Iterative Refinement of Semantic and Spatial Representations for Open-Vocabulary Camouflaged Object Segmentation

Fangyan Wang[1], Ge Jiao[1*], and Guowen Yue[1]

College of Computer Science and Technology, Hengyang Normal University, Hengyang, Hunan, China
wfy@hynu.edu.cn, jiaoge@126.com, ygw@hynu.edu.cn

Abstract. Open-Vocabulary Camouflaged Object Segmentation aims to segment camouflaged objects from unseen categories. Existing two-stage methods typically treat category semantics as a one-time prior for segmentation, thereby lacking deep and iterative interaction between textual semantics and visual spatial structures. To address this limitation, we propose an iterative two-stage refinement framework based on semantic context and spatial structure. Specifically, we first design a spatial-structure-aware category re-ranker that leverages segmentation masks generated by SAM2 to reorder CLIP's candidate categories according to region-level visual consistency. Furthermore, we introduce a semantic-context-based segmentation modulator that injects the refined category information into SAM2, guiding SAM2 to produce more precise and discriminative segmentation results. Notably, the proposed re-ranker and modulator are jointly optimized in an iterative manner, forming a closed-loop refinement process that enables mutual guidance between semantic representations and spatial features. Through iterative updates, semantic context and spatial structure are progressively enhanced. Extensive experiments demonstrate that the proposed method significantly outperforms existing open-vocabulary methods and camouflaged object segmentation approaches. In particular, the proposed method achieves a 14.4% improvement in the cS_m metric compared with OVCoser.

Keywords: Open-Vocabulary Camouflaged Object Segmentation · Segment Anything Model 2 · Multi-modal · Iterative Refinement

1 Introduction

Camouflaged Object Segmentation (COS) aims to segment objects that are visually similar to their surrounding backgrounds in complex scenes. As a fundamental and challenging vision problem, COS serves as an essential component in scene understanding [1,2], autonomous driving [3], medical image analysis [4,5],

* Corresponding Author.

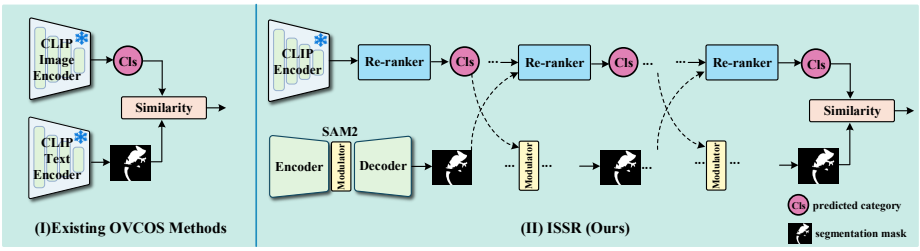

Fig. 1: Existing methods rely on a one-shot two-stage pipeline, limiting self-correction in COS scenarios(Fig. 1(I)). In contrast, the proposed ISSR framework treats category semantics and spatial masks as iteratively updated states rather than static outputs(Fig. 1(II)).

and intelligent robotics [6]. However, the success of existing COS methods largely relies on the closed-world assumption. In real-world environments, where novel categories continuously emerge, this assumption often breaks down, thereby limiting the scalability and generalization capability of models in practical applications.

To address this limitation, OVCoser [7] introduced the Open-Vocabulary Camouflaged Object Segmentation (OVCOS) task, along with a corresponding benchmark and dataset. By jointly leveraging multi-source information, OVCoser explores camouflaged object cues in the visual feature space. Subsequently, SuCLIP [8] proposed a context-aware prompting mechanism that enhances textual prompt representations using the implicit knowledge of visual encoders, while explicitly aligning textual semantics with local visual features to mitigate semantic ambiguity under camouflage conditions. Although these methods have facilitated the initial development of OVCOS, their overall performance remains relatively limited.

We observe that existing OVCOS approaches generally adopt a two-stage paradigm with one-shot semantic injection. Specifically, a vision-language model such as CLIP [9] is first employed for category-level semantic prediction, and the predicted category semantics are then treated as fixed priors for the downstream segmentation model to generate spatial masks, as illustrated in Fig. 1(I). This design implicitly assumes that category recognition and spatial localization can be effectively decoupled within a single forward pass, and that semantic predictions no longer require refinement once the segmentation stage begins. However, due to the high visual similarity between camouflaged objects and their backgrounds, there exists a strong coupling between category semantics and spatial localization. In such scenarios, even slight semantic deviations can be amplified during subsequent segmentation, leading to catastrophic prediction failures. More critically, existing methods generally lack a mechanism to leverage the spatial structural information formed during segmentation to refine semantic predictions in return. As a result, once early semantic errors occur, they are difficult to correct. This unidirectional, one-shot information flow severely

constrains the model's robustness and self-correction capability in complex camouflaged environments.

In light of the above analysis, we revisit the OVCOS task from the perspective of iterative learning. Specifically, we observe that although CLIP's Top-1 predictions are often unstable in camouflaged scenarios, the ground-truth category appears with significantly higher probability among the Top-N candidates. This observation indicates that semantic failure does not arise from semantic absence, but rather from unreliable one-shot ranking under insufficient spatial constraints. Furthermore, the spatial mask structures progressively formed during segmentation inherently contain discriminative cues for foreground-background separation, thereby providing valuable feedback for semantic refinement. These insights motivate a bidirectional semantic-spatial interactive iterative refinement framework, termed ISSR, as illustrated in Fig. 1(II). In our framework, category semantics and spatial masks are no longer treated as static outputs, but as dynamically updated state variables across iterations. At the t-th iteration, the current semantic representation guides the segmentation model to generate a spatial mask. The resulting spatial structural information, together with visual features, is then fed back to re-evaluate and re-rank semantic candidates, yielding a more reliable semantic representation for the $(t+1)$-th iteration. Through this closed-loop process, in which semantics guide spatial reasoning and spatial structures refine semantics, the model progressively corrects early semantic biases and continuously improves segmentation quality.

