Computational Visual Media
https://doi.org/10.26599/CVM.2025.9450512

Research Article

Open-vocabulary camouflaged object segmentation with cascaded vision language models

Kai Zhao[1], Wubang Yuan[1], Zheng Wang[1], Guanyi Li[1], Xiaoqiang Zhu[1] (✉), Deng-Ping Fan[2], and Dan Zeng[1]

Abstract Open-vocabulary camouflaged object segmentation (OVCOS) seeks to segment and classify camouflaged objects in arbitrary categories, presenting unique challenges due to visual ambiguity and unseen categories. Recent approaches typically adopt a two-stage paradigm: they first segment objects, and then classify the segmented regions using vision language models (VLMs). However, such methods (i) suffer from a domain gap caused by the mismatch between VLMs' full-image training and cropped-region inferencing, and (ii) depend on generic segmentation models optimized for well-delineated objects which are less effective for camouflaged objects. Without explicit guidance, generic segmentation models often overlook subtle boundaries, leading to imprecise segmentation. In this paper, we introduce a novel VLM-guided cascaded framework to address these issues in OVCOS. For segmentation, we leverage the segment anything model (SAM), guided by the VLM. Our framework uses VLM-derived features as explicit prompts to SAM, effectively directing attention to camouflaged regions and significantly improving localization accuracy. For classification, we avoid the domain gap introduced by *hard* cropping. Instead, we treat the segmentation output as a *soft* spatial prior using the alpha channel. This retains the full image context while providing precise spatial guidance, leading to more accurate and context-aware classification of camouflaged objects. The same VLM is shared between segmentation and classification to ensure efficiency and semantic consistency. Extensive experiments on both OVCOS and conventional camouflaged object segmentation benchmarks demonstrate the clear superiority of our method, highlighting the effectiveness of leveraging rich VLM semantics for both segmentation and classification of camouflaged objects. Our code and models are open-sourced at `https://github.com/intcomp/camouflaged-vlm`.

Keywords open-vocabulary segmentation; camouflaged objects; vision-language models (VLMs); CLIP; segment anything model (SAM)

1 Introduction

Open-vocabulary camouflaged object segmentation (OVCOS) is a challenging task that requires segmenting and classifying camouflaged objects which may belong to novel categories not seen during training [1]. Compared to traditional semantic segmentation [2–4], OVCOS faces additional challenges because it requires recognizing novel categories in visually ambiguous scenes, where camouflage leads to low contrast, indistinct boundaries, and high similarity between objects and their backgrounds. These challenges are particularly relevant in real-world applications such as medical image analysis [5] and agricultural monitoring [6].

Several existing open-vocabulary segmentation approaches [7–10] utilize vision-language models (VLMs), e.g., CLIP [11], to directly classify each pixel across the entire input image, thereby improving semantic generalization. Such approaches use a one-stage framework. However, VLMs are pre-trained

1 School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China. E-mail: K. Zhao, kz@kaizhao.net; W. Yuan, yuanwubang@shu.edu.cn; Z. Wang, zhengwang@shu.edu.cn; G. Li, kwunyatlee@shu.edu.cn; X. Zhu, xqzhu@shu.edu.cn (✉); D. Zeng, dzeng@shu.edu.cn.

2 Nankai International Advanced Research Institute (Shenzhen Futian), Shenzhen 518045, China, and VCIP, and CS Department, Nankai University, Tianjin 300071, China. E-mail: fdp@nankai.edu.cn.

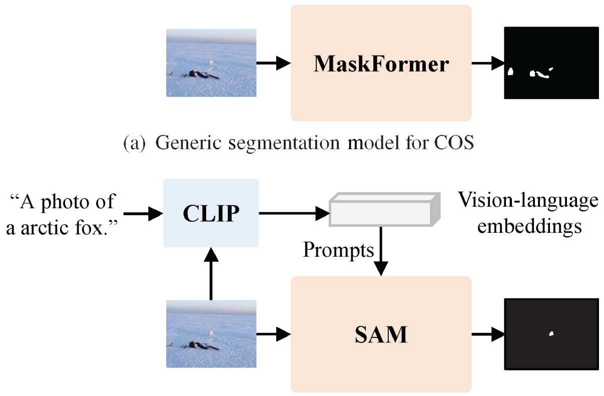
Manuscript received: 2025-06-24; accepted: 2025-09-09

清華大学出版社 Tsinghua University Press
Available on IEEE Xplore®

for image-level understanding, creating a granularity mismatch that hinders effective visual-semantic alignment and limits semantic transfer, often leading to suboptimal performance [12].

To bridge this gap, recent works [1, 12–15] first perform class-agnostic segmentation and then classify the segmented regions using VLMs. This pipeline forms a two-stage framework. The decoupling of segmentation and classification partially alleviates the granularity mismatch [12]. However, in the segmentation stage, (see Fig. 1(a)), many existing approaches typically rely on generic segmentation architectures [2, 12–15] to identify the target region. These generic segmentation models are primarily tailored for well-delineated objects and can fail to generalize effectively to camouflaged scenarios, where targets are subtle, indistinct, and visually embedded in complex backgrounds. The lack of alignment between the pretraining objectives and the demands of camouflaged segmentation leads to imprecise localization. In addition, most existing methods do not incorporate explicit edge-awareness mechanisms, which are crucial for accurately delineating objects with weak or ambiguous boundaries.

Recent advanced foundation models such as the segment anything model (SAM) [16] have shown remarkable ability to generalize to various segmentation tasks [17, 18], largely due to their ability to perform prompt-guided segmentation. By using prompts to specify target regions, SAM can adapt its attention to user-defined areas, making it particularly effective for specialized tasks such as camouflaged object segmentation. To address the limitations of generic segmentation architectures in handling camouflaged objects with weak or ambiguous boundaries, we propose an adapted SAM architecture tailored for camouflaged object segmentation. See Fig. 1(b). We integrate CLIP-derived visual and textual embeddings as prompts into the SAM mask decoder, providing task-specific semantic guidance that enhances the ability of the model to focus on the camouflaged targets. Additionally, we enhance the mask decoder with conditional multi-way attention and an edge-aware refinement module to improve boundary precision, effectively handling the indistinct contours characteristic of camouflage.



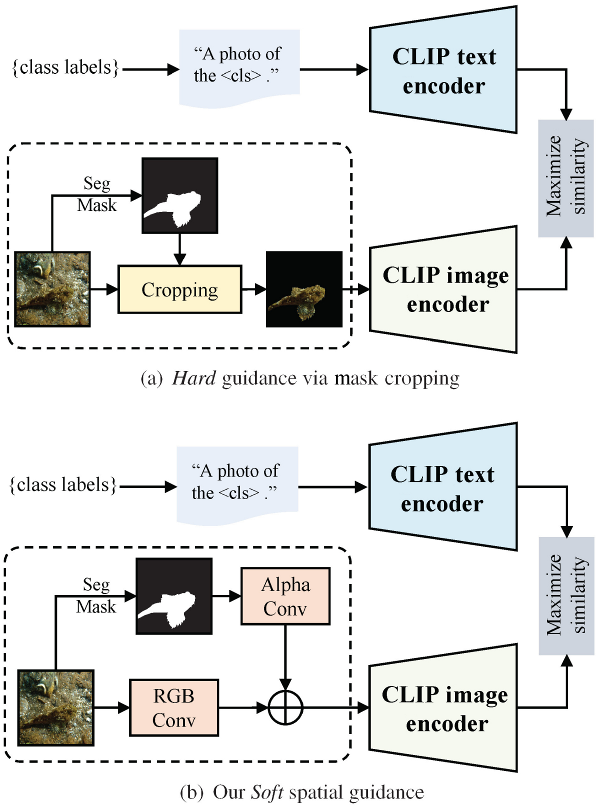
(a) Generic segmentation model for COS

(b) Visual-language prompted segmentation for COS

Fig. 1 Different segmentation paradigms in two-stage OVCOS. (a) Generic segmentation models, such as MaskFormer, typically operate directly on the input image without target-specific guidance, and are primarily designed to segment salient foreground objects. (b) Our segmentation model leverages vision-language embeddings from CLIP as prompts to guide the SAM model, directing attention to the camouflaged area.

In the classification stage, most existing methods crop the segmented regions for classification [1, 12, 13] (see Fig. 2(a)), introducing a domain gap since CLIP is pre-trained on full images. To mitigate the domain gap, we adopt a region-aware classification strategy that replaces hard cropping with a soft spatial prior derived from the segmentation mask, applied via the image's alpha channel. Our approach preserves the full image context while providing explicit spatial guidance. The predicted segmentation mask serves as a soft spatial prior and is fused with the input image via a lightweight integration module before being processed by the CLIP [11] image encoder. A comparison between *hard* and *soft* spatial guidance is presented in Fig. 2. Additionally, we fine-tune CLIP using a multi-modal prompting strategy inspired by Ref. [19], jointly optimizing both visual and textual prompts. This enhances semantic alignment and task-specific adaptability, enabling region-aware classification without disrupting global semantics.

Building on these ingredients, we introduce the cascaded open-vocabulary camouflaged understanding network (COCUS), a novel two-stage framework for the OVCOS task that explicitly decouples the process into *segmentation* and *classification*. In the first (segmentation) stage, we use CLIP [11] to extract visual and textual features. These features serve as prompts to the SAM [16] for segmentation. This prompt-based guidance allows SAM to focus more precisely on camouflaged target regions, enhancing localization in visually ambiguous scenes. In the second stage



(a) *Hard* guidance via mask cropping

(b) Our *Soft* spatial guidance

Fig. 2 Comparison of mask-guided classification strategies. (a) The mask cropping strategy applies the segmentation mask to crop the input image before feeding it into the CLIP image encoder. (b) Our method fuses the segmentation mask with the original image for region-aware classification while retaining full-image context.

(classification), the segmentation output serves as spatial guidance to refine the integration with the original image, allowing CLIP to perform open-vocabulary classification with improved focus on target regions. By disentangling segmentation and classification, our method enables more accurate semantic interpretation of camouflaged objects through prompt-based guidance segmentation and region-aware classification.

Extensive experiments on the OVCamo [1] benchmark demonstrate the effectiveness of the proposed framework for the OVCOS task. Compared to the strong baseline OVCoser [1], we achieve consistent improvements across all major evaluation metrics, establishing a new state-of-the-art on this challenging benchmark. Moreover, adapted SAM [16] demonstrates strong performance on the conventional COS task, confirming that CLIP-prompting and edge-aware segmentation are effective in standard closed-set scenarios. Besides, our method also shows strong cross-domain generalization, achieving competitive results on medical and agricultural datasets, further highlighting its robustness and practical applicability. In summary, the main contributions of this work are as follows:

- We propose a novel two-stage framework for OVCOS that explicitly decouples segmentation and classification. Our approach employs a prompt-guided segmentation model to generate a mask, which serves as *soft* spatial guidance for the classification stage while preserving full-image context.
- We propose an adapted SAM as the segmentation model, enhanced for camouflaged object localization by injecting CLIP-derived textual and visual embeddings as prompts. This design provides rich semantic guidance that steers attention toward visually ambiguous regions. Furthermore, we improve the SAM's mask decoder with conditional multi-way attention and edge-aware refinement, improving both spatial accuracy and boundary delineation.
- Extensive experiments on the OVCamo benchmark demonstrate that our method achieves state-of-the-art performance. Moreover, the adapted SAM exhibits strong generalization on the conventional COS task, validating the effectiveness of our framework in both open- and closed-set camouflaged segmentation scenarios.

