# From Language to Segmentation: Collaborative Category-Guided Unsupervised Camouflaged Object Detection with SAM3

**Huafeng Chen**[1] , **Yueming Lyu**[1,*] , **Caifeng Shan**[1,*]
[1]School of Intelligence Science and Technology, Nanjing University
{huafengchen@smail., ymlv@}nju.edu.cn, caifeng.shan@gmail.com

## Abstract

Camouflaged Object Detection (COD) aims to segment objects that are hidden within complex backgrounds. Due to the low visual contrast of camouflaged objects, annotations are costly, motivating unsupervised COD (UCOD) to eliminate labeling expenses. Most UCOD methods follow the "MLLMs + other foundation models + SAM" paradigm, which relies on spatial interactions that are unreliable in camouflaged scenarios, leading to fundamental performance bottlenecks. In this paper, we propose a novel UCOD framework that leverages SAM3 through category-level interaction with MLLMs, bypassing unreliable spatial prompts. To address SAM3's sensitivity to category granularity, we introduce Fine-grained Category Query, guiding MLLMs to generate full-granularity category chains for robust category prompting. To mitigate suboptimal segmentation caused by high camouflage and background confusion, we propose Semantic–Geometric Dual Confirmation, which jointly validates segmentation masks from semantic and spatial perspectives. Furthermore, we introduce Semantic–Geometric Reasoning Injection, which injects critical semantic and geometric cues into MLLMs to refine category reasoning and progressively correct segmentation errors under extreme camouflage or MLLM hallucinations. Extensive experiments show that our method significantly outperforms existing UCOD approaches and achieves performance comparable to weakly supervised COD.



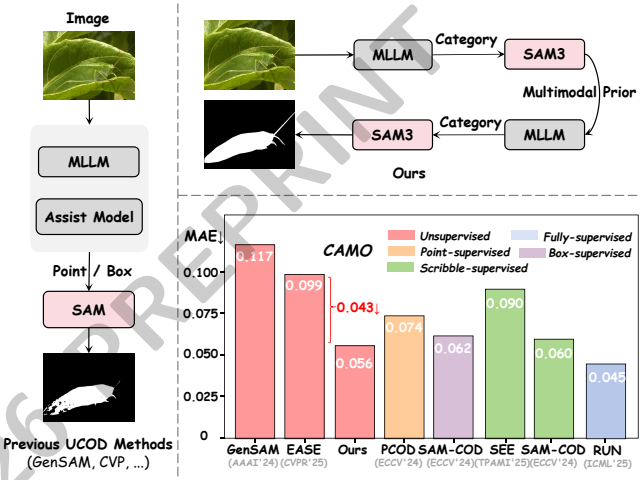


Figure 1: **Left:** Pipeline of traditional training-free UCOD methods. **Top-right:** Pipeline of our SAM3-based method with category-guided interaction between MLLMs and SAM3, producing the final mask. **Bottom-right:** Performance comparison of COD methods under varying supervision. Our method outperforms UCOD and weakly-supervised methods.

## 1 Introduction

Camouflaged object detection (COD) aims to detect objects that are highly concealed within their surrounding environments [Fan *et al.*, 2020a; Fan *et al.*, 2021; Fan *et al.*, 2023]. These objects exhibit visual appearances that are extremely similar to the background, particularly along object boundaries that almost blend into the surroundings, making the task highly challenging. Despite this difficulty, camouflaged object detection has broad application potential in real-world scenarios, such as rare species discovery [Fan *et al.*, 2020a], medical image segmentation (e.g., polyp segmentation [Fan *et al.*, 2020b], lung infection segmentation [Fan *et al.*, 2020c]), and agriculture (e.g., locust detection to prevent invasion [Fan *et al.*, 2020a]).

Considering that annotating camouflaged objects is particularly time-consuming and labor-intensive due to their low visibility [Fan *et al.*, 2020a; Chen *et al.*, 2024b], unsupervised camouflaged object detection (UCOD) has attracted increasing attention as a promising alternative. Early UCOD methods require training and generally suffer from limited performance (e.g., FOUND [Siméoni *et al.*, 2023], UCOS-DA [Zhang and Wu, 2023]), resulting in low practical usability. With the rapid development of foundation models, new opportunities have emerged for UCOD. Many recent UCOD approaches (e.g., GenSAM [Hu *et al.*, 2024a], ProMac [Hu *et al.*, 2024b], CVP [Tang *et al.*, 2024]) leverage foundation models to significantly improve performance without any