By unifying semantic prediction and spatial segmentation within an iterative refinement paradigm, we effectively overcome the inherent limitations of traditional one-shot semantic injection strategies under camouflaged conditions. The proposed iterative refinement mechanism endows the model with explicit semantic self-correction capability, thereby substantially enhancing the robustness and discriminative power of OVCOS in complex real-world scenarios. Our main contributions are summarized as follows:

- We propose a novel two-stage framework, termed ISSR, which introduces a re-ranker and a modulator to iteratively and jointly optimize semantic prediction and spatial segmentation, addressing the lack of deep bidirectional interaction between semantic guidance and structural representation in existing methods.
- To the best of our knowledge, this is the first OVCOS approach that explicitly models bidirectional iterative exchange between semantic and spatial representations.
- Extensive experiments demonstrate that ISSR exhibits stronger robustness and generalization ability in unseen-category scenarios. Compared with OVCoser, our method achieves a significant 14.4% improvement in terms of cS_m.

2 Related Work

2.1 Camouflaged Object Segmentation

COS aims to identify and segment objects that are visually indistinguishable from their surrounding environments, making it one of the most challenging tasks in computer vision. With the introduction of high-quality benchmarks such as CAMO [10], COD10K [11], and NC4K [12], substantial progress has been achieved, leading to a series of representative methods [13–21]. For example, ZoomNeXt [13] proposes a unified collaborative pyramid network for both static and dynamic COS. CamoDiffusion [16] formulates COS as conditional mask generation via diffusion modeling. VSCode [17] introduces two-dimensional prompt learning to enhance domain and task adaptability. EASE [19] leverages environmental prototype retrieval to assist segmentation. CFF-KDNet [21] employs cross-scale feature fusion with knowledge distillation to mitigate dataset imbalance. Despite these advances, existing methods are primarily developed under closed-set settings, relying on predefined categories and fixed data distributions. In open and dynamic real-world scenarios, where unseen categories and novel patterns continuously emerge, their performance degrades significantly, limiting generalization and scalability.

2.2 Open-Vocabulary Semantic Segmentation

Open-Vocabulary Semantic Segmentation (OVSS) aims to evaluate a model's ability to generalize to novel categories defined by free-form textual descriptions, beyond the closed-set assumption. Most existing approaches follow a two-stage paradigm. Methods such as SimSeg [22] generate mask proposals and rely on CLIP for open-vocabulary classification, while MaskCLIP [23] improves mask quality by integrating mask constraints into attention layers. CRTNet [24] further reformulates OVSS as a text generation problem, decoupling segmentation from predefined label spaces. To improve efficiency, single-stage methods have also emerged. ODISE [25] performs mask generation and classification jointly within a diffusion framework, and EOV-Seg [26] enhances spatial representation learning through dynamic embedding experts. Building upon OVSS, recent studies extend open-vocabulary learning to COS, forming the OVCOS task. OVCoser [7] establishes the first benchmark and baseline, while SuCLIP [8] alleviates semantic ambiguity via context-aware prompting. However, existing methods still suffer from semantic confusion and inaccurate spatial localization, and their performance remains far behind closed-set COS models.

To address these limitations, we revisit the core challenges of OVCOS and propose a bidirectional iterative framework consisting of a re-ranker and a modulator. By progressively coupling semantic reasoning with spatial refinement, our method mitigates semantic ambiguity and improves localization accuracy, substantially advancing OVCOS performance.

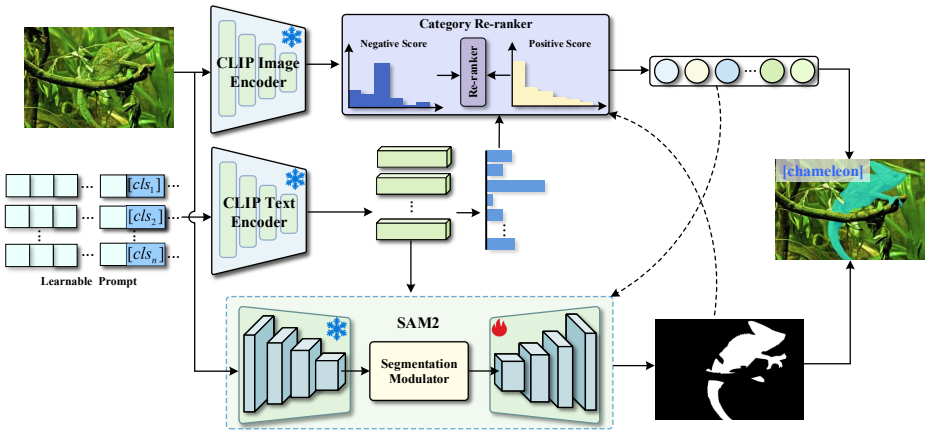

Fig. 2: Overall architecture of ISSR. Based on frozen CLIP and SAM2, it establishes a closed-loop interaction between a category re-ranker and a segmentation modulator, enabling progressive semantic refinement and spatial correction.

3 Proposed Method

3.1 Overview

The overall architecture of ISSR is illustrated in Fig. 2. The re-ranker leverages segmentation masks generated by Segment Anything Model 2 (SAM2) [27] as spatial priors to evaluate region-level semantic consistency and recalibrate category confidence scores. The updated scores guide the modulator to perform semantic-aware adjustment over multi-level visual features, enabling SAM2 to produce more accurate and semantically consistent masks. Through iterative refinement, the two modules form a closed-loop framework that progressively improves both category prediction and spatial localization. Furthermore, a learnable text prompting mechanism is introduced to adaptively reconstruct category embeddings, thereby alleviating semantic ambiguity in camouflaged scenes.

3.2 Category Re-ranker

In OVCOS, CLIP zero-shot similarities are typically used for direct Top-1 prediction. However, in camouflaged scenarios, strong foreground-background similarity often disturbs the ranking order. To address this issue, we propose a spatial structure-aware Top-N category re-ranker that corrects local ranking errors while preserving zero-shot generalization, as illustrated in Fig. 3. Given an input image I, we extract normalized visual features f using the frozen visual encoder of CLIP and compute class-level logits via similarity matching:

$$L = f^{\top} W, \tag{1}$$

where W denotes the normalized text embedding matrix. After applying the

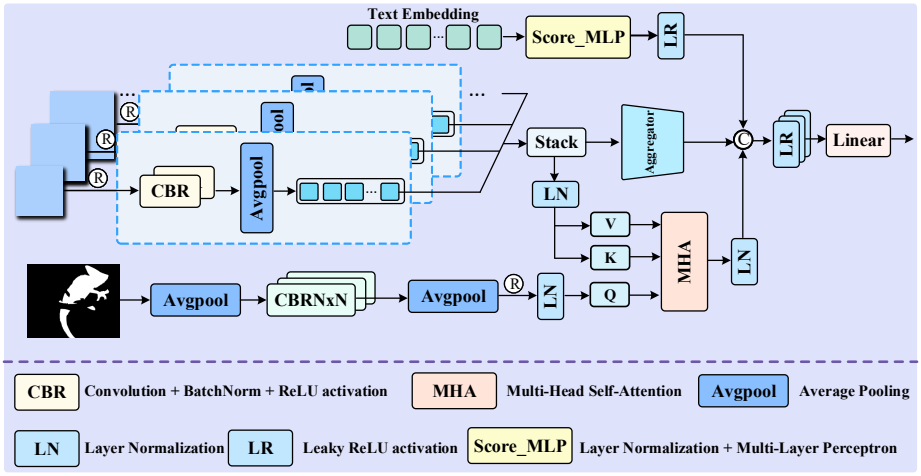

Fig. 3: Overall schematic diagram of the Category Re-ranker.