The remainder of this paper is organized as follows. Section 2 reviews recent advances in open-vocabulary segmentation and camouflaged object understanding. Section 3 presents the proposed framework, detailing its cascaded design, CLIP fine-tuning pipeline and adapted SAM segmentation model. Section 4 presents implementation details, including training settings and architectural configurations, followed by comprehensive experimental results and ablation studies.

2 Related work

2.1 Vision-language models

Vision-language models (VLMs) are neural architectures that learn joint visual–textual representations by embedding both image and text inputs in a shared semantic space. A seminal model in this domain, CLIP [11], jointly learns image and text representations via contrastive learning on large-scale web data, demonstrating strong generalization across open-vocabulary tasks such as object detection [20–23] and segmentation [7, 12–15, 24–27]. However, basic CLIP often performs poorly in downstream tasks without task-specific adaptation. To address this limitation, researchers have proposed a variety of fine-tuning approaches. Alpha-CLIP [28] introduces spatially adaptive attention to improve focus on semantically relevant image regions. CoOp [29] and Co-CoOp [30] optimize textual prompts for better few-shot performance and generalization, respectively. Visual prompt tuning [31] further enhances adaptability by injecting fine-grained prompts into the vision branch. To overcome the limitations of single-modality tuning, recent works [19, 32, 33] have adopted multi-modal strategies. FGVP [33] learns patch-level visual prompts to improve alignment across diverse tasks. MaPLe [19] jointly tunes prompts in both visual and textual encoders, preserving CLIP's generality while enabling task-specific adaptation. In this work, we adopt a multi-modal prompt tuning framework similar to MaPLe to fine-tune CLIP, enhancing semantic alignment for OVCOS.

2.2 Camouflaged object segmentation

Camouflaged object segmentation has emerged as a significant research focus in computer vision, with the aim of segmenting objects that visually blend into their surroundings. Unlike traditional tasks such as salient object detection [34–38] and semantic segmentation [39, 40], COS is inherently more challenging due to low object-background contrast, ambiguous boundaries, and high background similarity. It holds practical value in domains such as medical image analysis [5] and agricultural monitoring [6]. COS is typically formulated as a class-agnostic task, focusing on segmenting camouflaged regions within complex visual scenes. Existing studies [5, 41–46] have demonstrated strong performance on established datasets [5, 42, 47]. Recent advances have introduced several SAM-based methods [16, 48, 49] adapted for COS, which use prompt tuning and architectural modifications to improve segmentation performance in complex scenes.

2.3 Open-vocabulary camouflaged object segmentation

Open-vocabulary camouflaged object segmentation is a specialized subtask of open-vocabulary segmentation, to segment and recognize camouflaged objects belonging to arbitrary textual categories. Open-vocabulary segmentation aims to align visual and textual representations in a shared embedding space, enabling pixel-level segmentation for unseen or novel categories. Early methods [50] used semantic hierarchies and concept graphs to bridge word concepts and semantic relations. With the rise of VLMs like CLIP [11], recent open-vocabulary segmentation methods have shifted toward leveraging pretrained VLMs to directly connect visual regions with text queries. These approaches follow one-stage and two-stage paradigms. One-stage methods such as MaskCLIP [14] adapt CLIP for segmentation without additional training. SAN [24] enhances feature representations via adapters. CAT-Seg [7] introduces cost aggregation between image and text embeddings. FC-CLIP [25] employs hierarchical feature fusion. However, these methods often suffer from suboptimal alignment due to CLIP's image-level representations. The two-stage methods address this by decoupling segmentation and classification. For example, SimSeg [12] uses a cascaded design with MaskFormer [2] for class-agnostic mask generation and CLIP for classification. OVSeg [15] fine-tunes CLIP on diverse and noisy data to improve generalization. In Ref. [51], a text-to-image diffusion model is employed for mask generation. While these two-stage framework methods work well on generic objects, they fall short in camouflaged scenarios. OVCOS is especially difficult because low contrast visuals, ambiguous edges, and visually similar backgrounds all contribute to degraded segmentation and classification results. OVCoser [1] was the first to address this task by combining a dedicated camouflaged segmentation model with a CLIP-based classifier in a two-stage pipeline. However, it relies on cropped inputs for classification and does not fully exploit VLM semantics in segmentation.

3 Method

3.1 Problem definition

Open-vocabulary camouflaged object segmentation aims to *segment and classify* camouflaged objects belonging to novel categories unseen during training. Formally, let $\mathcal{C}_{\text{seen}}$ denote the set of categories available during training, and $\mathcal{C}_{\text{unseen}}$ represent the disjoint set of target categories at inferencing time, such that $\mathcal{C}_{\text{seen}} \cap \mathcal{C}_{\text{unseen}} = \emptyset$. Given an input image I and novel class labels $\mathcal{C}_{\text{unseen}}$, the model is required to produce a segmentation mask M highlighting the camouflaged object and predict its corresponding class label $\hat{y} \in \mathcal{C}_{\text{unseen}}$.

To address this task, we adopt a *segment-and-classify* strategy. In the first stage, a class-agnostic segmentation model localizes camouflaged regions guided by visual and textual semantics. In the second stage, a vision–language model performs open-vocabulary classification by comparing the visual representation of the segmented regions with textual embeddings of the novel class labels, supporting recognition in an open-set setting.

3.2 Overview

Fig. 3 illustrates the proposed two-stage framework for OVCOS. During inferencing, the first stage generates a class-agnostic camouflaged segmentation mask, while the second stage performs open-vocabulary classification based on the segmented regions. We use the same CLIP model for both stages. Our CLIP model accepts a triplet $\{I \in \mathbb{R}^{H\times W\times 3}$, M, text$\}$ as input, where I and M are image and mask, and text is a description of the input, in the format "a photo of <*something*>". The CLIP model outputs visual and textual embeddings, E_v and E_t, which serve as prompts to guide segmentation in the first stage and are used for similarity-based open-vocabulary classification in the second stage. Notably, to ensure a consistent input format across stages, we use an all-one mask in the first stage, while in the second stage, the predicted segmentation mask is used as input.

In the first stage, as shown in Fig. 3(left), we perform segmentation guided by textual and visual embeddings. The inputs consist of an RGB image $I \in \mathbb{R}^{H\times W\times 3}$ and a set of class labels $\mathcal{C} = \{c_1, \ldots, c_N\}$, where N denotes the number of candidate classes. They are processed by the CLIP [11] model to produce a textual embedding E_t and a visual embedding E_v optimized for camouflaged object understanding. These embeddings serve as prompts and, together with the image I, are input into the adapted SAM model to guide the prediction of a class-agnostic camouflaged segmentation mask $M \in [0,1]^{H\times W\times 1}$, effectively localizing the camouflaged object.

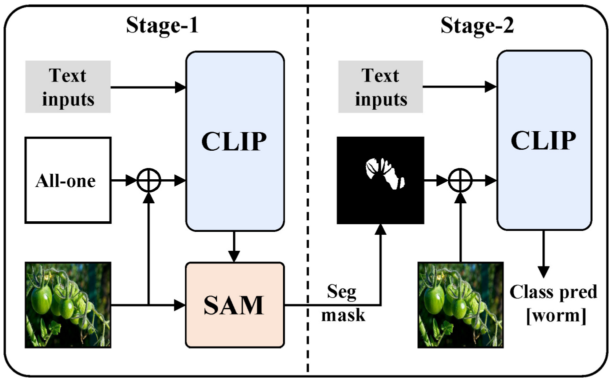


Fig. 3 Overview of the cascaded *segment and classify* framework. In Stage 1, the adapted SAM model generates a class-agnostic camouflaged segmentation mask using textual and visual embeddings as prompts. In Stage 2, we use the generated segmentation mask to enable region-aware open-vocabulary classification.

In the second stage (see Fig. 3(right)), we perform open-vocabulary classification guided by the segmented result. The inputs include the same RGB image I and class labels $\mathcal{C}_{\text{unseen}}$ from the first stage, while the predicted segmentation mask M which is used as an additional input to the CLIP model, providing spatial guidance. These inputs are processed by the CLIP model as stage one, which now focuses more precisely on the localized object area. The model then outputs a predicted class label $\hat{y} \in \mathcal{C}_{\text{unseen}}$, identifying the category of the camouflaged object. Let $E_t^N \in \mathbb{R}^{N\times d}$, and $E_v \in \mathbb{R}^{1\times d}$ be the textual and visual embeddings, where $d = 768$ is the feature dimension. We first calculate the similarity scores $S \in \mathbb{R}^N$:

$$S = E_t^N \cdot (E_v)^{\mathrm{T}} \tag{1}$$

During training, we first fine-tune our CLIP [11] model by optimizing learnable prompts in both the language and vision branches to enhance its sensitivity to camouflaged objects, with all encoder parameters frozen. Figure 4 illustrates the fine-tuning pipeline of our CLIP. After fine-tuning, we freeze the CLIP model as a feature extractor and train the SAM [16] using visual–textual features from CLIP as prompts. The details of the CLIP fine-tuning process are provided in Section 3.3, and the architecture of the SAM is described in Section 3.4.

3.3 CLIP fine-tuning pipeline

We fine-tune the CLIP model using a multi-modal prompting strategy to enhance its ability to capture subtle semantic cues for camouflaged object segmentation, as shown in Fig. 4. Our CLIP variant is a modified version of Alpha-CLIP [28]. Previous prompting strategies in CLIP [11] typically operate on the visual or textual modality. Language-only prompt tuning methods [12, 29, 30] optimize learnable prompts solely in the language branch, while visual-only approaches [14, 15, 33] inject prompts exclusively into the vision branch. In this work, we adopt a multi-modal prompting strategy, following Ref. [19], which jointly optimizes both textual and visual prompts to enhance multi-modal alignment and better adapt to task-specific objectives.

In particular, we append learnable textual prompts P_t to the language branch and generate the corresponding visual prompts P_v, which are produced by conditioning on the textual prompts through an MLP injector. The language and textual prompts P_t and P_v are shown in Fig. 4(center). During fine-tuning, only the textual prompts and injector parameters are updated, while the rest of the CLIP model remains frozen. This lightweight strategy promotes efficient adaptation and enables improved semantic alignment across modalities.