*Corresponding Authors

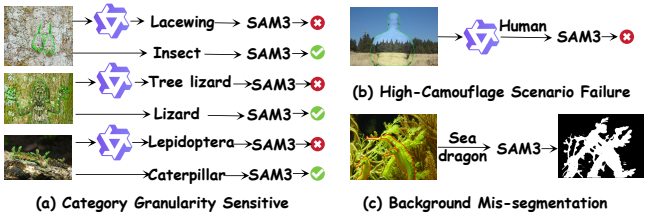


Figure 2: Issues arising from SAM3 in UCOD, *i.e.,* (a) Category granularity sensitivity: Failures caused by misalignment of category granularity. (b) High-camouflage scenarios failure: Failures due to low confidence in highly camouflaged regions. (c) Background mis-segmentation: Incorrect segmentation of regions with similar background.

training. However, a substantial performance gap with supervised models still remains, which to some extent limits their real-world applicability. Most of these UCOD methods follow a common paradigm of "MLLM + other foundation models + SAM". The core idea is to combine MLLMs with other foundation models to generate spatial prompts for target objects, which are then fed into SAM [Kirillov *et al.*, 2023] for segmentation. We argue that this paradigm may suffer from fundamental bottlenecks in camouflaged scenarios. This bottleneck mainly arises from two aspects. First, the spatial localization capability of MLLMs is limited in camouflaged scenarios. The boundaries between camouflaged objects and the background are highly ambiguous, making it difficult for MLLMs to provide accurate spatial prompts. Second, SAM is highly sensitive to spatial prompts in camouflaged scenes. Inaccurate prompts can severely degrade segmentation quality, forming a vicious cycle between unreliable localization and fragile segmentation. Consequently, we argue that spatial information interaction between MLLMs and SAM is likely a key factor constraining performance. The recent emergence of SAM3 [Carion *et al.*, 2025] offers a new opportunity by supporting category-level prompts for the first time. This inspires us to explore a new direction—bypassing unreliable spatial interactions and instead leveraging more robust categories for model interaction, potentially breaking the performance bottleneck of existing paradigms, as shown in Fig. 1.

Although SAM3 can be directly paired with MLLMs to provide a potential solution for UCOD, effectively leveraging it remains non-trivial. First, in UCOD, MLLMs are required to obtain the category via task-specific prompts (e.g., "Return the camouflaged object category"). However, this process is inherently uncontrollable, as MLLMs may return categories at varying levels of granularity. Correspondingly, SAM3 is highly sensitive to textual prompts in camouflaged scenarios, where even slight variations in category granularity for the same image can significantly affect segmentation results, as shown in Fig. 2. As a result, the category granularity produced by MLLMs may be suboptimal for SAM3, leading to segmentation failures. To address this issue, we propose Fine-grained Category Query, which explicitly instructs MLLMs to consider categories at multiple granularity levels in the prompt and dynamically return a full-granularity category chain for the camouflaged object. This design effectively mitigates the negative impact of category granularity on SAM3 and improves segmentation robustness.

After obtaining the full-granularity category chain, it is used as a prompt set and fed into SAM3 to generate segmentation masks. Unfortunately, SAM3 is not trained on COD datasets and has a limited understanding of the camouflage concept. As a result, two typical failure cases may occur: 1) even with correct category prompts, SAM3 may lack confidence when segmenting highly camouflaged objects, leading to segmentation failure; and 2) due to the absence of explicit constraints on the camouflage concept during segmentation, SAM3 may easily include surrounding background regions with similar appearances, as shown in Fig. 2. To address these issues, we first introduce Text Prompt Augmentation tailored for UCOD, which explicitly incorporates the camouflage concept into the prompts, thereby improving the segmentation of highly camouflaged objects. In addition, we propose Semantic-Geometric Dual Confirmation, which jointly evaluates the generated masks from both semantic and spatial perspectives to filter out masks that correspond to background regions.

Despite these designs substantially improving the performance of the proposed method, failure cases remain unavoidable. In particular, when MLLMs generate inaccurate responses due to the extreme ambiguity of camouflaged samples or inherent hallucination effects, such errors can propagate to the subsequent SAM3 stage and ultimately lead to erroneous segmentation results. To further enhance reliability, we introduce Semantic-Geometric Reasoning Injection that supplements MLLMs with critical semantic and geometric cues, enabling more reliable reasoning in camouflaged scenarios. The refined category information is then fed back to SAM3 to produce more robust segmentation results. Overall, this design allows the system to progressively correct errors and significantly improves segmentation stability under challenging camouflage conditions.