softmax operation to L, we obtain the category probability distribution and select the Top-N candidate categories $\{(c_i, s_i)\}_{i=1}^{N}$, where c_i and s_i denote the category index and confidence score of the i-th candidate, respectively.

Relying solely on CLIP's global visual representation is insufficient to capture fine-grained differences among candidate categories in complex camouflaged scenes. Therefore, we further introduce intermediate visual features from multiple layers of the CLIP visual encoder. After spatial pooling, we obtain compact representations $v^{(m)} \in \mathbb{R}^{1\times C}$, $m \in \{12, 16, 18\}$, to provide complementary local and mid-level semantic cues. For each candidate category c_i in Top-N, we construct a low-dimensional ranking feature vector:

$$r_i = \begin{bmatrix} s_i \\ \frac{i}{N} \\ s_i - s_1 \end{bmatrix}, \tag{2}$$

where s_i represents the absolute confidence score, $\frac{i}{N}$ encodes its relative ranking position within Top-N, and $s_i - s_1$ measures the score gap with respect to the Top-1 candidate, thereby explicitly modeling inter-candidate relationships.

Based on the ranking cue r_i, the hierarchically fused CLIP features $\{v^{(m)}\}$, and the spatial mask M, we design a category re-ranker $\mathcal{R}(\cdot)$ to produce the refined ranking score $\widehat{S}_i$:

$$\widehat{S}_i = \mathcal{R}\Big(r_i,\, \{v^{(m)}\},\, M\Big)\,. \tag{3}$$

The mask M guides the model to focus on potential foreground regions, alleviating background interference introduced by global pooling. During training, ground-truth masks are used, while at inference we adopt masks generated by SAM2 for annotation-free deployment. The re-ranker is optimized using a

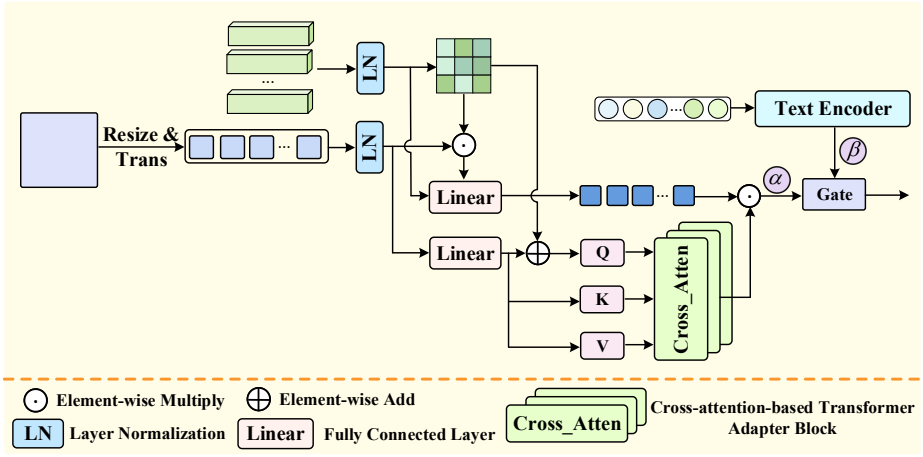

Fig. 4: Overall schematic diagram of the Segmentation Modulator.

margin-based pairwise ranking loss within the Top-N candidate set:

$$\mathcal{L}_{\text{cls}} = \frac{1}{|\mathcal{P}|} \sum_{(p,n)\in\mathcal{P}} \max\left(0,\, 1 - \widehat{s}_p + \widehat{s}_n\right), \tag{4}$$

where $\widehat{s}_p$ and $\widehat{s}_n$ denote the scores of the positive and negative categories, respectively. Restricting optimization to Top-N candidates reduces noise from irrelevant global negatives and stabilizes ranking refinement.

3.3 Segmentation Modulator

In OVCOS, semantic guidance plays two complementary roles: fine-grained cross-modal semantics localize concept-consistent regions, while category-level priors impose global constraints to suppress ambiguous responses. However, their relative importance varies across scenes, and fixed fusion strategies often cause either over-reliance on priors or insufficient semantic supervision. To address this issue, we propose a learnable semantic-allocation segmentation modulator inserted between the SAM2 image encoder and mask decoder, as shown in Fig. 4.

The modulator takes intermediate visual features F_{sam} from the SAM2 encoder and textual features F_{clip} from the CLIP text encoder as input. To reduce cross-modal discrepancy, modality-specific gating functions are applied:

$$V_i = \mathrm{G}_{\text{sam}}(F_{sam}), \quad S_t = \mathrm{G}_{\text{clip}}(F_{clip}), \tag{5}$$

where V_i and S_t denote aligned visual and semantic features, respectively. To model fine-grained semantic-spatial correspondence, the aligned features are first fused via element-wise addition:

$$F_{fus} = V_i + S_t, \tag{6}$$

and further modulated using a multi-head attention mechanism:

$$F_{att} = \mathrm{MHA}\left(Q = F_{fus},\ K = V_i,\ V = V_i\right). \tag{7}$$

This attention selectively reweights spatial features according to semantic cues. To further enforce category-level consistency, we compute a cross-modal relevance map:

$$R = V_i \odot S_t, \tag{8}$$

and use it to modulate the attention-enhanced representation:

$$F_{rel} = F_{att} \odot R. \tag{9}$$

Meanwhile, the Top-1 category predicted by the structure-aware re-ranker is encoded by the CLIP text encoder and reshaped into a category-level semantic vector:

$$c_{top} = \mathrm{Flatten}\big(\mathrm{CLIP}(\text{Top-1})\big), \tag{10}$$

The category-level semantics are then used to perform a final conditional modulation:

$$F_{cat} = F_{rel} \odot (1 + c_{top}). \tag{11}$$

Instead of manually balancing fine-grained semantic-spatial interaction and category-level semantic constraints, we introduce a learnable semantic allocation mechanism. A scalar parameter s is optimized during training and normalized via a sigmoid function:

$$\lambda = \sigma(s), \quad \alpha = \lambda, \quad \beta = 1 - \lambda. \tag{12}$$

Finally, the outputs of semantic-spatial attention interaction and category-aware refinement are fused via weighted summation:

$$F_{mod} = \alpha \cdot F_{att} + \beta \cdot F_{cat}, \tag{13}$$

which is fed into the SAM2 mask decoder to produce semantically consistent and spatially precise segmentation results.