Next, we outline the fine-tuning pipeline of the CLIP model. The fine-tuning pipeline begins with the language branch, where the base class labels $\mathcal{C}_{\text{seen}}$ are formatted using the prompt template “A photo of the <*class*> camouflaged in the background.” and enriched with learnable textual prompts P_t. These are processed by the frozen CLIP text encoder to produce textual embeddings $E_t^N \in \mathbb{R}^{N\times 768}$, where N is the number of base classes.

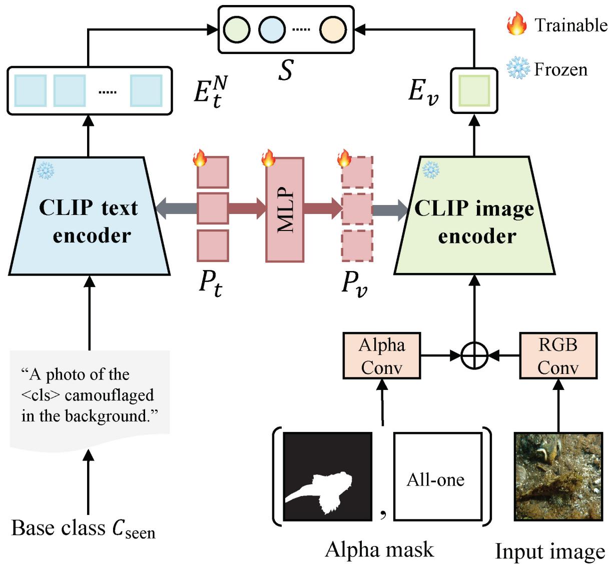


Fig. 4 CLIP fine-tuning pipeline. The language branch encodes base class labels C_{seen} with a camouflage-specific prompt template and learnable textual prompts P_t to obtain textual embeddings E_t^N. The vision branch fuses features from the input image and alpha mask, combined with visual prompts P_v injected via an MLP, and passes them to the frozen CLIP image encoder to obtain the visual embedding E_v. Similarity scores S are computed by aligning E_t^N and E_v in a shared space.

Currently, in the vision branch, the input RGB image $I \in \mathbb{R}^{H\times W\times 3}$ is augmented with an auxiliary alpha mask $A \in \mathbb{R}^{H\times W\times 1}$. The alpha mask A is randomly selected as either the all-one mask A_J or the ground-truth segmentation mask A_{gt}, each with equal probability. This enables the CLIP model to optionally accept a mask as input, defaulting to an all-one matrix when no explicit mask is provided—for example, during the first stage of segmentation.

The image I and the alpha mask A are separately processed through dedicated convolutional layers, e.g., AlphaConv and RGBConv in Fig. 4, to extract modality-specific features, which are then fused to form the visual representation. This fused representation, along with the injected visual prompts P_v generated by a lightweight MLP-based injector, is fed into the frozen CLIP image encoder to obtain the visual embedding $E_v \in \mathbb{R}^{1\times 768}$.

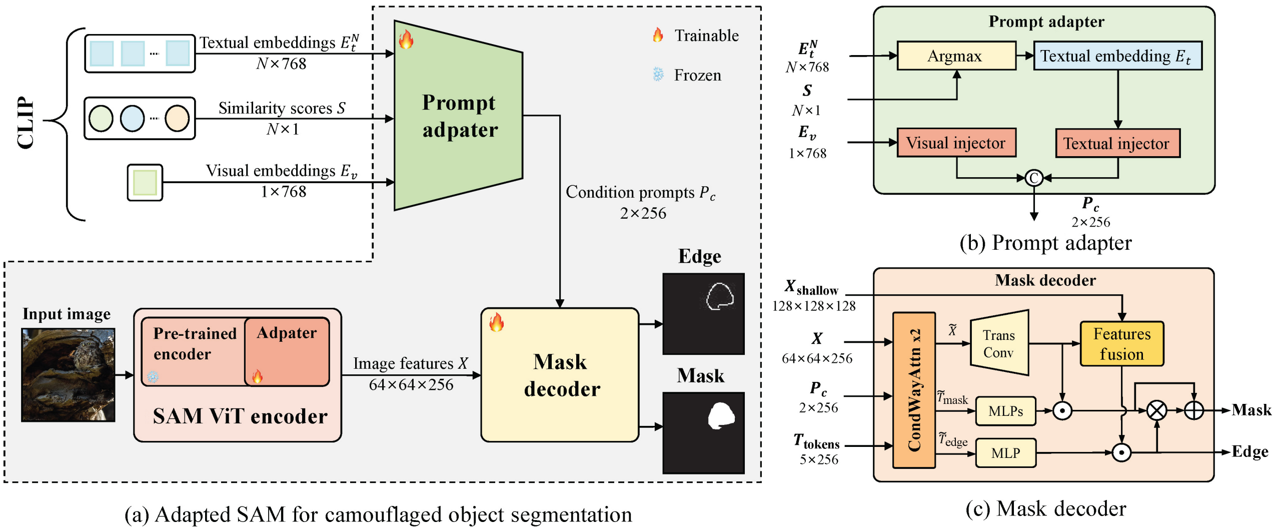
Finally, the textual and visual embeddings are used to compute the similarity score as defined in Eq. (1), which is used to calculate a cross-entropy loss against the ground-truth class labels.

3.4 Adapted SAM

We build upon SAM [16] to address the unique challenges of COS. While SAM excels at general-purpose segmentation, it struggles with the subtle visual cues and semantic ambiguities inherent to camouflaged objects. To overcome these limitations (see Fig. 5(a)), we adapt SAM by incorporating textual and visual embedding guidance and edge-aware enhancements for improved segmentation.

Specifically, we integrate our fine-tuned CLIP model with SAM to provide semantic context. The CLIP model produces textual embeddings $E_t^N \in \mathbb{R}^{N\times 768}$, a visual embedding $E_v \in \mathbb{R}^{1\times 768}$, and similarity scores $S \in \mathbb{R}^{N\times 1}$. These embeddings are further processed by a prompt adapter, which projects them into condition prompts $P_c \in \mathbb{R}^{2\times 256}$, providing high-level semantic guidance in the segmentation pipeline.

In parallel, the SAM ViT encoder extracts image features $X \in \mathbb{R}^{64\times 64\times 256}$ from the input image. To adapt SAM to camouflage-specific cues, we introduce



(a) Adapted SAM for camouflaged object segmentation

(b) Prompt adapter

(c) Mask decoder

Fig. 5 Overview of the adapted SAM framework. (a) *Adapted SAM for COS:* Our fine-tuned CLIP provides textual embeddings E_t^N, a visual embedding E_v, and similarity scores S, which are projected into condition prompts P_c via a prompt adapter. Image features X extracted by an SAM ViT encoder are refined by adapters. The mask decoder integrates X and P_c to predict the segmentation mask M and edge map E, enabling precise localization. (b) The *prompt adapter* selects the most relevant textual embedding based on S, and projects both E_t and E_v into a unified condition space via lightweight MLPs to guide the decoder. (c) The *adapted mask decoder* combines image features X, condition prompts P_c, and output tokens T_{tokens} to produce accurate masks and edge maps, improving segmentation of camouflaged scenes.

lightweight adapter modules that refine the image features X while keeping the backbone frozen.

Finally, the refined image features X and condition prompts P_c are fused within a mask decoder, which outputs a segmentation mask $M \in \mathbb{R}^{H\times W\times 1}$ and an edge map $E \in \mathbb{R}^{H\times W\times 1}$. The integration of refined image features and condition prompts within the decoder ensures accurate object localization and precise boundary delineation.

The *prompt adapter* refines textual and visual embeddings from our fine-tuned CLIP to generate condition prompts for segmentation guidance, as shown in Fig. 5(b). Given textual embeddings $E_t^N = \{e_t^1, \ldots, e_t^N\}$, visual embedding E_v, and similarity scores $S = \{s_1, \ldots, s_N\}$, the adapter selects the textual embedding corresponding to the highest similarity score:

$$i^* = \arg\max_i s_i, \quad E_t = e_t^{i^*} \tag{2}$$

The selected textual embedding E_t and visual embedding E_v are projected into a shared 256-dimensional condition space using lightweight MLP-based injectors. The resulting condition prompts $P_c \in \mathbb{R}^{2\times256}$ provide high-level semantic and visual guidance to the segmentation mask decoder, enhancing object localization and boundary accuracy. Formally, this is written:

$$P_t = \text{MLP}_{\text{text}}(E_t), \quad P_v = \text{MLP}_{\text{vis}}(E_v) \tag{3}$$

$$P_c = [P_t, P_v] \in \mathbb{R}^{2\times256} \tag{4}$$

where $\text{MLP}_{\text{text}}(\cdot)$ and $\text{MLP}_{\text{vis}}(\cdot)$ denote the projection functions for textual and visual features, respectively.

We adapt the original SAM [16] *mask decoder* to address the specific challenges of camouflaged object segmentation by introducing semantic conditioning and edge-aware enhancements. The modified decoder integrates multi-level image features X, condition prompts P_c, and output tokens T_{tokens}, including mask tokens T_{mask} and an edge token T_{edge}, to precisely localize objects and accurately refine boundaries (see Fig. 5(c)).

We first apply two conditional multi-way attention (CondWayAttn × 2) modules to model the interactions between image features, condition prompts, and tokens. Each block enables dense bidirectional information flow between these components. Specifically, it includes image-to-token and image-to-condition attention to incorporate visual context, token-to-condition and token-to-image attention to align output tokens with semantic and spatial cues, and token self-attention and an MLP layer to capture intra-token dependencies and to perform feature transformation. The enhanced outputs are computed as

$$\tilde{X}, \tilde{T}_{\text{mask}}, \tilde{T}_{\text{edge}} = \text{CondWayAttn}(X, P_c, T_{\text{token}}) \tag{5}$$

The attention-enhanced features $\tilde{X}$ are first upsampled using a transposed convolution to restore spatial resolution. To recover fine details, these features are fused with shallow image features X_{shallow} through a fusion block defined as

$$\begin{aligned} X_{\text{fusion}} &= \text{TConv}(\tilde{X}) \\ &\quad + \text{Conv}\big(\text{ReLU}\big(\text{Norm}\big(\text{Conv}(X_{\text{shallow}})\big)\big)\big) \end{aligned} \tag{6}$$

The attention-enhanced mask and edge tokens are then projected via task-specific MLPs. The coarse segmentation mask is computed by element-wise multiplication of the mask token with the upsampled features

$$M_{\text{coarse}} = \text{MLPs}(\tilde{T}_{\text{mask}}) \odot \text{TConv}(\tilde{X}) \tag{7}$$

Similarly, the edge map is predicted by combining the edge token with the fused feature map:

$$E = \text{MLP}(\tilde{T}_{\text{edge}}) \odot X_{\text{fusion}} \tag{8}$$

The final refined mask incorporates edge guidance by multiplying the coarse mask by the edge map, followed by residual addition:

$$M_{\text{fine}} = M_{\text{coarse}} + (M_{\text{coarse}} \otimes E) \tag{9}$$

This edge-guided refinement enhances boundary accuracy while preserving regional consistency, effectively handling low-contrast and subtle camouflaged structures. The effectiveness of this module is evidenced by the ablation study in Section 4.3.5.