Our contributions are summarized as follows:

- We propose a novel training-free framework for UCOD. To the best of our knowledge, this is the first work to leverage SAM3 in the UCOD setting by interacting with MLLMs through category-level prompts, rather than spatial prompts.
- We introduce Fine-grained Category Queries, enabling MLLMs to generate full-granularity category chains, thereby mitigating the adverse impact of SAM3's sensitivity to category granularity on UCOD.
- We propose Semantic–Geometric Dual Confirmation and Semantic–Geometric Reasoning Injection. The former jointly validates segmentation results from semantic and geometric perspectives, while the latter injects semantic and geometric priors into MLLMs to enhance reasoning over failure cases, thereby improving suboptimal outcomes in challenging camouflaged scenarios.
- Experimental results demonstrate that our method significantly outperforms existing UCOD approaches, achieves performance comparable to WSCOD methods, and even approaches that of fully supervised methods.

## 2 Related Work

**Camouflaged Object Detection.** Camouflaged Object Detection (COD) focuses on detecting camouflaged objects within an image. Research on Camouflaged Object Detection (COD) has progressed rapidly, yet most existing methods rely on fully supervised annotations [Pang *et al.*, 2022; Luo *et al.*, 2024; He *et al.*, 2025b], which incur substantial labeling costs. To reduce annotation overhead, many works explore weakly supervised settings, using annotations such as points [Chen *et al.*, 2024a], bounding boxes [Chen *et al.*, 2024b], and scribbles [He *et al.*, 2025a; He *et al.*, 2023b] as alternatives to pixel-level labels. Although these approaches alleviate labeling effort to some extent, they still require per-image annotation. Beyond weak supervision, a growing body of work investigates COD under unsupervised settings. Existing UCOD methods can be broadly categorized into two groups. The first group consists of training-based UCOD methods, such as UCOS-DA [Zhang and Wu, 2023] and UCOD-DPL [Yan *et al.*, 2025], which follow a paradigm of generating pseudo labels and then training segmentation models. The second group includes training-free UCOD methods [Tang *et al.*, 2024; Hu *et al.*, 2024a; Hu *et al.*, 2024b; Du *et al.*, 2025], which rely on foundation models including MLLMs, diffusion models, DINOv2 [Oquab *et al.*, 2023], CLIP [Radford *et al.*, 2021], and SAM. These methods typically follow a paradigm of combining MLLMs with other foundation models to generate spatial prompts, which are then used by SAM for segmentation. Benefiting from foundation models, this category generally achieves better performance than training-based UCOD methods.

**Multimodal Large Language Models for COD**. Multimodal Large Language Models (MLLMs) achieve unified vision-language understanding through joint visual-textual token alignment. These MLLMs, like GPT-4V [Achiam *et al.*, 2023], LLaVA-1.5 [Liu *et al.*, 2024] and Qwen2.5-VL [Bai *et al.*, 2025], demonstrate unparalleled adaptability across diverse tasks. Many studies employ MLLMs for UCOD. CVP [Tang *et al.*, 2024] proposes CoVP to improve MLLMs' perception in COD. GenSAM [Hu *et al.*, 2024a] tries to map text prompts onto image heatmaps using MLLMs, acquiring visual prompts. ProMaC [Hu *et al.*, 2024b] leverages hallucinations from large models to obtain instance-specific prompts. EASE [Du *et al.*, 2025] utilizes MLLMs to extract category-level textual information from images, which is then input into a Diffusion Model to generate corresponding images.

**Segment Anything Model for COD.** Segment Anything Model (SAM) [Kirillov *et al.*, 2023] is an interactive segmentation model that supports multiple types of prompts. Owing to this flexibility, SAM has been widely adopted in both weakly supervised and unsupervised COD [Chen *et al.*, 2024b; Tang *et al.*, 2024; Hu *et al.*, 2024a; Hu *et al.*, 2024b; Du *et al.*, 2025]. In WSCOD, methods [He *et al.*, 2025a; Chen *et al.*, 2024b] use weak annotations as prompts to generate pseudo masks via SAM. In UCOD, approaches [Tang *et al.*, 2024; Hu *et al.*, 2024a; Hu *et al.*, 2024b; Du *et al.*, 2025] obtain spatial prompts by combining MLLMs with other foundation models and feed them into SAM for segmentation. Recently, SAM3 [Carion *et al.*, 2025] further extends SAM by supporting category-level prompts, enabling segmentation to be conditioned on semantic categories rather than explicit spatial cues.

## 3 Approach

The overall architecture of the proposed framework is illustrated in Fig. 3. The Fine-grained Category Query (FCQ) explicitly guides MLLMs to comprehensively reason about the camouflaged object categories at multiple semantic granularities. Semantic–Geometric Dual Confirmation (SGDC) jointly validates the masks produced by SAM3 from both semantic and geometric perspectives. Semantic–Geometric Reasoning Injection (SGRI) further enhances the reasoning capability of MLLMs by incorporating prior knowledge from both semantic and geometric domains, enabling robust handling of hard samples caused by segmentation failures.