3.4 Learnable Prompt

Handcrafted prompts are prone to human bias and limited adaptability, restricting generalization in open-vocabulary and complex camouflage scenarios. To address this limitation, we adopt a learnable continuous prompt representation that replaces discrete templates. Let C_B denote the set of base categories. For any category $c \in C_B$, its prompt representation is defined as:

$$P_c = [p_1, p_2, \ldots, p_l, cls_c], \tag{14}$$

where $\{p_i\}_{i=1}^{l}$ denotes a sequence of l learnable context vectors, and cls_c represents the fixed class-name embedding of category c, obtained from the token

embedding layer of CLIP. All context vectors p_i are randomly initialized with the same dimensionality as the CLIP text embeddings and are optimized via backpropagation during training.

The context prompts are shared across categories. For a novel category, only the corresponding class-name embedding is replaced, naturally supporting the open-vocabulary setting.

3.5 Iterative Refinement

To formalize the proposed iterative mechanism, we formulate the inference process of OVCOS as a semantic-spatial alternating optimization process. Given an input image I and its Top-N candidate set $\mathcal{C} = \{c_1, \ldots, c_N\}$, we define the semantic state at iteration t as $s^{(t)} \in \mathbb{R}^d$ and the spatial state as $M^{(t)} \in [0,1]^{H \times W}$. The semantic state is initialized using CLIP by aggregating Top-N text embeddings weighted by confidence scores:

$$s^{(0)} = \sum_{i=1}^{N} p_i \, e(c_i), \quad p_i = \mathrm{softmax}(\mathrm{CLIP}(I, c_i)), \tag{15}$$

where $e(c_i)$ denotes the text embedding of category c_i. At iteration t, the semantic state is updated via a category re-ranker $\mathcal{R}(\cdot)$:

$$s^{(t+1)} = \mathcal{R}\big(\{e(c_i)\}_{i=1}^{N},\, f_I,\, M^{(t)}\big), \tag{16}$$

where f_I represents CLIP visual features. The re-ranker refines semantic representation under spatial constraints.

The refined semantic state is then injected into a semantically modulated segmentation network, modeled as a spatial update operator $\mathcal{S}(\cdot)$:

$$M^{(t+1)} = \mathcal{S}\big(I,\, s^{(t+1)}\big). \tag{17}$$

The re-ranker and modulator form an explicit iterative optimization loop, which can be summarized as:

$$\begin{cases} s^{(t+1)} = \mathcal{R}(\mathcal{C}, f_I, M^{(t)}), \\ M^{(t+1)} = \mathcal{S}(I, s^{(t+1)}). \end{cases} \tag{18}$$

This alternating update progressively refines both semantic prediction and spatial segmentation in a closed-loop manner.

3.6 Loss Function

For the segmentation modulator, we simultaneously introduce the weighted binary cross-entropy loss [28] ($\mathcal{L}^{\omega}_{BCE}$), the weighted intersection-over-union loss [29] ($\mathcal{L}^{\omega}_{IoU}$), and the Dice loss [30] ($\mathcal{L}_{dice}$) to constrain the quality of the predicted masks from multiple perspectives. In addition, the category re-ranker is supervised by a margin-based pairwise ranking loss ($\mathcal{L}_{cls}$). The overall training objective is formulated as follows:

$$\mathcal{L}_{\text{total}} = \mathcal{L}^{\omega}_{BCE} + \mathcal{L}^{\omega}_{IoU} + \mathcal{L}_{dice} + \mathcal{L}_{cls}. \tag{19}$$

Table 1: Quantitative comparison on the OVCamo dataset under different training settings. The best performance is highlighted in bold.

Model	VLM	Text Prompt	$cS_m \uparrow$	$cF_\beta^w \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
Test on OVCamo with weights trained on COCO								
SimSeg [22]	CLIP-ViT-B/16 [31]	[32]	0.128	0.105	0.838	0.112	0.143	0.094
OVSeg [33]	CLIP-ViT-L/14 [31]	[34]	0.341	0.306	0.584	0.325	0.384	0.273
ODISE [25]	CLIP-ViT-L/14 [31]	[35]	0.409	0.339	0.500	0.341	0.421	0.302
SAN [36]	CLIP-ViT-L/14 [31]	[34]	0.414	0.343	0.489	0.357	0.456	0.319
CAT-Seg [37]	CLIP-ViT-L/14 [31]	[9]	0.430	0.344	0.448	0.366	0.459	0.310
FC-CLIP [38]	CLIP-ConvNeXt-L [39]	[34]	0.374	0.306	0.539	0.320	0.409	0.285
Finetune on OVCamo with weights trained on COCO								
SimSeg [22]	CLIP-ViT-B/16 [31]	[32]	0.098	0.071	0.852	0.081	0.128	0.066
OVSeg [33]	CLIP-ViT-L/14 [31]	[34]	0.164	0.131	0.763	0.147	0.208	0.123
ODISE [25]	CLIP-ViT-L/14 [31]	[35]	0.182	0.125	0.691	0.219	0.309	0.189
SAN [36]	CLIP-ViT-L/14 [31]	[34]	0.321	0.216	0.550	0.236	0.331	0.204
CAT-Seg [37]	CLIP-ViT-L/14 [31]	[9]	0.185	0.094	0.702	0.110	0.185	0.088
FC-CLIP [38]	CLIP-ConvNeXt-L [39]	[34]	0.124	0.074	0.798	0.088	0.162	0.072
Train on OVCamo								
SimSeg [22]	CLIP-ViT-B/16 [31]	[32]	0.053	0.049	0.921	0.056	0.098	0.047
OVSeg [33]	CLIP-ViT-L/14 [31]	[34]	0.024	0.046	0.954	0.056	0.130	0.046
ODISE [25]	CLIP-ViT-L/14 [31]	[35]	0.187	0.119	0.700	0.211	0.298	0.167
SAN [36]	CLIP-ViT-L/14 [31]	[34]	0.275	0.202	0.612	0.220	0.318	0.189
CAT-Seg [37]	CLIP-ViT-L/14 [31]	[9]	0.181	0.106	0.719	0.123	0.196	0.094
FC-CLIP [38]	CLIP-ConvNeXt-L [39]	[34]	0.080	0.076	0.872	0.090	0.191	0.072
OVCoser [7]	CLIP-ConvNeXt-L [39]	CamoPrompts [7]	0.579	0.490	0.336	0.520	0.616	0.443
SuCLIP [8]	CLIP-ConvNeXt-L [39]	CAP [8]	0.667	0.594	0.242	0.633	0.722	0.540
Ours	**CLIP-ViT-L/14** [31]	**Learnable**	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**

