains only the lightweight text adapter (0.103M trainable parameters) while keeping the CLIP backbone frozen, and finishes in about 30 minutes on a single RTX 4090. In FP16 profiling (batch size 16, 200 iterations with 50 warmup), our method has 390.303M parameters vs. 390.200M for the no-adapter baseline (+0.103M, $\sim$0.03%) and adds 0.253 GFLOPs per forward; thanks to caching text features after warmup, the steady-state runtime is nearly unchanged (0.1142 s/iter vs. 0.1143 s/iter).

B. Dataset and Evaluation Protocol

Evaluation is conducted on the OVCamo benchmark for OVCOS [1],which contains 11,483 images across 75 categories.The dataset split follows [1] with 7,713 images from 14 seen categories for training and 3,770 images from 61 unseen categories for testing.

For classification evaluation, top-1 classification accuracy is reported on correctly identified camouflaged regions. For segmentation evaluation, standard metrics including cIoU (class-aware Intersection over Union), cSm (class-aware S-measure), cMAE (class-aware Mean Absolute Error), cF_β (class-aware F-measure), $\text{cF}_{\omega\beta}$ (class-aware weighted F-measure), and cEm (class-aware E-measure) are employed to comprehensively assess segmentation performance.

TABLE I
Quantitative comparison of classification results with existing methods on OVCamo dataset. Numbers show absolute Top-1 accuracy (%) with Δ indicating gain over Alpha-CLIP baseline (in pp).

Method	Truth (GT mask)	All-black
Alpha-CLIP (2024) [3]	74.60%	69.88%
CLIP-Adapter (2021) [14]	74.06%	71.06%
CoT-LLM w/ CoVP (2024) [15]	75.40%	75.40%
MMA (2024) [16]	75.63%	73.19%
MaPLe (2023) [17]	77.34%	75.27%
Standard Adapter (Ours)	74.83%	73.51%
Standard Adapter + LAI (Ours)	78.95%	75.87%
Standard Adapter + LAI + TTA (Ours)	**79.75%**	**76.85%**

C. Classification Performance Comparison

This section evaluates the classification performance of different methods under various inference scenarios to demonstrate the effectiveness of our approach. The results are presented in Table I.

The proposed method achieves significant improvements in classification accuracy across both inference scenarios. With ground-truth masks, our complete method (LAI + TTA) reaches 79.75%, representing a 5.15 percentage point improvement over the Alpha-CLIP baseline. In the more challenging all-black scenario without spatial guidance, the improvement is even more substantial at 6.97 percentage points (from 69.88% to 76.85%).

The LAI adapter alone provides the majority of performance gains, demonstrating its effectiveness in enhancing semantic understanding for camouflaged objects. TTA contributes additional consistent improvements of approximately 0.8-1.0 percentage points across both scenarios.

D. Segmentation Performance Comparison

This section compares segmentation performance across different methods to validate that improved classification translates to better segmentation results. Table II shows the comparison, and Table III demonstrates the impact of classification enhancement on segmentation metrics.

The proposed method achieves substantial improvements across all segmentation metrics compared to existing approaches. The cIoU increases from 0.443 to 0.493, cSm from 0.579 to 0.658, and cMAE reduces from 0.336 to 0.239 when compared to the OVCoser baseline. These improvements demonstrate that enhanced classification capabilities directly translate to better segmentation performance.

Compared to general open-vocabulary segmentation methods, the specialized approach shows significant advantages in handling camouflaged objects, highlighting the importance of domain-specific adaptations for this challenging task.

TABLE II
QUANTITATIVE COMPARISON OF SEGMENTATION RESULTS WITH EXISTING METHODS ON OVCAMO DATASET. METRICS: CSM↑/CF$\omega\beta$↑/CMAE↓/CFβ↑/CEM↑/CIOU↑.

Method	cSm	cF$\omega\beta$	cMAE	cFβ	cEm	cIoU
SimSeg'21 [5]	0.053	0.049	0.921	0.056	0.098	0.047
OVSeg'22 [6]	0.024	0.046	0.954	0.056	0.130	0.046
ODISE'23 [9]	0.187	0.119	0.700	0.211	0.298	0.167
SAN'23 [8]	0.275	0.202	0.612	0.220	0.318	0.189
CAT-Seg'23 [10]	0.181	0.106	0.719	0.123	0.196	0.094
FC-CLIP'23 [11]	0.080	0.076	0.872	0.090	0.191	0.072
OVCoser'24 [1]	0.579	0.490	0.336	0.520	0.616	0.443
Ours	**0.658**	**0.547**	**0.239**	**0.582**	**0.696**	**0.493**

TABLE III
ABLATION STUDY ON THE EFFECT OF CLASSIFICATION ENHANCEMENT ON SEGMENTATION PERFORMANCE.

Setting	cSm	cF$\omega\beta$	cMAE	cFβ	cEm	cIoU
CLIP Classifier	0.584	0.513	0.336	0.538	0.616	0.447
Ours (LAI + TTA)	**0.658**	**0.547**	**0.239**	**0.582**	**0.696**	**0.493**
Oracle Classifier	**0.800**	**0.628**	**0.036**	**0.672**	**0.856**	**0.570**

E. Ablation Study

1) Impact of Classification Enhancement on Segmentation Performance: The classification experiment in Table III provides strong evidence for the proposed approach. When perfect classification is achieved, segmentation performance improves dramatically across all metrics, with cMAE dropping from 0.336 to 0.036 and notable gains in other measures. This validates the hypothesis that improving classification accuracy leads to meaningful overall performance improvements in open-vocabulary camouflaged object segmentation.