### 3.1 Fine-grained Category Query

We employ a task prompt as the input to MLLMs to obtain the camouflaged object category, which is then used as the prompt for SAM3 to generate segmentation masks. However, the granularity of the categories predicted by MLLMs is highly variable, and SAM3 is sensitive to category specifications in camouflaged scenarios. This mismatch often leads to sub-optimal segmentation results, as illustrated in Fig. 2. To address this issue, we propose Fine-grained Category Query (FCQ), which explicitly enforces MLLMs to reason about camouflaged object categories at multiple levels of granularity within the task prompt, as shown in Fig. 3. Specifically, conventional UCOD methods adopt a simple task prompt for MLLMs, e.g., "Return the camouflaged object localization and category." In contrast, we augment the prompt by explicitly requiring MLLMs to identify camouflaged objects at different granularity levels and to dynamically return a complete chain of categories across these levels according to each input image. This design effectively mitigates the negative impact of category-granularity sensitivity on SAM3. In addition, we further require MLLMs to output background category and localization information in the prompt, which will be utilized by subsequent modules. The detailed prompt design is provided in the supplementary material.

### 3.2 Semantic-Geometric Dual Confirmation

After obtaining the full-granularity category chain, it is used as a prompt set for SAM3 to generate segmentation masks. However, since SAM3 is not trained on COD datasets, it has a limited understanding of the concept of camouflage, which leads to two major issues: 1) SAM3 exhibits low confidence when segmenting highly camouflaged objects. 2) Due to the lack of explicit constraints on the camouflaged concept, it tends to mistakenly segment visually similar background regions. To address these problems, we propose Text Prompt Augmentation and Semantic–Geometric Dual Confirmation (SGDC). The former enhances SAM3's ability to segment camouflaged objects by strengthening its understanding of the camouflage concept, while the latter performs a comprehensive quality assessment of the generated segmentation masks.