For the *loss function*, we adopt two supervised losses: a mask loss for segmentation and an edge loss for boundary refinement.

Following Ref. [52], the predicted mask M is rescaled to the input resolution and compared to the ground-truth mask G_m. Supervision combines binary cross-entropy (BCE) [53] and intersection-over-union (IoU) losses [54]:

$$\mathcal{L}_{\text{mask}} = \mathcal{L}_{\text{bce}}(M, G_m) + \mathcal{L}_{\text{iou}}(M, G_m) \tag{10}$$

For edge estimation, we follow Ref. [1] and supervise the predicted edge E against the ground-truth edge G_e using the Dice loss [55]:

$$\mathcal{L}_{\text{edge}} = \mathcal{L}_{\text{dice}}(E, G_e) \tag{11}$$

The overall objective function is defined as the unweighted sum of the two losses:

$$\mathcal{L} = \mathcal{L}_{\text{mask}} + \mathcal{L}_{\text{edge}} \tag{12}$$

4 Experiments

4.1 Implementation details

4.1.1 Datasets

We evaluated our method on two tasks: camouflaged object segmentation (COS) and open-vocabulary COS (OVCOS). For the OVCOS task, all experiments were conducted on the OVCamo [1] dataset, a benchmark specifically designed for this setting. It comprises 11,483 images sourced from various publicly available datasets, covering 75 camouflaged object categories embedded in complex natural scenes. To enable open-vocabulary evaluation, the dataset is divided into two disjoint subsets by category: the training set $\mathcal{D}_{\text{train}}$ includes 7713 images in 14 seen categories, while the test set $\mathcal{D}_{\text{test}}$ contains 3770 images in 61 unseen categories, following an approximate 7:3 split.

For the COS task, we evaluated on three widely used benchmarks: CAMO [47], COD10K [5], and NC4K [42]. A total of 4040 images from CAMO and COD10K were used for training. We conducted our evaluation on the remaining images from these datasets, as well as the entire NC4K set. Detailed statistics for all datasets, including training/testing splits, are presented in Table 1.

4.1.2 Evaluation metrics

To ensure fair and comprehensive evaluation of OVCOS performance, we adopted a set of evaluation metrics tailored for OVCOS, which are adapted from those originally proposed for the camouflaged scene understanding task [5, 56]. Specifically, we use six metrics: class structure measure cS_m, class weighted F-measure cF^w_β, class mean absolute error cMAE, class standard F-measure cF_β, class enhanced alignment measure cE_m, and class intersection over union cIoU. These metrics are standard in the open-vocabulary segmentation literature [7, 12, 15, 24, 25, 57], jointly assessing classification accuracy and segmentation quality for a balanced evaluation of model performance.

For the COS task, we followed established protocols [5] and adopted four commonly used metrics: structure measure S_α, enhanced alignment measure E_ϕ, weighted F-measure F^ω_β, and mean absolute error MAE. The first three evaluate structural and region-aware similarity between predictions and ground truth, where higher values

Table 1 Summary of datasets used for OVCOS and COS tasks

Dataset	Task	Total	Train	Test	Categories
OVCamo	OVCOS	11,483	7713	3770	75 (14/61)
CAMO	COS	1250	1000	250	—
COD10K	COS	5066	3040	2026	—
NC4K	COS	4121	—	4121	—

indicate better performance. Conversely, MAE measures pixel-wise error, with lower values indicating better accuracy.

4.1.3 Training details

We carried out all experiments on a workstation equipped with two NVIDIA RTX 3090Ti GPUs, using Ubuntu 20.04. Our framework was implemented in PyTorch and used CUDA 11.8 for GPU acceleration.

During CLIP model fine-tuning, we adopted a multi-modal prompting strategy on the pre-trained ViT-L/14 Alpha-CLIP model [28]. The model was trained on the OVCamo [1] dataset for 10 epochs using stochastic gradient descent (SGD) with a learning rate of 0.0035 and a batch size of 8 on a single GPU, following the setup in Ref. [19]. Additionally, the input alpha mask was randomly selected as either an all-one mask or the ground-truth segmentation mask with equal probability, balancing global context encoding and localized focus.

During adapted SAM training, the fine-tuned CLIP was integrated into our adapted SAM architecture, based on the ViT-H variant of SAM [16]. The network was trained for 20 epochs using the Adam optimizer with an initial learning rate of 2×10^{-4}, and decayed by cosine annealing. Training was conducted on two GPUs with a batch size of 2 and completed in approximately 24 h.

4.2 Comparison to the state-of-the-art

In this section, we compare our method to state-of-the-art approaches on both OVCOS and COS tasks, providing qualitative, quantitative, and efficiency comparisons.

4.2.1 Qualitative comparisons on OVCOS

We first present sample open-vocabulary camouflaged object segmentation and classification results using the OVCamo [1] dataset in Fig. 6, Our method consistently delivers superior segmentation quality,

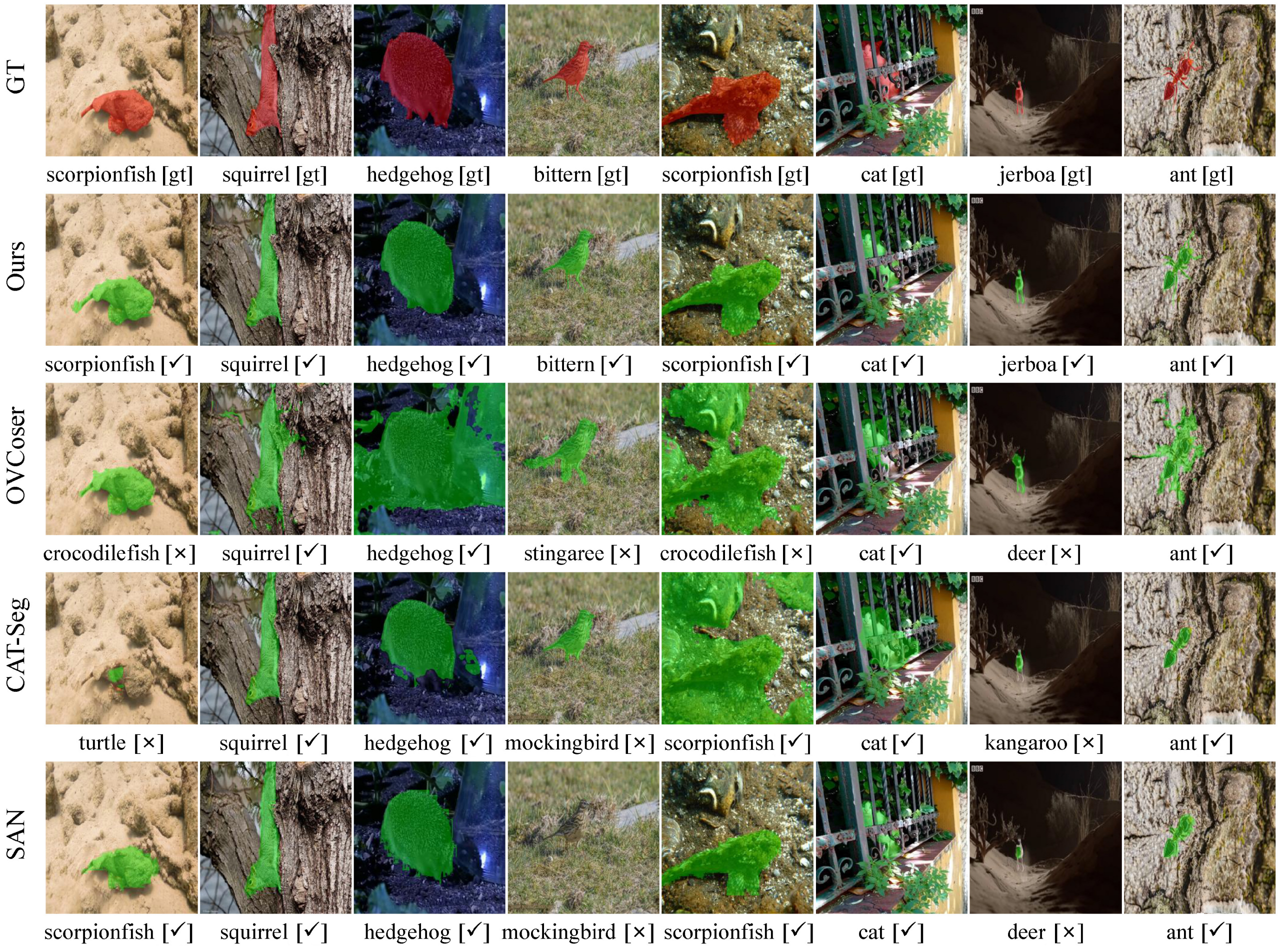


Fig. 6 Qualitative comparison between our method and CLIP-based baselines on OVCamo. Columns depicts input images with segmentation result and predicted label. The predicted label is shown below each segmentation result, where [✓] indicates correct prediction, and [✗] denotes an incorrect one.

accurately delineating camouflaged objects with well-preserved shapes and precise boundaries—even in low-contrast settings with cluttered backgrounds. Compared to other methods, our approach better maintains object integrity and minimizes background leakage, demonstrating enhanced robustness in camouflage scenarios.

In terms of classification, our method reliably predicts correct categories across diverse samples, outperforming prior methods that often misclassify visually ambiguous targets. This classification accuracy improvement stems from our region-aware classification strategy, which integrates segmentation masks as alpha masks into the fine-tuned CLIP model. Combined with multi-modal prompting and edge-aware decoding, our method achieves accuracy of both localization and recognition under open-vocabulary conditions.

4.2.2 Quantitative comparisons on OVCOS

To thoroughly evaluate the effectiveness of our proposed framework, we compared it to recent state-of-the-art open-vocabulary segmentation methods, including CAT-Seg [7], SAN [24], SimSeg [12], OVSeg [15], FC-CLIP [25], ODISE [51], and the baseline OVCoser [1]. For a fair comparison, all models were trained or fine-tuned on the OVCamo [1] dataset. We adopted the large variants of such approaches wherever available, except for SimSeg, which is only released in its base form. As Table 2 shows, our method consistently outperforms all competitors across multiple evaluation metrics. Table 2 summarizes quantitative results on the OVCamo [1] dataset. While open-vocabulary segmentation methods such as SAN [24], OVSeg [15], and CAT-Seg [7] benefit from large-scale pretraining, they lack task-specific adaptation, resulting in limited performance on OVCOS (e.g., OVSeg: 0.164 cS_m, 0.123 cIoU). The baseline OVCoser [1] improves results by integrating camouflage segmentation with CLIP-based classification, achieving 0.579 cS_m and 0.443 cIoU, but it does not fine-tune vision-language embeddings or incorporate semantic guidance into segmentation.