4 Experiments

4.1 Experimental Settings

Datasets and Metrics. We evaluate the proposed method on both the OVCOS and conventional COS tasks. For OVCOS task, experiments are conducted on the OVCamo [7] dataset. Consistent with prior works [7, 8, 40, 41], we adopt six OVCOS metrics: cS_m, cF_β^ω, $cMAE$, cF_β, cE_m, and $cIoU$. For the conventional COS task, we evaluate on three widely used benchmarks: CAMO [10], COD10K [11], and NC4K [12]. For COS evaluation, we report four commonly used metrics: Structure Measure [42] (S_m), Enhanced Alignment Measure [43] (E_m), weighted F-measure [44] (F_β^ω), and Mean Absolute Error [45] (MAE).

Implementation Details. All experiments are implemented in PyTorch and conducted on a single NVIDIA RTX 3090 GPU. The CLIP encoders and SAM2 encoder are frozen during training, while the remaining trainable parameters are randomly initialized and optimized end-to-end. We use the AdamW optimizer with an initial learning rate of 1×10^{-6} and weight decay of 5×10^{-4}. The batch size is set to 4, and input images are resized to 384×384.

4.2 Comparisons with State-of-the-Art Methods

Quantitative Evaluation. We compare ISSR with representative OVSS and OVCOS approaches. As shown in Table 1, ISSR achieves consistent improve-

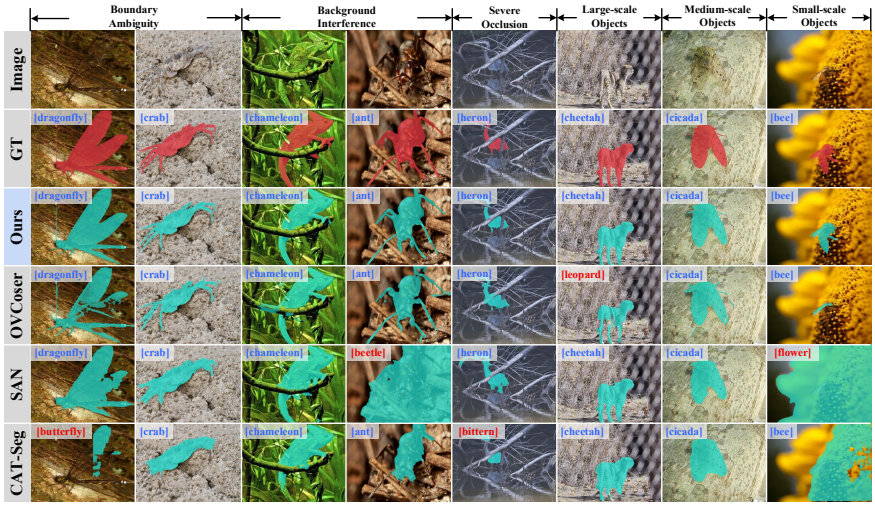

Fig. 5: Qualitative visualization results on the OVCamo dataset. Blue text denotes correctly predicted categories, while red text denotes incorrectly predicted categories.

Table 2: Quantitative comparison between the proposed method and 9 state-of-the-art methods on the COS task. The best performance is highlighted in bold.

Method	CAMO (250)				COD10K (2,026)				NC4K (4,121)			
	$S_m \uparrow$	$F_\beta^w \uparrow$	$E_m \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$E_m \uparrow$	$MAE \downarrow$	$S_m \uparrow$	$F_\beta^w \uparrow$	$E_m \uparrow$	$MAE \downarrow$
SINet [11]	0.745	0.712	0.804	0.092	0.776	0.667	0.864	0.043	0.808	0.768	0.871	0.058
ACUMEN [46]	0.886	0.850	0.939	**0.039**	0.852	0.761	0.930	0.026	0.874	0.826	0.932	0.036
VSCode [17]	0.873	0.820	0.925	0.046	0.869	0.780	0.931	0.025	0.882	0.841	0.935	0.032
CamoFocus [47]	0.873	0.842	0.926	0.044	0.873	0.802	0.935	0.022	0.889	0.853	0.936	0.030
ZoomNeXt [13]	0.874	0.839	0.931	0.047	0.887	0.818	0.948	**0.019**	0.892	0.852	0.943	0.030
RUN [48]	0.877	-	0.934	0.045	0.878	-	0.941	0.021	0.892	-	0.940	0.030
SAM2-UNet [49]	0.884	0.861	0.932	0.042	0.880	0.789	0.936	0.021	0.901	0.863	0.941	0.029
CamoDiffusion [16]	0.878	0.853	0.940	0.042	0.881	0.814	0.944	0.020	0.893	0.859	0.942	0.029
CFF-KDNet-P2 [21]	0.870	0.833	0.920	0.048	0.889	0.821	0.937	0.021	0.897	0.854	0.936	0.030
Ours	**0.896**	**0.870**	**0.945**	**0.039**	**0.898**	**0.832**	**0.953**	**0.019**	**0.905**	**0.883**	**0.959**	**0.028**

ments across all settings. Compared with OVCoser [7], ISSR improves cS_m, cF_β^ω, cF_β, cE_m, and $cIoU$ by 14.4%, 17.3%, 17.3%, 14.6%, and 16.8%, respectively, while reducing $cMAE$ by 12.5%. Compared with the state-of-the-art SuCLIP [8], ISSR still achieves gains of 5.6%, 6.9%, 6.0%, 4.0%, and 7.1% on the above metrics, respectively, while reducing $cMAE$ by 3.1%. The results indicate that the proposed iterative re-ranking and modulation mechanism effectively captures semantic variability in camouflaged scenes, leading to consistent improvements across all evaluation metrics.