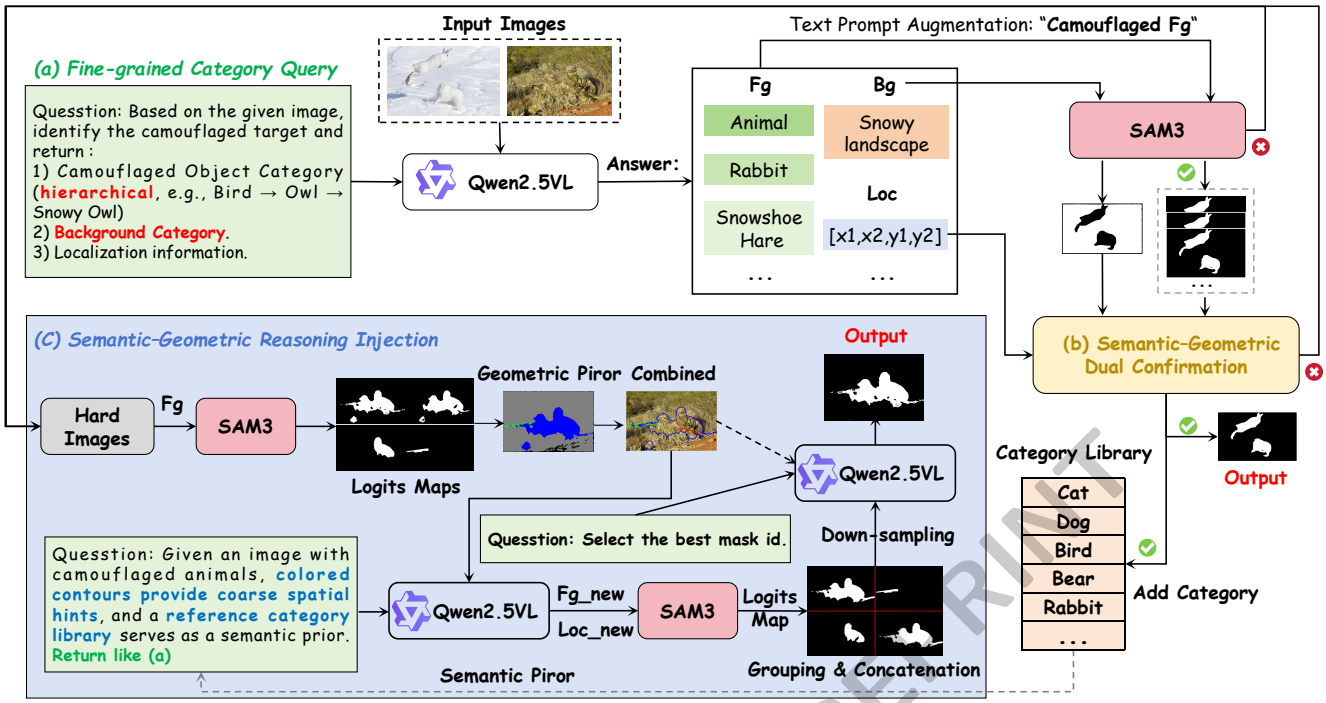


Figure 3: **The main framework of our proposed method.** The MLLM first obtains the category granularity chain, background categories, and bounding boxes through the Fine-grained Category Query (FCQ). The category granularity chain and background categories are then used with SAM3 to generate the corresponding masks. The Semantic-Geometric Dual Confirmation (SGDC) module is subsequently applied to identify difficult samples with poor segmentation. Finally, Semantic-Geometric Reasoning Injection (SGRI) refines these difficult samples by injecting semantic and geometric priors into the MLLM.



Specifically, MLLMs take the task prompt and the input image described in the previous section as input, and output a full-granularity category chain, the background category, and the corresponding localization coordinates:

$$\{C_{fg}, C_{bg}, B\} = \text{MLLMs}(I, P_{task}), \tag{1}$$

where $I$ denotes the input image, $P_{task}$ is the task prompt, $B$ represents the bounding box, $C_{fg}$ denotes the category chain, and $C_{bg}$ represents the background category. The category chain $C_{fg}$ is further augmented in a COD-specific manner:

$$\hat{C}^i_{fg} = \{\, C^i_{fg},\ \text{“Camouflaged”} \oplus C^i_{fg} \,\}, \tag{2}$$

where $\oplus$ denotes string concatenation, $i$ represents the index of the granularity level in the full-granularity category chain $C_{fg} = \{C^1_{fg}, C^2_{fg}, \ldots, C^N_{fg}\}$, ordered from coarse to fine. $\hat{C}^i_{fg}$ and $C_{bg}$ are used as category prompts for SAM3, yielding multiple foreground masks and a single background mask, respectively. The corresponding segmentation masks are obtained as:

$$\mathcal{M}_{fg} = \{\, \text{SAM3}(I, \hat{C}^i_{fg}) \mid i = 1, \ldots, N \,\}, \tag{3}$$

$$\mathcal{M}_{bg} = \text{SAM3}(I, C_{bg}). \tag{4}$$

where $\mathcal{M}_{fg}$ and $\mathcal{M}_{bg}$ denote the sets of foreground masks and the background mask obtained from SAM3, respectively. It is worth noting that $\mathcal{M}_{fg}$ may contain zero or more masks, as SAM3 may produce no output for a given category prompt.

For the non-empty cases of $\mathcal{M}_{fg}$, we perform a semantic and geometric dual assessment. From the semantic perspective, well-segmented samples should exhibit complementary foreground and background masks with low overlap. Therefore, we use the Intersection-over-Union (IoU) to evaluate the degree of overlap between $\mathcal{M}_{fg}$ and $\mathcal{M}_{bg}$:

$$\text{Sig}^i_{sem} = \begin{cases} 1, & \text{if IoU}(m^i_{fg}, \mathcal{M}_{bg}) < \tau, \\ 0, & \text{otherwise}, \end{cases} \quad \forall m^i_{fg} \in \mathcal{M}_{fg}, \tag{5}$$

where $m^i_{fg}$ denotes the $i$-th foreground mask in $\mathcal{M}_{fg}$, $\tau$ is a threshold controlling the allowed overlap, and $\text{Sig}^i_{sem}$ indicates whether $m^i_{fg}$ satisfies the semantic complementarity constraint. From the geometric perspective, although the bounding boxes predicted by MLLMs are not perfectly aligned at the boundaries, they can still roughly indicate the location of the identified camouflaged objects. Therefore, we extract the enclosing bounding box of each foreground mask in $\mathcal{M}_{fg}$ as the predicted object location, and compare it with the box provided by MLLMs:

$$\text{Sig}^i_{geo} = \begin{cases} 1, & \text{if IoU}\Big(\text{Box}(m^i_{fg}), B\Big) > \eta, \\ 0, & \text{otherwise}, \end{cases} \quad \forall m^i_{fg} \in \mathcal{M}_{fg}, \tag{6}$$

where $\text{Box}(\cdot)$ denotes the bounding box of mask, $\eta$ is a threshold, and $\text{Sig}^i_{geo}$ indicates whether $m^i_{fg}$ is geometrically consistent. Only masks with $\text{Sig}^i_{sem} = 1$ and $\text{Sig}^i_{geo} = 1$ are