Unlike existing methods, our framework leverages fine-tuned CLIP and a task-adapted SAM to enhance both segmentation and classification. It achieves state-of-the-art results, surpassing the baseline OVCoser [1] by notable margins across all metrics: +8.9% in cS_m, +12.5% in cIoU, +12.5% in cF_β^w, +11.1% in cF_β, +8.1% in cE_m, and a reduction of 7.1% in cMAE. These results highlight the effectiveness of our cascaded design and cross-modal semantic conditioning in tackling the OVCOS challenge.

4.2.3 Quantitative comparisons on COS

As Table 3 shows, our adapted SAM model achieves competitive performance across three widely used COS benchmarks: CAMO [47], COD10K [5], and NC4K [42]. Compared to both traditional non-SAM-based methods [5, 41–46] and recent SAM-based approaches [16, 48, 49], our model consistently outperforms others across all datasets.

Specifically, the adapted SAM ranks first on 11 out of 12 evaluation metrics and second on the remaining one, demonstrating strong generalization and robustness in diverse camouflage scenarios. Our method achieves notable improvements in structure-aware metrics (S_α, E_ϕ), region-aware precision (F_β^ω), and pixel-level accuracy (MAE), particularly on the COD10K and NC4K datasets. These results highlight the effectiveness of our edge-enhanced architecture and prompt-guided segmentation in capturing fine-grained boundaries and ensuring semantic consistency.

Table 2 Comparison of our method to state-of-the-art CLIP-based OVCOS approaches on the OVCamo dataset. Bold values indicate the results of our method, which achieves the best overall performance. The second best is underlined

Model	VLM	Train set	Fine tune	cS_m ↑	cF_β^w ↑	cMAE ↓	cF_β ↑	cE_m ↑	cIoU ↑
SimSeg	CLIP-ViT-B/16	COCO-Stuff	OVCamo	0.098	0.071	0.852	0.081	0.128	0.0
OVSeg	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.164	0.131	0.763	0.147	0.208	0.123
ODISE	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.182	0.125	0.691	0.219	0.309	0.189
SAN	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.321	0.216	0.550	0.236	0.331	0.204
FC-CLIP	CLIP-ConvNeXt-L	COCO-Stuff	OVCamo	0.124	0.074	0.798	0.088	0.162	0.072
CAT-Seg	CLIP-ViT-L/14	COCO-Stuff	OVCamo	0.185	0.094	0.702	0.110	0.185	0.088
OVCoser	CLIP-ConvNeXt-L	OVCamo	—	0.579	0.490	0.336	0.520	0.616	0.443
Ours	Our Fine-Tuned CLIP	OVCamo	—	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

Table 3 COS performance on CAMO, COD10K, and NC4K datasets. The best performance per metric is highlighted in bold, and the second best is underlined

Method	CAMO				COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE $\downarrow$
SINet	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051	0.808	0.883	0.768	0.058
RankNet	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045	0.840	0.904	0.802	0.048
PFNet	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040	0.829	0.887	0.784	0.053
SINetV2	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.770	0.048
ZoomNet	0.820	0.892	0.752	0.066	0.838	0.911	0.729	0.029	0.853	0.912	0.784	0.043
SegMaR	0.815	0.872	0.742	0.071	0.833	0.895	0.724	0.033	0.841	0.905	0.781	0.046
DGNet	0.839	0.901	0.769	0.057	0.822	0.896	0.693	0.033	0.857	0.911	0.784	0.042
SAM	0.684	0.687	0.606	0.132	0.783	0.798	0.701	0.050	0.767	0.776	0.696	0.078
SAM-Adapter	0.847	0.873	0.765	0.070	0.883	0.918	0.801	0.025	—	—	—	—
MedSAM	0.820	**0.904**	0.779	0.065	0.841	0.917	0.751	0.033	0.866	0.929	0.821	0.041
Ours	**0.865**	0.902	**0.789**	**0.057**	**0.905**	**0.947**	**0.845**	**0.019**	**0.904**	**0.933**	**0.852**	**0.031**

4.2.4 Model size and runtime

We compare inferencing time and memory usage of models in Table 4, with results reported per image on a single NVIDIA RTX 3090 Ti (24 GB). Our method is efficient: SAM-ViT-B requires only 240 ms and 9.8 GB, offering a favorable balance of speed and accuracy. Larger backbones (SAM-ViT-L/H) deliver further performance gains at the expense of runtime and memory, enabling flexible trade-offs under different resource constraints.

As Table 5 shows, even the lightweight SAM-ViT-B already surpasses the baseline OVCoser [1], while SAM-ViT-H achieves the best overall results. This confirms that our method scales effectively, adapting to both efficiency-critical and accuracy-oriented applications.

Table 4 Comparison of inferencing time and memory requirement of different models. Inferencing time measured with batch size = 1 on a single NVIDIA RTX 3090 Ti (24 GB)

Model	Backbone	Time (ms)	Mem (GB)
SimSeg	ResNet101	350	6.5
OVSeg	Swin-B	980	15.2
ODISE	StableDiffusion	860	16.8
SAN	ViT Adapter	160	5.4
CAT-Seg	Swin-B	140	4.8
FC-CLIP	CLIP-CvNeXt-L	210	7.2
OVCoser	CLIP-CvNeXt-L	125	3.9
Ours	SAM-ViT-B	240	9.8
Ours	SAM-ViT-L	370	13.5
Ours	SAM-ViT-H	530	19.7

4.3 Ablation and related studies

4.3.1 Effectiveness of the fine-tuned CLIP

To evaluate the effectiveness of our fine-tuned CLIP, we first compared different VLMs under multiple cropping strategies.

Table 6 presents the results for CLIP-ConvNeXt-L [58], CLIP-ViT-L/14 [11], Alpha-CLIP [28], and our fine-tuned CLIP. For the same backbone, replacing *full-image* classification with *hard* cropping degrades performance, underscoring the mismatch between full-image pretraining and cropped inference [1]. Our fine-tuned CLIP with soft cropping achieves the strongest results across all metrics, for all settings, demonstrating that the soft prior effectively mitigates the mismatch while further benefiting from the enhanced semantic representations learned during fine-tuning.

To verify the classification performance of our fine-tuned CLIP, we tested our CLIP with both *all one* and *ground-truth masks* as the soft prior on the OVCamo [1] test set. The results in Table 7 clearly show that our method enhances classification accuracy in both the *all-one* and *gt* settings. These

Table 5 Performance comparison with different backbones, showing that even the lightweight SAM-ViT-B achieves significant gains over OVCoser while maintaining low computational cost

Model	Backbone	$cS_m \uparrow$	$cF_\beta^w \uparrow$	cMAE $\downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	cIoU $\uparrow$
OVCoser	CLIP-ConvNeXt-L	0.579	0.490	0.336	0.520	0.616	0.443
Ours	SAM-ViT-B	0.614	0.519	0.278	0.552	0.650	0.461
Ours	SAM-ViT-L	0.659	0.600	0.267	0.614	0.691	0.549
Ours	SAM-ViT-H	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

Table 6 Ablation study on mask cropping strategies for VLM-based OVCOS. *Full-image* uses the entire image for classification without cropping. *Hard* cropping directly removes surrounding context using the segmentation mask, while *soft* cropping blends the segmentation mask with the original image, preserving contextual cues

Model	VLM	Crop	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE ↓	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU ↑
COCUS	CLIP-ConvNeXt-L	*Full-image*	0.573	0.524	0.365	0.544	0.607	0.495
COCUS	CLIP-ConvNeXt-L	*Hard*	0.567	0.518	0.375	0.534	0.591	0.481
COCUS	CLIP-ViT-L/14	*Full-image*	0.591	0.545	0.343	0.562	0.629	0.515
COCUS	CLIP-ViT-L/14	*Hard*	0.580	0.536	0.353	0.551	0.617	0.503
COCUS	Alpha-CLIP	*Soft*	0.639	0.589	0.299	0.603	0.668	0.545
COCUS	Our fine-tuned CLIP	*Soft*	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

Table 7 Classification performance of different CLIP models on the OVCamo test set

Model	Alpha	Top-1↑	Top-5↑
CLIP-ConvNeXt-L	—	0.6944	0.8918
CLIP-ViT-L/14	—	0.7040	0.8915
Alpha-CLIP	all one	0.6934	0.8849
Alpha-CLIP	gt	0.7467	0.9456
Ours	all one	0.7462	0.9003
Ours	gt	**0.7859**	**0.9497**

results highlight the effectiveness of task-specific CLIP fine-tuning, while maximizing performance when reliable masks are available.

4.3.2 Alpha mask selection probability

We conducted a study to evaluate the impact of the selection probability P between *all-one* and *gt* masks during CLIP fine-tuning. From Table 8, we observe that the choice of P has minimal effect on the performance metrics across the board. The results are relatively stable across all values of P, with $P = 0.5$ achieving the best performance in all metrics. Based on these results, we adopted $P = 0.5$ as the default configuration in our framework.

4.3.3 Impact of different text templates

We conducted additional experiments to evaluate the impact of different text prompt templates on classification performance. Specifically, we compared a generic prompt: "*A photo of a <cls>.*" with our proposed camouflage-specific prompt: "*A photo of the <cls> camouflaged in the background.*". As Table 9 shows, the camouflage-specific template consistently outperformed the generic one.

Table 8 Effect of selection probability P between *all-one* and *gt* masks during CLIP fine-tuning

P	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE ↓	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU ↑
0.00	0.659	0.607	0.266	0.624	0.688	0.559
0.25	0.660	0.610	0.266	0.619	0.690	0.558
0.50	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**
0.75	0.663	0.611	0.267	0.621	0.693	0.560
1.00	0.661	0.611	0.268	0.623	0.691	0.561

Table 9 Comparison of generic and camouflage-specific prompt templates for the fine-tuned CLIP

Model	Prompt template	Top-1↑
Fine-tuned CLIP	A photo of a <cls>.	0.7762
Fine-tuned CLIP	A photo of the <cls> camouflaged in the background.	**0.7859**

4.3.4 Benefit of the two-stage pipeline

We tested the effectiveness of the two-stage *segmentation + classification* pipeline by comparing it to a single-stage pipeline that performs both tasks simultaneously. See Table 10. The metrics: cS_m, cF_β^w, cMAE, cF_β, cE_m, and cIoU, reflect both segmentation quality and classification accuracy. Compared to the one-stage pipeline, the two-stage pipeline consistently achieves higher performance across all metrics. These results demonstrate that, though the two-stage framework might accumulate errors, its overall benefits outweigh the drawbacks.