Qualitative Evaluation. Fig. 5 presents visual comparisons between ISSR and recent state-of-the-art methods under diverse challenging scenarios, including blurred boundaries (Cols. 1–2), background interference (Cols. 3–4), severe occlusion (Col. 5), and multi-scale objects (Cols. 6–8). CAT-Seg [37] and SAN [36]

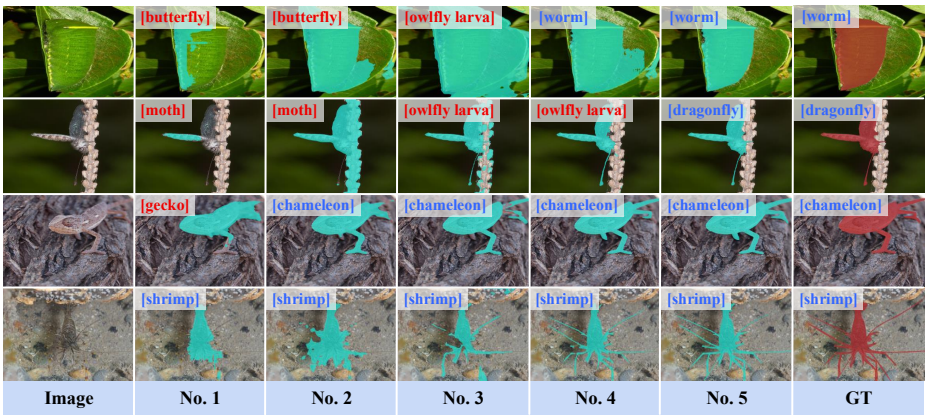

Fig. 6: Visualization results of different components, corresponding to the ablation study settings in Table 3. Blue text denotes correctly predicted categories, while red text denotes incorrectly predicted categories.

Table 3: Ablation experiments of ISSR on the OVCamo dataset. R: category re-ranker; M: segmentation modulator; L: learnable prompt; I: iterative refinement. The best performance is highlighted in bold.

No.	Module	OVCamo					
		$cS_m \uparrow$	$cF_\beta^\omega \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
1	Baseline	0.658	0.615	0.268	0.620	0.711	0.534
2	$+R$	0.678	0.634	0.256	0.645	0.730	0.557
3	$+R, M$	0.697	0.642	0.242	0.660	0.743	0.576
4	$+R, M, L$	0.711	0.651	0.230	0.679	0.749	0.599
5	$\boldsymbol{+R, M, L, I}$	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**

often exhibit inaccurate localization and incorrect category predictions under blurred boundaries and complex backgrounds, leading to semantic confusion. OVCoser fails to recover complete object contours in occluded or small-scale cases, resulting in incomplete masks. In contrast, ISSR produces more accurate localization and more complete object structures across all scenarios.

More Comparisons. We further evaluate the proposed method on the conventional COS task. As shown in Table 2, compared with SAM2-UNet [49], ISSR improves the F_β^ω and E_m metrics by 0.9% and 1.3% on CAMO, 4.3% and 1.7% on COD10K, and 2.0% and 1.8% on NC4K, respectively. In comparison with CamoDiffusion [16], the proposed method consistently improves the F_β^ω metric by 1.7%, 1.8%, and 2.4% on CAMO, COD10K, and NC4K, respectively. Furthermore, compared with the recently proposed CFF-KDNet-P2 [21], our method achieves improvements of 3.7% and 2.5% in terms of F_β^ω and E_m on CAMO, 1.1% and 1.6% on COD10K, and 2.9% and 2.3% on NC4K, respectively.

Table 4: Ablation study on the learnable prompt context length l on the OVCamo dataset. The best performance is highlighted in bold.

l	OVCamo					
	$cS_m \uparrow$	$cF_\beta^w \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
4	0.714	0.653	0.223	0.683	0.752	0.600
8	0.718	0.657	0.218	0.689	0.756	0.607
16	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**
32	0.717	0.657	0.217	0.688	0.754	0.605
64	0.713	0.651	0.221	0.683	0.751	0.601

Table 5: Ablation study of the Top-N candidate set size in the re-ranker on the OVCamo dataset. The best performance is highlighted in bold.

Top-N	OVCamo					
	$cS_m \uparrow$	$cF_\beta^w \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
2	0.703	0.649	0.227	0.675	0.744	0.594
4	0.707	0.651	0.225	0.678	0.749	0.597
6	0.712	0.654	0.221	0.681	0.753	0.603
8	0.717	0.657	0.219	0.687	0.757	0.608
10	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**
12	0.719	0.655	0.221	0.677	0.754	0.607
14	0.714	0.652	0.225	0.679	0.751	0.604

4.3 Ablation Study

Effectiveness of Proposed Modules. We conduct ablation studies by progressively incorporating each module into the baseline model, which consists of a CLIP-based vision-language encoder and SAM2, where the default CLIP Top-1 prediction is used to guide SAM2. The results are reported in Table 3. Introducing the category re-ranker (No. 2) improves cS_m, cF_β^ω, cF_β, cE_m, and $cIoU$ by 2.0%, 1.9%, 2.5%, 1.9%, and 2.3%, respectively, while reducing $cMAE$ by 1.2%, demonstrating its effectiveness in alleviating semantic ambiguity. Further incorporating the segmentation modulator (No. 3) yields additional gains of 1.9%, 0.8%, 1.5%, 1.3%, and 1.9% on these metrics, respectively, together with a 1.4% reduction in $cMAE$, indicating enhanced semantic-spatial alignment. Adding the learnable prompt mechanism (No. 4) further improves performance across all evaluation metrics, showing the benefit of adapting semantic representations to the open-vocabulary setting. Finally, incorporating the iterative refinement strategy (No. 5) achieves the best overall performance, yielding consistent gains across all metrics and validating the effectiveness of jointly refining semantic prediction and spatial segmentation. Qualitative comparisons under different configurations are shown in Fig. 6, where progressively integrating the proposed modules results in increasingly accurate localization and category prediction.