| Methods | Backbones | Sup. | CAMO | | | | COD10K | | | | NC4K | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
| *Fully-Supervised Methods* | | | | | | | | | | | | | | |
| SINet [Fan *et al.*, 2020a] CVPR'20 | ResNet50 | F | 0.092 | 0.745 | 0.804 | 0.644 | 0.043 | 0.776 | 0.864 | 0.631 | 0.058 | 0.808 | 0.871 | 0.723 |
| UGTR [Yang *et al.*, 2021] ICCV'21 | ResNet50 | F | 0.086 | 0.784 | 0.822 | 0.684 | 0.036 | 0.817 | 0.852 | 0.666 | 0.052 | 0.839 | 0.874 | 0.747 |
| ZoomNet [Pang *et al.*, 2022] CVPR'22 | ResNet50 | F | 0.066 | 0.820 | 0.892 | 0.752 | 0.029 | 0.838 | 0.911 | 0.729 | 0.043 | 0.853 | 0.896 | 0.784 |
| FEDER [He *et al.*, 2023a] CVPR'23 | Res2Net50 | F | 0.066 | 0.836 | 0.897 | 0.806 | 0.029 | 0.844 | 0.911 | 0.736 | 0.042 | 0.862 | 0.913 | 0.823 |
| SAM-Ada. [Chen *et al.*, 2023] ICCVW'23 | VIT-H | F | 0.070 | 0.847 | 0.873 | 0.765 | 0.025 | 0.883 | 0.918 | 0.801 | - | - | - | - |
| HitNet [Hu *et al.*, 2023] AAAI'23 | PVT V2 | F | 0.060 | 0.834 | 0.892 | 0.801 | 0.027 | 0.847 | 0.922 | 0.798 | 0.042 | 0.858 | 0.911 | 0.825 |
| CamoFocus [Khan *et al.*, 2024] WACV'24 | PVT V2 | F | **0.044** | 0.870 | 0.924 | 0.842 | 0.022 | 0.868 | 0.931 | 0.802 | 0.031 | 0.886 | 0.932 | 0.853 |
| RUN [He *et al.*, 2025b] ICML'25 | PVT V2 | F | 0.045 | **0.877** | 0.934 | **0.853** | **0.021** | 0.878 | 0.941 | 0.808 | **0.030** | **0.892** | 0.940 | **0.866** |
| SAM-COD[1] [Liu *et al.*, 2025] ICCV'25 | PVT V2, SAM-B | F | 0.047 | 0.866 | **0.948** | 0.839 | 0.023 | **0.881** | **0.969** | **0.817** | 0.032 | 0.889 | **0.963** | 0.857 |
| *Weakly-supervised Methods* | | | | | | | | | | | | | | |
| CRNet [He *et al.*, 2023b] AAAI'23 | ResNet50 | S | 0.092 | 0.735 | 0.815 | 0.641 | 0.049 | 0.733 | 0.832 | 0.576 | 0.063 | 0.775 | 0.855 | 0.688 |
| MiNet [Niu *et al.*, 2024] MM'24 | ResNet50 | S | 0.091 | 0.750 | 0.840 | 0.669 | 0.049 | 0.749 | 0.840 | 0.596 | 0.061 | 0.793 | 0.869 | 0.709 |
| SSE [He *et al.*, 2025a] TPAMI'25 | ResNet50 | S | 0.090 | 0.765 | 0.826 | 0.742 | 0.036 | 0.807 | 0.883 | 0.724 | 0.051 | 0.836 | 0.891 | 0.805 |
| SAM-COD [Chen *et al.*, 2024b] ECCV'24 | PVT V2 | S | **0.060** | 0.836 | **0.903** | 0.779 | 0.029 | 0.833 | 0.904 | 0.728 | 0.039 | 0.859 | 0.912 | 0.795 |
| PCOD [Chen *et al.*, 2024a] ECCV'24 | PVT V2 | P | 0.074 | 0.798 | 0.872 | 0.721 | 0.042 | 0.784 | 0.859 | 0.650 | 0.051 | 0.822 | 0.889 | 0.748 |
| SAM-COD [Chen *et al.*, 2024b] ECCV'24 | PVT V2 | B | 0.062 | **0.837** | 0.901 | **0.786** | **0.028** | **0.842** | **0.914** | **0.745** | **0.037** | **0.867** | **0.923** | **0.813** |
| *Unsupervised Methods (with model training)* | | | | | | | | | | | | | | |
| FOUND [Siméoni *et al.*, 2023] CVPR'23 | DINO V1 | U | 0.127 | 0.701 | 0.784 | 0.606 | 0.086 | 0.689 | 0.740 | 0.513 | 0.085 | 0.755 | 0.819 | 0.656 |
| UCOS-DA [Zhang and Wu, 2023] ICCV'23 | DINO V1 | U | 0.129 | 0.685 | 0.782 | 0.584 | 0.085 | 0.670 | 0.751 | 0.482 | 0.084 | 0.741 | 0.824 | 0.637 |
| SdalsNet [Shou *et al.*, 2025] AAAI'25 | DINO V1 | U | 0.117 | 0.696 | - | - | 0.071 | 0.696 | - | - | 0.084 | 0.738 | - | - |
| UCOD-DPL [Yan *et al.*, 2025] CVPR'25 | DINO V1 | U | 0.108 | 0.706 | 0.801 | 0.621 | 0.059 | 0.727 | 0.822 | 0.577 | 0.074 | 0.761 | 0.851 | 0.680 |
| *Unsupervised Methods (without training)* | | | | | | | | | | | | | | |
| GenSAM [Hu *et al.*, 2024a] AAAI'24 | BLIP-$2_{6.7B}$, CLIP, SAM | U | 0.117 | 0.730 | - | 0.655 | 0.069 | 0.731 | - | 0.584 | 0.087 | 0.754 | - | 0.665 |
| CVP [Tang *et al.*, 2024] MM'24 | $Shikra_{7B}$, DINO V2, SAM-HQ | U | 0.100 | 0.749 | - | 0.680 | 0.065 | 0.733 | - | 0.592 | 0.082 | 0.768 | - | 0.681 |
| ProMaC [Hu *et al.*, 2024b] NIPS'24 | LLaVA-$1.5_{13B}$, CLIP, SAM, SD V2 | U | 0.098 | 0.775 | 0.832 | 0.698 | 0.045 | 0.795 | 0.861 | 0.624 | 0.073 | 0.779 | 0.845 | 0.707 |
| EASE [Du *et al.*, 2025] CVPR'25 | LLaVA-$1.5_{7B}$, DINO V2, SD V1.5, SAM | U | 0.099 | 0.777 | 0.841 | 0.748 | **0.028** | 0.856 | 0.914 | 0.801 | 0.051 | 0.849 | 0.902 | 0.815 |
| Ours | Qwen2.5-$VL_{3B}$, SAM3 | U | **0.056** | **0.842** | **0.892** | **0.810** | 0.034 | **0.859** | **0.918** | **0.802** | **0.035** | **0.891** | **0.935** | **0.873** |