4.3.5 Impact of the adapted mask decoder

In Table 11, we report ablation studies on the OVCamo [1] dataset to assess the effectiveness of the proposed conditional multi-way attention (CMA) and edge enhancement (EDE) modules in our adapted mask decoder. The baseline is built upon SAM [16] with a lightweight adapter, corresponding to the original SAM mask decoder without either

Table 10 Comparison of one-stage and two-stage pipelines

Pipeline	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE ↓	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU ↑
One stage	0.657	0.605	0.274	0.615	0.685	0.554
Two stage	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

Table 11 Ablation study on conditional multi-way attention (CMA) and edge enhancement (EDE) in the adapted mask decoder

Model	$cS_m\uparrow$	$cF_\beta^w\uparrow$	cMAE ↓	$cF_\beta\uparrow$	$cE_m\uparrow$	cIoU ↑
Baseline	0.644	0.599	0.281	0.610	0.651	0.549
+ EDE	0.650	0.605	0.278	0.615	0.666	0.554
+ CMA	0.652	0.607	0.273	0.621	0.683	0.551
Ours	**0.668**	**0.615**	**0.265**	**0.631**	**0.697**	**0.568**

enhancement. Despite its simplicity, the baseline delivers strong performance due to two key factors: (i) SAM provides powerful, pre-trained feature representations on large and diverse data, offering robust and generalizable priors, and (ii) our two-stage design enables even a lightweight adapter to effectively exploit these priors.

Based on the strong baseline, our method further improves the performance by near 10% in cMAE and cS_m. Specifically, adding the EDE module (+ EDE) leads to notable gains in contour-sensitive metrics such as cF_β^w and cIoU, indicating that explicit edge modeling enhances boundary precision. In contrast, introducing the CMA module (+ CMA) results in stronger improvements in semantic-aware metrics like cE_m and cF_β, demonstrating that conditional attention effectively enriches textual-visual feature fusion.

When both modules are combined, our method achieves the best performance across all metrics, underscoring the complementary strengths of semantic conditioning and edge-aware refinement. These findings confirm the importance of both enhancements in improving segmentation quality for challenging camouflaged object scenarios.

4.3.6 *Sensitivity of classification to mask quality*

The segmentation mask of the first stage serves as a soft prior for the second-stage classification. To assess the sensitivity of classification to mask quality, we simulated erroneous segmentation masks by applying morphological operations to a correctly predicted mask. Specifically, the predicted segmentation mask was dilated and eroded by a certain number of pixels, generating expanded and shrunk variants. These simulated erroneous masks were used as the soft prior in the second-stage classification. We demonstrate an example in Fig. 7, and summarize quantitative results in Fig. 8.

In Fig. 7, the dilated mask (top row) still preserved most of the camouflaged target, enabling correct classification (*scorpionfish*). However, the eroded mask (bottom row) excluded critical target regions, resulting in misclassification (*egyptian nightjar*). This indicates that while soft cropping is tolerant to moderate inaccuracies (e.g., dilation), it remains sensitive to severe erosion that removes essential object content.

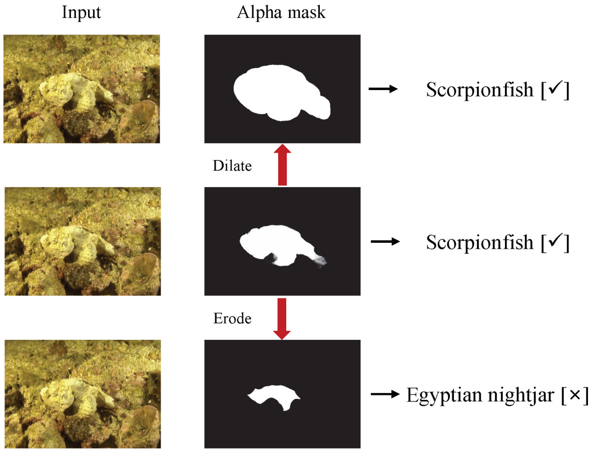


Fig. 7 Classification result with erroneous segmentation predictions. Left: input image. Center: dilated, original, and eroded masks. Right: corresponding classification results. The dilated and original masks retain the target object (scorpionfish) and lead to correct classification. The eroded mask loses key target regions, causing misclassification (egyptian nightjar).

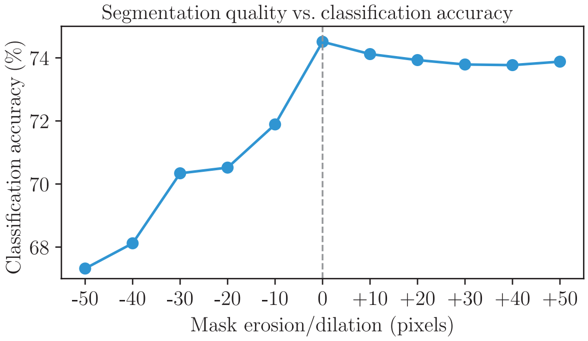


Fig. 8 Classification accuracy for varying degrees of mask erosion (-) and dilation (+). Classification accuracy degrades significantly when the mask is severely eroded, while it remains relatively stable with moderate dilation.

To quantify this effect, we analyzed classification accuracy for various degrees of mask erosion and dilation. See Fig. 8. Accuracy peaks with the original predicted mask. Performance remains stable under moderate dilation but declines sharply with severe erosion. These results confirm that the soft prior mask ensures reliable classification as long as key target regions are preserved, with significant degradation of accuracy when the mask undergoes substantial structural loss.

4.3.7 *Contribution of boundary refinement and soft prior*

We further investigated the contribution of the proposed edge enhancement module (EDE) and the two-stage pipeline by conducting an interpretability analysis. Specifically, instead of merely verifying whether they improve performance, we visualized and compared the segmentation results with and without EDE, as well as the classification outcomes from single-stage and two-stage pipelines, in order to better reveal how these components influence the decision process.

As Fig. 9 shows, segmentation masks predicted with the EDE module exhibit sharper and more accurate boundaries compared to those without it.

Figure 10 visualizes classification results with and without the soft prior mask. Without the mask (*w/o*), classifying the full image leads to a misclassification as a crocodilefish. With the mask (*w/*), the soft cropping strategy focuses the model on the target region, enabling correct recognition of the scorpionfish.

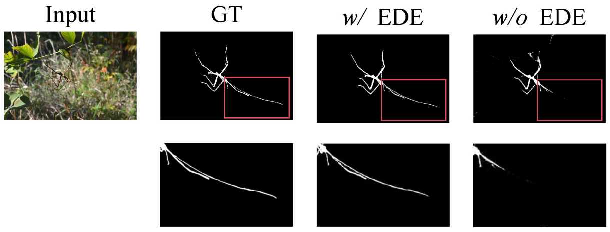


Fig. 9 Predicted segmentation masks with and without the EDE module.

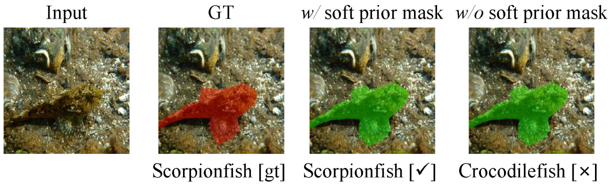


Fig. 10 Effect of the first-stage soft prior mask on classification. [✓] indicates correct prediction, and [✗] denotes an incorrect one.

4.3.8 Effects of choices on segmentation

We present visual results to illustrate the benefits of our proposed modules in both segmentation and classification. Figure 11 presents results for the baseline, CMA only, EDE only, and the full model (CMA + EDE), alongside the ground truth. The baseline fails to accurately localize and delineate the camouflaged target, producing incomplete or noisy masks. Incorporating CMA enhances semantic focus by better identifying the target region, while EDE improves boundary quality, yielding sharper and more continuous contours. Our complete framework, combining CMA and EDE, produces results most consistent with the ground truth, confirming their complementary contributions.

4.3.9 Choice of classification strategy

Figure 12 compares three classification strategies: *full-image*, *hard crop*, and our proposed *soft crop*. The full-image approach suffers from background distractions, often leading to misclassification. Hard cropping

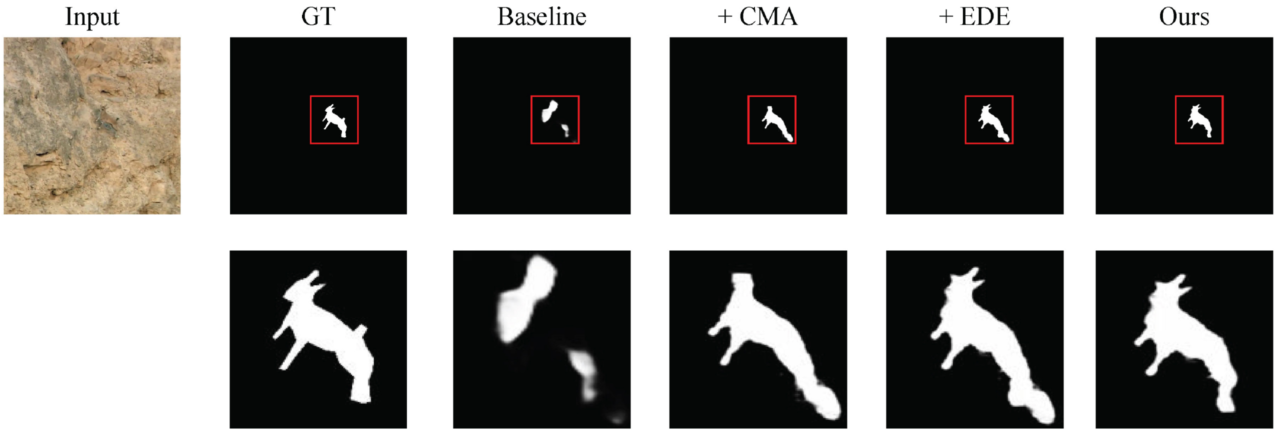


Fig. 11 Contribution of CMA and EDE: results for the baseline, CMA only, EDE only, and the full model compared to the ground truth.

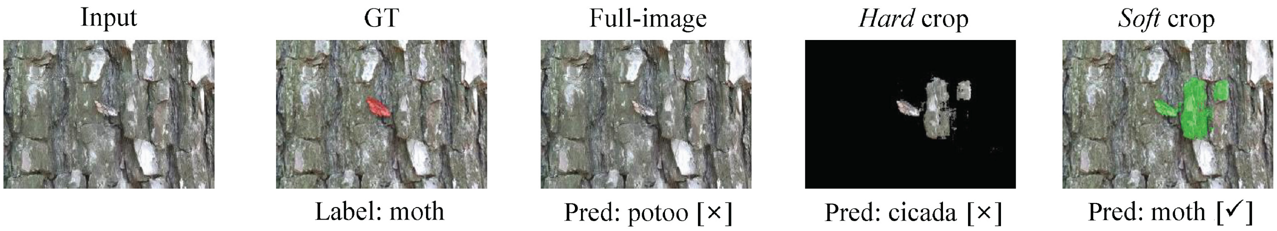


Fig. 12 Comparison of classification strategies: *full-image*, *hard crop*, and our proposed *soft crop*. [✓] indicates correct prediction, and [✗] denotes an incorrect one.

introduces a domain gap, as CLIP [11] is pre-trained on full images. By directly cropping around the mask, it discards useful contextual information and may retain irrelevant regions, thereby distorting image structure and reducing classification accuracy. In contrast, our soft cropping strategy preserves global context while emphasizing the target region, enabling accurate recognition without disrupting CLIP's pre-training assumptions.