Impact of Learnable Prompt Context Length. We investigate the effect of the learnable prompt context length l on model performance, where

Table 6: Ablation study on the number of iterative refinement steps t in ISSR on the OVCamo dataset. The best performance is highlighted in bold.

t	OVCamo					
	$cS_m \uparrow$	$cF_\beta^w \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
1	0.714	0.650	0.231	0.679	0.751	0.597
2	0.715	0.651	0.224	0.682	0.755	0.603
3	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**
4	0.717	0.654	0.223	0.684	0.751	0.599
5	0.711	0.651	0.233	0.680	0.747	0.586

Table 7: Ablation study on different input composition schemes of the re-ranker on the OVCamo dataset. S: ranking scores; I: image visual features; M: mask features. The best performance is highlighted in bold.

Input	OVCamo					
	$cS_m \uparrow$	$cF_\beta^w \uparrow$	$cMAE \downarrow$	$cF_\beta \uparrow$	$cE_m \uparrow$	$cIoU \uparrow$
S	0.711	0.650	0.221	0.682	0.753	0.602
S, I	0.715	0.654	0.219	0.686	0.758	0.606
S, M	0.717	0.656	0.218	0.687	0.757	0.605
$\boldsymbol{S, I, M}$	**0.723**	**0.663**	**0.211**	**0.693**	**0.762**	**0.611**

$l \in \{4, 8, 16, 32, 64\}$. As shown in Table 4, shorter prompt contexts with l set to 4 or 8 provide limited semantic capacity, making it difficult to encode sufficient task-specific contextual information. In contrast, overly long prompt contexts with l set to 32 or 64 introduce redundant learnable tokens and increase optimization difficulty. The best performance is achieved when l is set to 16, indicating that a moderate context length provides a favorable balance between semantic expressiveness and training stability.

Top-N Selection Strategy of the Re-ranker. We analyze the impact of the candidate set size N on re-ranking performance, where $N \in \{2, 4, 6, 8, 10, 12, 14\}$. As reported in Table 5, smaller candidate sets with N set to 2 or 4 may exclude the ground-truth category from the candidate set, thereby limiting re-ranking effectiveness. Increasing N improves coverage and progressively enhances performance. The best results are achieved at $N = 10$, suggesting a favorable balance between candidate diversity and noise suppression. However, further increasing N to 12 or 14 leads to slight performance degradation. An excessively large candidate set introduces numerous low-relevance or semantically mismatched categories, increasing the difficulty of re-ranking and weakening the discriminative power between positive and negative samples. Thus, $N = 10$ is adopted in subsequent experiments.

Effectiveness of Iterative Refinement. We further evaluate the impact of the iteration number t in the re-ranker–modulator loop, where $t \in \{1, 2, 3, 4, 5\}$ under identical training settings. The results are presented in Table 6. Performance improves steadily as t increases from 1 to 3, reaching the optimum at

$t = 3$. This observation demonstrates that multi-round semantic–spatial interaction effectively refines both category prediction and segmentation quality. However, further increasing t to 4 or 5 leads to slight performance degradation due to error accumulation and reduced semantic diversity across excessive iterations. Accordingly, $t = 3$ is selected as the default configuration.

Evaluation of Input Composition Schemes of the Re-ranker. We analyze the impact of different input compositions for the re-ranker, and the results are shown in Table 7. Using only ranking scores (Row 1) yields limited improvement, indicating that ranking priors alone are insufficient for reliable re-ranking. Incorporating visual features (Row 2) or mask features (Row 3) leads to consistent performance gains, demonstrating the benefit of semantic and spatial cues. The best performance is achieved when ranking scores, visual features, and mask features are jointly used (Row 4).

5 Conclusion

This paper presents ISSR, a two-stage iterative refinement framework for OV-COS. By establishing a bidirectional closed-loop interaction between category re-ranking and mask modulation, ISSR jointly addresses semantic ambiguity and spatial structural uncertainty. Through iterative semantic-spatial refinement, category-level semantics and pixel-level structural cues progressively enhance each other, leading to more reliable mask-label alignment. Extensive experiments on multiple COS benchmarks demonstrate that ISSR consistently outperforms existing OVSS methods and dedicated COS approaches, and exhibits strong robustness and generalization to unseen categories. These results highlight the importance of treating semantic guidance as a dynamic process refined by spatial feedback, rather than as a fixed prior.

Despite the significant improvements achieved by ISSR, its extension to multi-category scenarios remains worthy of further exploration. In particular, ISSR still has limitations in handling images containing multiple different categories, where more explicit instance decomposition and multi-label semantic reasoning are required to distinguish category-specific regions.

Acknowledgements

This work was supported by the Postgraduate Scientific Research Innovation Project of Hunan Province (CX20251738), the Open Fund of Hunan Engineering Research Center for Cyberspace Security Technology and Application at Hengyang Normal University (2025HSKFJJ032), the Hengyang City Science and Technology Innovation Program Project (202550037822), the Hunan Provincial University Key Laboratory of Artificial Intelligence Application and System Security (Document No. Xiangjiaotong [2025] 233), the Science and Technology Plan Project of Hunan Province (2016TP1020), and the 14th Five-Year Plan Key Disciplines and Application- oriented Special Disciplines of Hunan Province (Document No. Xiangjiaotong [2022] 351).

References

1. Qian Yu, Xiaoqi Zhao, Youwei Pang, Lihe Zhang, and Huchuan Lu. Multi-view aggregation network for dichotomous image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3921–3930, 2024.
2. Jiacheng Ruan, Wenzhen Yuan, Zehao Lin, Ning Liao, Zhiyu Li, Feiyu Xiong, Ting Liu, and Yuzhuo Fu. Mm-camobj: A comprehensive multimodal dataset for camouflaged object scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6740–6748, 2025.
3. Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020.
4. Feng Li, Zetao Huang, Lu Zhou, Haixia Peng, and Yimin Chu. Semi-supervised spatial-temporal calibration and semantic refinement network for video polyp segmentation. *Biomedical Signal Processing and Control*, 100:107127, 2025.
5. Hao Shao, Yang Zhang, and Qibin Hou. Polyper: Boundary sensitive polyp segmentation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 4731–4739, 2024.
6. Minyoung Hwang, Jaeyeon Jeong, Minsoo Kim, Yoonseon Oh, and Songhwai Oh. Meta-explore: Exploratory hierarchical vision-and-language navigation using scene object spectrum grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6683–6693, 2023.
7. Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *European Conference on Computer Vision*, pages 476–495. Springer, 2024.
8. Peng Ren, Tian Bai, Jing Sun, and Fuming Sun. Seeing the unseen: A semantic alignment and context-aware prompt framework for open-vocabulary camouflaged object segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23657–23666, 2025.
9. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.
10. Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.*, 184:45–56, 2019.
11. Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 2774–2784. Computer Vision Foundation / IEEE, 2020.
12. Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 11591–11601. Computer Vision Foundation / IEEE, 2021.

13. Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom-next: A unified collaborative pyramid network for camouflaged object detection. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):9205–9220, 2024.
14. Yitong Liu, Jindong Zhang, Yiming Wang, and Jingyi Jin. Camossr: A unified framework with energy-based scanning and cognitive reasoning for camouflaged object detection. *Neurocomputing*, 663:131977, 2026.
15. Xinyu Yan, Meijun Sun, Yahong Han, and Zheng Wang. Camouflaged object segmentation based on matching–recognition–refinement network. *IEEE Transactions on Neural Networks and Learning Systems*, 35(11):15993–16007, 2023.
16. Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 1272–1280, 2024.
17. Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. Vscode: General visual salient and camouflaged object detection with 2d prompt learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 17169–17180, 2024.
18. Xu Li, Xiaosheng Yu, and Peng Chen. Dsanet: Deep surrounding-aware network for camouflaged object detection via cross-refinement mirror strategy. *Expert Syst. Appl.*, 298:129605, 2026.
19. Ji Du, Fangwei Hao, Mingyang Yu, Desheng Kong, Jiesheng Wu, Bin Wang, Jing Xu, and Ping Li. Shift the lens: Environment-aware unsupervised camouflaged object detection. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19271–19282, 2025.
20. Baoyu Wang, Mao Yang, Pingping Cao, and Aihong Shen. Dwgnet: A dual-path wavelet guidance framework for salient object detection. *Inf. Sci.*, 724:122699, 2026.
21. Bo Cai, Houjie Li, Yanping Yang, and Jin Yan. Cff-kdnet: Cross-scale feature fusion network with knowledge distillation for camouflaged object detection. *Expert Syst. Appl.*, 299:130209, 2026.
22. Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for open-vocabulary semantic segmentation with pre-trained vision-language model. In *European Conference on Computer Vision*, pages 736–753. Springer, 2022.
23. JeongYeon Nam, Jinbae Im, Wonjae Kim, and Taeho Kil. Extract free dense misalignment from clip. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6173–6181, 2025.
24. Jiannan Ge, Lingxi Xie, Hongtao Xie, Pandeng Li, Sun-Ao Liu, Xiaopeng Zhang, Qi Tian, and Yongdong Zhang. Clip-adapted region-to-text learning for generative open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 24034–24044, 2025.
25. Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-vocabulary panoptic segmentation with text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2955–2966, 2023.
26. Hongwei Niu, Jie Hu, Jianghang Lin, Guannan Jiang, and Shengchuan Zhang. Eov-seg: Efficient open-vocabulary panoptic segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 6254–6262, 2025.
27. Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloé Rolland, Laura Gustafson, Eric

Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross B. Girshick, Piotr Dollár, and Christoph Feichtenhofer. SAM 2: Segment anything in images and videos. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025.

28. Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(10):6024–6042, 2021.
29. Jun Wei, Shuhui Wang, and Qingming Huang. F^3net: fusion, feedback and focus for salient object detection. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 12321–12328, 2020.
30. Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XIII*, volume 12358 of *Lecture Notes in Computer Science*, pages 696–711. Springer, 2020.
31. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.
32. Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. Learning to prompt for vision-language models. *Int. J. Comput. Vis.*, 130(9):2337–2348, 2022.
33. Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023.
34. Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.
35. Golnaz Ghiasi, Xiuye Gu, Yin Cui, and Tsung-Yi Lin. Scaling open-vocabulary image segmentation with image-level labels. In Shai Avidan, Gabriel J. Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 540–557. Springer, 2022.
36. Mengde Xu, Zheng Zhang, Fangyun Wei, Han Hu, and Xiang Bai. Side adapter network for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2945–2954, 2023.
37. Seokju Cho, Heeseong Shin, Sunghwan Hong, Anurag Arnab, Paul Hongsuck Seo, and Seungryong Kim. Cat-seg: Cost aggregation for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4113–4123, 2024.
38. Qihang Yu, Ju He, Xueqing Deng, Xiaohui Shen, and Liang-Chieh Chen. Convolutions die hard: Open-vocabulary segmentation with single frozen convolutional clip. *Advances in Neural Information Processing Systems*, 36:32215–32234, 2023.
39. Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings*

of the IEEE/CVF conference on computer vision and pattern recognition, pages 2818–2829, 2023.

40. Qi Chen, Lingxiao Yang, Yun Chen, Nailong Zhao, Jianhuang Lai, Jie Shao, and Xiaohua Xie. Training-free class purification for open-vocabulary semantic segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23124–23134, 2025.
41. Jiayun Luo, Siddhesh Khandelwal, Leonid Sigal, and Boyang Li. Emergent open-vocabulary semantic segmentation from off-the-shelf vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4029–4040, 2024.
42. Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, pages 4548–4557, 2017.
43. Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 698–704. ijcai.org, 2018.
44. Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 248–255, 2014.
45. Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, pages 733–740. IEEE Computer Society, 2012.
46. Hong Zhang, Yixuan Lyu, Qian Yu, Hanyang Liu, Huimin Ma, Ding Yuan, and Yifan Yang. Unlocking attributes' contribution to successful camouflage: A combined textual and visual analysis strategy. In *European Conference on Computer Vision*, pages 315–331. Springer, 2024.
47. Abbas Khan, Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El Saddik, Giulia De Masi, and Fakhri Karray. Camofocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement. In *Proceedings of the IEEE/cvf winter conference on applications of computer vision*, pages 1434–1443, 2024.
48. Chunming He, Rihan Zhang, Fengyang Xiao, Chengyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. RUN: reversible unfolding network for concealed object segmentation. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu, editors, *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, volume 267 of *Proceedings of Machine Learning Research*. PMLR / OpenReview.net, 2025.
49. Xinyu Xiong, Zihuang Wu, Shuangyi Tan, Wenxue Li, Feilong Tang, Ying Chen, Siying Li, Jie Ma, and Guanbin Li. Sam2-unet: Segment anything 2 makes strong encoder for natural and medical image segmentation. *Visual Intelligence*, 4(1):2, 2026.