Table 1: Comparison of our methods with recent methods. "F", "S", "B", "P" and "U" denote fully-supervised, scribble-supervised, box-supervised, point-supervised and unsupervised methods, respectively. **Bold** indicates the best result in group settings, and underline indicates the second-best result.



retained as valid candidates, and all valid masks are finally fused by mean aggregation to generate the final output mask.

### 3.3 Semantic-Geometric Reasoning Injection

Although most samples are successfully segmented after SGDC, a subset of hard cases still fails. This is mainly due to inaccurate predictions from MLLMs, which consequently lead to failures in SAM3. To address this issue, we propose Semantic–Geometric Reasoning Injection (SGRI), which injects semantic and geometric priors into MLLMs to enhance their reasoning ability on hard samples. For semantic priors, we collect the category information from successfully segmented samples to construct a category library, which is then provided to MLLMs as a set of highly probable candidate categories. For geometric priors, we extract the logits maps from SAM3 for the failed samples. Although these logits maps may be noisy and incomplete, they still contain informative boundary cues that are valuable for identifying camouflaged objects, as illustrated in Fig. 3. We therefore extract boundary information from the logits maps and combine it with the original image to inject geometric priors into MLLMs. With both semantic and geometric priors, MLLMs are prompted to re-reason and generate a new foreground category chain. To ensure a reliable final mask, the category chain is fed back into SAM3 to obtain a set of logits maps. We then group the logits maps according to their IoU similarity. Logits maps within the same group are first fused to form a group-level mask. The fused masks from different groups are subsequently concatenated to construct a composite mask image. This is further downsampled to reduce computational cost and is concatenated with the input image as the input to MLLMs, which performs a final selection to produce the ultimate segmentation result.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We evaluate our method on three widely used camouflaged object detection benchmarks: CAMO [Le *et al.*, 2019], COD10K [Fan *et al.*, 2020a], and NC4K [Lv *et al.*, 2021], which contain 250, 2,026, and 4,121 test images, respectively.