4.3.10 Attention and edge feature visualizations

We provide visualizations in Fig. 13 to highlight the effects of CMA and EDE.

The attention map in Fig. 13(b) is derived from the cross-attention between semantic condition prompts and SAM [16] image features within CMA. It clearly focuses on the camouflaged target while suppressing irrelevant background textures, thereby enhancing semantic focus under low contrast conditions.

For EDE, the visualization in Fig. 13(c) corresponds to the edge-related feature map X_{fusion} in Eq. (8). After channel aggregation, this feature emphasizes thin, continuous contours aligned with object boundaries, while maintaining low responses in the background. This demonstrates improved boundary quality prior to edge prediction.

4.4 Cross-domain generalization

To test the generalizability of the proposed method, we extended our experiments to two downstream tasks: (i) medical polyp segmentation using the Kvasir [59] dataset, and (ii) agricultural concealed crop detection using the ACOD-12K [60] dataset. This evaluation provided a rigorous examination of our method's adaptability and practical utility beyond the scope of the primary benchmark setting.

4.4.1 Dataset overview

The Kvasir [59] dataset is a widely used benchmark for polyp segmentation in gastrointestinal endoscopy. It contains high-resolution colonoscopy images with pixel-accurate polyp annotations. It poses challenges due to substantial variations in polyp size, shape, and texture.

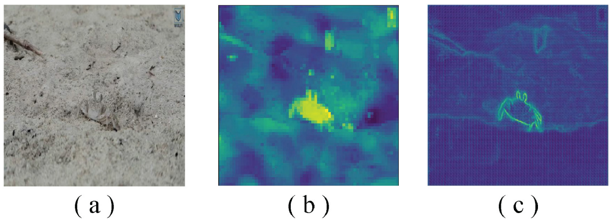


Fig. 13 Visualization of CMA and EDE: (a) input image containing a low-contrast, camouflaged target, (b) cross-attention map between semantic condition prompts and SAM image features in CMA, (c) channel-aggregated edge feature map from EDE.

The ACOD-12K [60] dataset is a large-scale agricultural dataset for concealed crop detection in dense, real-world greenhouse environments. It contains images in which crops are capture crops partially or fully occluded by leaves, stems, and support structures, with complex backgrounds, variable lighting, and subtle visual cues—making detection highly challenging.

4.4.2 Experimental setup

To evaluate our method in downstream tasks, we fine-tuned our model for two scenarios:

(1) *Medical domain:* Following Ref. [61], our model was fine-tuned on the Kvasir [59] dataset.

(2) *Agricultural domain:* Following Ref. [60], our model was fine-tuned on the ACOD-12K [60] dataset.

Example segmentation results are visualized in Fig. 14 and quantitative comparisons are summarized in Tables 12 and 13. Our method achieved the best results on both datasets compared to strong baselines and state-of-the-art methods. These results validate the robustness and practical applicability of our approach across diverse real-world domains.

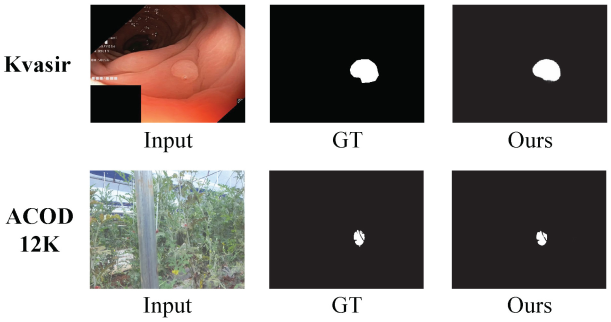


Fig. 14 Qualitative comparison on the Kvasir (medical) and ACOD-12K (agricultural) datasets showing original images, ground truth masks, and our predicted segmentation.

Table 12 Quantitative comparison on the Kvasir medical polyp segmentation dataset. Bold values indicate best results, and underlined values indicate second-best results

Model	wFm ↑	Sm ↑	Emax ↑	MAE ↓
U-Net	0.7940	0.8580	0.8930	0.0550
U-Net++	0.8080	0.8620	0.9100	0.0480
SFA	0.6700	0.7820	0.8490	0.0750
Ours	**0.8551**	**0.9057**	**0.9223**	**0.0370**

Table 13 Quantitative comparison on the ACOD-12K agricultural concealed crop detection dataset. Bold values indicate the best results, and underlined values indicate the second-best results

Model	Sm ↑	wFm ↑	Emax ↑
SINet	0.7450	0.4740	0.8260
PFNet	0.8050	0.6850	0.9420
ZoomNet	0.8320	0.7470	0.9400
FSPNet	0.7190	0.5260	0.8190
Ours	**0.8455**	**0.8002**	**0.9625**

5 Conclusions

In this paper, we have presented COCUS, a two-stage framework for OVCOS that explicitly decouples segmentation and classification. In the first stage, visual and textual embeddings are extracted using our fine-tuned CLIP model. These embeddings guide an adapted SAM with a redesigned mask decoder to enhance object localization and boundary precision. In the second stage, the predicted segmentation mask is fused with the input image to guide the attention of the model toward the target regions, enabling region-aware classification without relying on cropped inputs. Extensive experiments on both OVCOS and COS benchmarks show that our method outperforms existing open-vocabulary segmentation methods. The adapted SAM also achieves superior results on the COS benchmarks. These experiments confirm the benefits of our two-stage framework and edge-aware enhancements in complex camouflage scenarios.

Availability of data and materials

We provide links to the code, models, and datasets used in our experiments to facilitate reproducibility and further research:

- The code and models of this paper are available at `https://github.com/intcomp/camouflaged-vlm`
- The OVCamo dataset can be accessed at `https://github.com/lartpang/OVCamo`
- The COD10K dataset is available at `https://dengpingfan.github.io/pages/COD.html`
- The CAMO dataset can be found at `https://sites.google.com/view/ltnghia/research/camo`
- The NC4K dataset is accessible at `https://github.com/JingZhang617/COD-Rank-Localize-and-Segment`

Competing interests

The authors have no competing interests to declare.

Funding

This work was partially supported by the National Natural Science Foundation of China (62372284, 62476143), the National Key Laboratory of Science and Technology on Space-Born Intelligent Information Processing (TJ-02-22-01), and the Pioneer R&D Program of Zhejiang Province (2024C01024).

Author contributions

K.Z. conceived the study and led the implementation. W.Y. contributed to implementation, experiments, and formal analysis. Z.W. and G.L. assisted with data processing and result interpretation. X.Z. supervised the project and revised the manuscript. D.F. contributed to methodology and evaluation design. D.Z. provided overall supervision and strategic guidance. All authors reviewed and approved the final manuscript.

References

[1] Pang, Y.; Zhao, X.; Zuo, J.; Zhang, L.; Lu, H. Open-vocabulary camouflaged object segmentation. In: *Computer Vision – ECCV 2024. Lecture Notes in Computer Science, Vol. 15105.* Leonardis, A.; Ricci, E.; Roth, S.; Russakovsky, O.; Sattler, T.; Varol, G. Eds. Springer Cham, 476–495, 2025.

[2] Cheng, B.; Schwing, A.; Kirillov, A. Per-pixel classification is not all you need for semantic segmentation. *arXiv preprint* arXiv:2107.06278, 2021.

[3] Xie, E.; Wang, W.; Yu, Z.; Anandkumar, A.; Alvarez, J. M.; Luo, P. SegFormer: Simple and efficient design for semantic segmentation with Transformers. *arXiv preprint* arXiv:2105.15203, 2021.

[4] Chen, L. C.; Zhu, Y.; Papandreou, G.; Schroff, F.; Adam, H. Encoder–decoder with atrous separable convolution for semantic image segmentation. In: *Computer Vision – ECCV 2018. Lecture Notes in Computer Science, Vol. 11211.* Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 833–851, 2018.

[5] Fan, D. P.; Ji, G. P.; Cheng, M. M.; Shao, L. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 44, No. 10, 6024–6042, 2022.

[6] Liu, L.; Wang, R.; Xie, C.; Yang, P.; Wang, F.; Sudirman, S.; Liu, W. PestNet: An end-to-end deep learning approach for large-scale multi-class pest detection and classification. *IEEE Access* Vol. 7, 45301–45312, 2019.

[7] 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, 4113–4123, 2024.

[8] Li, B.; Weinberger, K. Q.; Belongie, S.; Koltun, V.; Ranftl, R. Language-driven semantic segmentation. *arXiv preprint* arXiv:2201.03546, 2022.

[9] Bucher, M.; Vu, T. H.; Cord, M.; Perez, P. Zero-shot semantic segmentation. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 466–477, 2019.

[10] Xian, Y.; Choudhury, S.; He, Y.; Schiele, B.; Akata, Z. Semantic projection network for zero- and few-label semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8248–8257, 2020.

[11] 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. *arXiv preprint* arXiv:2103.00020, 2021.

[12] Xu, M.; Zhang, Z.; Wei, F.; Lin, Y.; Cao, Y.; Hu, H.; Bai, X. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13689.* Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 736–753, 2022.

[13] Ding, J.; Xue, N.; Xia, G.; Dai, D. Decoupling zero-shot semantic segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11573–11582, 2022.

[14] Ding, Z.; Wang, J.; Tu, Z. Open-vocabulary universal image segmentation with MaskCLIP. *arXiv preprint* arXiv:2208.08984, 2022.

[15] Liang, F.; Wu, B.; Dai, X.; Li, K.; Zhao, Y.; Zhang, H.; Zhang, P.; Vajda, P.; Marculescu, D. Open-vocabulary semantic segmentation with mask-adapted CLIP. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7061–7070, 2023.

[16] Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W. Y.; et al. Segment anything. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 4015–4026, 2023.

[17] Zhang, P.; Yan, T.; Liu, Y.; Lu, H. Fantastic animals and where to find them: Segment any marine animal with dual SAM. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2578–2587, 2024.

[18] Yan, T.; Wan, Z.; Deng, X.; Zhang, P.; Liu, Y.; Lu, H. MAS-SAM: Segment any marine animal with aggregated features. In: Proceedings of the 33rd Conference on Artificial Intelligence, 6886–6894, 2024.

[19] 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, 19113–19122, 2023.

[20] Zang, Y.; Li, W.; Zhou, K.; Huang, C.; Loy, C. C. Open-vocabulary DETR with conditional matching. In: *Computer Vision – ECCV 2022. Lecture Notes in Computer Science, Vol. 13669.* Avidan, S.; Brostow, G.; Cissé, M.; Farinella, G. M.; Hassner, T. Eds. Springer Cham, 106–122, 2022.