2) Effectiveness of Different Components: This section analyzes the contribution of each component in the framework through systematic ablation experiments. The results are shown in Table IV.

The ablation results demonstrate the effectiveness of each proposed component. The layered asymmetric initialization (LAI) provides the most significant contribution, achieving 4.35 percentage point improvement in the Truth setting and 5.99 points in the All-black setting compared to the Alpha-CLIP baseline.

The superiority of LAI over standard adapter initialization is evident: LAI achieves 78.95% vs 74.83% (+4.12pp) in the Truth setting and 75.87% vs 73.51% (+2.36pp) in the All-black setting. This substantial performance difference highlights the importance of proper initialization strategies for lightweight adapters, where each parameter carries significant influence due to the compact parameter space.

Test-time augmentation consistently provides additional improvements of approximately 0.8-1.0 percentage points across both scenarios, demonstrating its effectiveness as a complementary enhancement technique.

TABLE IV
ABLATION STUDY ON OVCAMO. NUMBERS SHOW ABSOLUTE TOP-1 ACCURACY (%) WITH Δ INDICATING GAIN OVER ALPHA-CLIP BASELINE (IN PP).

Components	Truth (GT mask)	All-black
Alpha-CLIP (baseline) [3]	74.60%	69.88%
+ Standard Adapter	74.83% (+0.23)	73.51% (+3.63)
+ LAI	78.95% (+4.35)	75.87% (+5.99)
+ TTA	**79.75%** (+**5.15**)	**76.85%** (+**6.97**)

The segmentation-guided inference approach (using predicted masks from segmentation models) achieves 79% accuracy, representing a practical middle ground between the all-black scenario (76.85%) and the ground-truth scenario (79.75%). This demonstrates the value of iterative mask-guided refinement in real-world applications where ground-truth masks are unavailable.

IV. CONCLUSION

It is discovered that improving classification accuracy can lead to meaningful improvements in open-vocabulary camouflaged object segmentation through a classifier-centric adaptive framework. The proposed lightweight text adapters with empirically effective layered asymmetric initialization and iterative mask-guided refinement enable efficient semantic enhancement while preserving pre-trained knowledge. On the OVCamo benchmark, substantial improvements in segmentation metrics are achieved compared to existing methods: cIoU increases from 0.443 to 0.493, cSm from 0.579 to 0.658, and cMAE reduces from 0.336 to 0.239, demonstrating that targeted classification enhancement provides an effective approach for advancing OVCOS capabilities. The significant impact of the initialization strategy underscores the importance of proper parameter initialization in lightweight adapter architectures; nevertheless, rare classes with extreme camouflage remain challenging, and although our adapter partially alleviates class bias, generalizing to extreme long-tail categories (e.g., mongoose) is still difficult.

While the LAI strategy shows consistent empirical benefits, future work could explore the theoretical foundations of these initialization strategies and their connection to adapter optimization dynamics in vision-language models. Additionally, investigating the integration of this framework with other advanced segmentation architectures and extending the approach to other challenging vision tasks could further validate its generalizability.

REFERENCES

[1] Y. Pang *et al.*, "Open-vocabulary camouflaged object segmentation," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2024.

[2] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.

[3] Z. Sun *et al.*, "Alpha-CLIP: A CLIP model focusing on wherever you want," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.

[4] X. He and K. He, "Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2015, pp. 1026–1034.
[5] S. Ghiasi *et al.*, "Scaling open-vocabulary image segmentation with image-level labels," in *Proc. Eur. Conf. Computer Vision (ECCV)*, 2022, pp. 540–557.
[6] J. Ding *et al.*, "Decoupling zero-shot semantic segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 11593–11602.
[7] X. Gu *et al.*, "Open-vocabulary semantic segmentation with mask-adapted CLIP," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 7061–7070.
[8] M. Xu *et al.*, "Side adapter network for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 2945–2954.
[9] J. Yu *et al.*, "ODISE: Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 950–960.
[10] S. Cho *et al.*, "CAT-Seg: Cost aggregation for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023, pp. 13991–14001.
[11] D. Xian *et al.*, "FC-CLIP: Frozen convolutional CLIP for open-vocabulary semantic segmentation," in *Proc. Adv. Neural Inf. Process. Syst. (NeurIPS)*, 2023, pp. 28742–28754.
[12] D.-P. Fan *et al.*, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 2777–2787.
[13] H. Mei *et al.*, "Camouflaged object segmentation with distraction mining," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2021, pp. 8772–8781.
[14] P. Gao *et al.*, "CLIP-Adapter: Better vision-language models with feature adapters," *arXiv preprint arXiv:2110.04544*, 2021.
[15] L. Tang *et al.*, "Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection," in *Proc. 32nd ACM Int. Conf. Multimedia*, 2024, pp. 6887–6896.
[16] L. Yang, R.-Y. Zhang, Y. Wang, and X. Xie, "MMA: Multi-modal adapter for vision-language models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024.
[17] M. U. Khattak *et al.*, "MaPLe: Multi-modal prompt learning," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2023.

ACKNOWLEDGMENT

This work was supported by the National Natural Science Foundation of China (Grant No. 62573123), the Research Fund for Advanced Ocean Institute of Southeast University, Nantong (GP202411), the Fundamental Research Funds for the Central Universities, and the Undergraduate Training Programs for Innovation of Jiangsu Province. We also thank the Big Data Computing Center of Southeast University for providing facility support for the numerical calculations in this paper.