[21] Gu, X.; Lin, T. Y.; Kuo, W.; Cui, Y. Open-vocabulary object detection via vision and language knowledge distillation. In: Proceedings of the International Conference on Learning Representations, 2021.

[22] Zareian, A.; Dela Rosa, K.; Hu, D. H.; Chang, S. F. Open-vocabulary object detection using captions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 14388–14397, 2021.

[23] Li, S.; Li, M.; Wang, P.; Zhang, L. OpenSD: Unified open-vocabulary segmentation and detection. *arXiv preprint* arXiv:2312.06703, 2023.

[24] 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, 2945–2954, 2023.

[25] Yu, Q.; He, J.; Deng, X.; Shen, X.; Chen, L. C. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional CLIP. *arXiv preprint* arXiv:2308.02487, 2023.

[26] Luo, H.; Bao, J.; Wu, Y.; He, X.; Li, T. SegCLIP: Patch aggregation with learnable centers for open-vocabulary semantic segmentation. *arXiv preprint* arXiv:2211.14813, 2022.

[27] Xu, J.; Hou, J.; Zhang, Y.; Feng, R.; Wang, Y.; Qiao, Y.; Xie, W. Learning open-vocabulary semantic segmentation models from natural language supervision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2935–2944, 2023.

[28] Sun, Z.; Fang, Y.; Wu, T.; Zhang, P.; Zang, Y.; Kong, S.; Xiong, Y.; Lin, D.; Wang, J. Alpha-CLIP: A CLIP model focusing on wherever you want. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13019–13029, 2024.

[29] Zhou, K.; Yang, J.; Loy, C. C.; Liu, Z. Learning to prompt for vision-language models. *International Journal of Computer Vision* Vol. 130, No. 9, 2337–2348, 2022.

[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, 16795–16804, 2022.

[31] Bahng, H.; Jahanian, A.; Sankaranarayanan, S.; Isola, P. Exploring visual prompts for adapting large-scale models. *arXiv preprint* arXiv:2203.17274, 2022.

[32] Khattak, M. U.; Wasim, S. T.; Naseer, M.; Khan, S.; Yang, M. H.; Khan, F. S. Self-regulating prompts: Foundational model adaptation without forgetting. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 15144–15154, 2024.

[33] Yang, L.; Wang, Y.; Li, X.; Wang, X.; Yang, J. Fine-grained visual prompting. In: Proceedings of the 37th Conference on Neural Information Processing Systems, 24993–25006, 2023.

[34] Borji, A.; Cheng, M. M.; Hou, Q.; Jiang, H.; Li, J. Salient object detection: A survey. *Computational Visual Media* Vol. 5, No. 2, 117–150, 2019.

[35] Pang, Y.; Zhao, X.; Zhang, L.; Lu, H. Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9410–9419, 2020.

[36] Ji, W.; Yan, G.; Li, J.; Piao, Y.; Yao, S.; Zhang, M.; Cheng, L.; Lu, H. DMRA: Depth-induced multi-scale recurrent attention network for RGB-D saliency detection. *IEEE Transactions on Image Processing* Vol. 31, 2321–2336, 2022.

[37] Liu, J. J.; Hou, Q.; Liu, Z. A.; Cheng, M. M. PoolNet+: Exploring the potential of pooling for salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 45, No. 1, 887–904, 2023.

[38] Li, J.; Ji, W.; Zhang, M.; Piao, Y.; Lu, H.; Cheng, L. Delving into calibrated depth for accurate RGB-D salient object detection. *International Journal of Computer Vision* Vol. 131, No. 4, 855–876, 2023.

[39] Ji, W.; Li, J.; Bian, C.; Zhou, Z.; Zhao, J.; Yuille, A.; Cheng, L. Multispectral video semantic segmentation: A benchmark dataset and baseline. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1094–1104, 2023.

[40] Zhao, X.; Pang, Y.; Yang, J.; Zhang, L.; Lu, H. Multi-source fusion and automatic predictor selection for zero-shot video object segmentation. In: Proceedings of the 29th ACM International Conference on Multimedia, 2645–2653, 2021.

[41] Fan, D. P.; Ji, G. P.; Sun, G.; Cheng, M. M.; Shen, J.; Shao, L. Camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2774–2784, 2020.

[42] Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; Fan, D. P. Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 11586–11596, 2021.

[43] Mei, H.; Ji, G. P.; Wei, Z.; Yang, X.; Wei, X.; Fan, D. P. Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8768–8777, 2021.

[44] Pang, Y.; Zhao, X.; Xiang, T. Z.; Zhang, L.; Lu, H. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2150–2160, 2022.

[45] Jia, Q.; Yao, S.; Liu, Y.; Fan, X.; Liu, R.; Luo, Z. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4703–4712, 2022.

[46] Ji, G. P.; Fan, D. P.; Chou, Y. C.; Dai, D.; Liniger, A.; Van Gool, L. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research* Vol. 20, No. 1, 92–108, 2023.

[47] Le, T. N.; Nguyen, T. V.; Nie, Z.; Tran, M. T.; Sugimoto, A. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding* Vol. 184, 45–56, 2019.

[48] Chen, T.; Zhu, L.; Ding, C.; Cao, R.; Wang, Y.; Zhang, S.; Li, Z.; Sun, L.; Zang, Y.; Mao, P. SAM-adapter: Adapting segment anything in underperformed scenes. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops, 3359–3367, 2023.

[49] Ma, J.; He, Y.; Li, F.; Han, L.; You, C.; Wang, B. Segment anything in medical images. *Nature Communications* Vol. 15, Article No. 654, 2024.

[50] Zhao, H.; Puig, X.; Zhou, B.; Fidler, S.; Torralba, A. Open vocabulary scene parsing. In: Proceedings of the IEEE International Conference on Computer Vision, 2021–2029, 2017.

[51] Xu, J.; Liu, S.; Vahdat, A.; Byeon, W.; Wang, X.; De Mello, S. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2955–2966, 2023.

[52] Yin, B.; Zhang, X.; Fan, D. P.; Jiao, S.; Cheng, M. M.; Van Gool, L.; Hou, Q. CamoFormer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 46, No. 12, 10362–10374, 2024.

[53] De Boer, P. T.; Kroese, D. P.; Mannor, S.; Rubinstein, R. Y. A tutorial on the cross-entropy method. *Annals of Operations Research* Vol. 134, No. 1, 19–67, 2005.

[54] Máttyus, G.; Luo, W.; Urtasun, R. DeepRoadMapper: Extracting road topology from aerial images. In: Proceedings of the IEEE International Conference on Computer Vision, 3458–3466, 2017.

[55] Milletari, F.; Navab, N.; Ahmadi, S. A. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In: Proceedings of the 4th International Conference on 3D Vision, 565–571, 2016.

[56] Cheng, X.; Xiong, H.; Fan, D. P.; Zhong, Y.; Harandi, M.; Drummond, T.; Ge, Z. Implicit motion handling for video camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13854–13863, 2022.

[57] Zhu, C.; Chen, L. A survey on open-vocabulary detection and segmentation: Past, present, and future. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 46, No. 12, 8954–8975, 2024.

[58] Cherti, M.; Beaumont, R.; Wightman, R.; Wortsman, M.; Ilharco, G.; Gordon, C.; Schuhmann, C.; Schmidt, L.; Jitsev, J. Reproducible scaling laws for contrastive language-image learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2818–2829, 2023.

[59] Pogorelov, K.; Randel, K. R.; Griwodz, C.; Eskeland, S. L.; de Lange, T.; Johansen, D.; Spampinato, C.; Dang-Nguyen, D. T.; Lux, M.; Schmidt, P. T.; et al. KVASIR: A multi-class image dataset for computer aided gastrointestinal disease detection. In: Proceedings of the 8th ACM on Multimedia Systems Conference, 164–169, 2017.

[60] Wang, L.; Yang, J.; Zhang, Y.; Wang, F.; Zheng, F. Depth-aware concealed crop detection in dense agricultural scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 17201–17211, 2024.

[61] Fan, D. P.; Ji, G. P.; Zhou, T.; Chen, G. PraNet: Parallel reverse attention network for polyp segmentation. In: *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020. Lecture Notes in Computer, Vol. 12266.* Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; Shao, L. Eds. Springer Cham, 263–273, 2020.

Kai Zhao (Member, IEEE) is an associate professor at the School of Communication and Information Engineering, Shanghai University. He previously conducted postdoctoral research at the University of California, Los Angeles, and earned his Ph.D. degree in computer science from Nankai University in 2020. He received his B.S. and M.S. degrees from Shanghai University in 2014 and 2017, respectively. His research interests include computer vision, geometry, and deep learning.

Wubang Yuan received his bachelor degree in electronic information engineering from the School of Communication and Information Engineering, Shanghai University in 2023. He is currently pursuing a master degree in signal and information processing in the same school. His research interests include computer vision, machine learning, and camouflaged object detection.

Zheng Wang received his B.E. degree in communication engineering from Shanghai University in 2022. He is currently working towards an M.E. degree in the Key Laboratory of Specialty Fiber Optics and Optical Access Networks and Shanghai Institute of Advanced Communication and Data Science, Shanghai University. His research interests include computer vision and visual generation.

Guanyi Li received his bachelor degree in computer science and technology from Zhengzhou University in 2018 and his master degree in computer technology from Zhengzhou University of Light Industry in 2022. He is currently pursuing his Ph.D. degree in the Key Laboratory of Specialty Fiber Optics and Optical Access Networks and Shanghai Institute of Advanced Communication and Data Science, Shanghai University. His research interests include computer vision and camouflaged object detection.

Xiaoqiang Zhu is an associate professor in the School of Communication and Information Engineering, Shanghai University. He received his Ph.D. degree from the State Key Laboratory of CAD&CG at Zhejiang University. In 2023, he was a visiting scholar at the National Centre for Computer Animation, Bournemouth University, UK. His research interests include computer graphics and intelligent information processing.

Deng-Ping Fan (Senior Member, IEEE) joined Nankai International Advanced Research Institute (SHENZHEN-FUTIAN) as a faculty member in 2024. Before he was a full professor and deputy director of the Media Computing Lab (MCLab) in the College of Computer Science, Nankai University, China, and earlier, he was a postdoctoral researcher with Prof. Luc Van Gool in the Computer Vision Lab at ETH Zurich. He was one of the core technical members of the TRACE-Zurich project on automated driving.

Dan Zeng (Senior Member, IEEE) received her B.S. degree in electronic science and technology and her Ph.D. degree in circuits and systems from the University of Science and Technology of China, Hefei. She is a full professor and the Dean of the Department of Communication Engineering, Shanghai University, and directs the Computer Vision and Pattern Recognition Laboratory. Her main research interests include computer vision, multimedia analysis, and machine learning. She currently serves as an associate editor for *IEEE Transactions on Multimedia* and *IEEE Transactions on Circuits and Systems for Video Technology*, as a TC Member for IEEE MSA, and as an associate TC Member for IEEE MMSP.