**Evaluation Metrics.** We employ four commonly used metrics for evaluation: Mean Absolute Error (MAE), S-measure ($S_m$) [Fan *et al.*, 2017], E-measure ($E_m$) [Fan *et al.*, 2018], weighted F-measure ($F_\beta^w$) [Margolin *et al.*, 2014].

**Implementation Details.** We implement our method in PyTorch and conduct all experiments on a single NVIDIA RTX A6000 GPU. We adopt the official release of SAM3 [Carion *et al.*, 2025] for all experiments. For the MLLM component, we use Qwen2.5-VL-3B [Bai *et al.*, 2025], and the detailed prompts are provided in the supplementary material. The logits maps are binarized with a fixed threshold of 128 before being used.

### 4.2 Compare with State-of-the-art Methods

**Quantitative Comparison.** As shown in Tab. 1, we conduct a comprehensive comparison with COD methods of different learning paradigms. Compared with fully supervised methods, our approach outperforms several strong fully supervised baselines. Compared with weakly supervised methods, our method achieves comparable performance without any training. Compared with training-free unsupervised methods, our

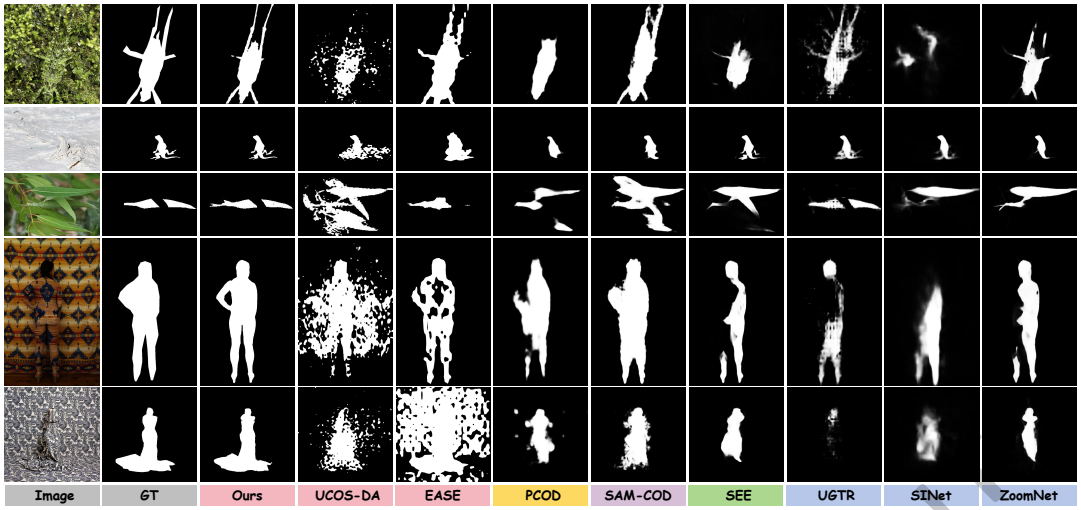


Figure 4: Visual comparison with state-of-the-art fully supervised, weakly supervised, and unsupervised methods. The colors pink, yellow, purple, green, and blue indicate unsupervised, point-supervised, box-supervised, scribble-supervised, and fully supervised methods, respectively.

| Methods | Foundation Models. | Iterations | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
|---|---|---|---|---|---|---|
| GenSAM | BLIP-$2_{6.7B}$, CLIP, SAM | 12 (Fully) | 0.117 | 0.730 | - | 0.655 |
| ProMac | LLaVA-$1.5_{13B}$, CLIP, SAM, SD V2 | 6 (Fully) | 0.098 | 0.775 | 0.832 | 0.698 |
| Ours | Qwen2.5-$VL_{3B}$, SAM3 | 3 (Partly) | **0.056** | **0.842** | **0.895** | **0.810** |

Table 2: Comparison of UCOD methods in terms of model complexity and performance.

| Components | | | | | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
|---|---|---|---|---|---|---|---|---|
| MLLM | SAM3 | FCQ | SGDC | SGRI | | | | |
| ✓ | ✓ | | | | 0.144 | 0.684 | 0.696 | 0.568 |
| ✓ | ✓ | ✓ | | | 0.157 | 0.724 | 0.779 | 0.689 |
| ✓ | ✓ | ✓ | ✓ | | 0.142 | 0.741 | 0.806 | 0.704 |
| ✓ | ✓ | ✓ | ✓ | ✓ | **0.056** | **0.842** | **0.895** | **0.810** |

Table 3: Ablation study of our proposed method.

approach delivers superior performance while using fewer foundation models and significantly fewer parameters. Compared with unsupervised methods that require training, our method not only eliminates the need for training but also substantially surpasses these methods.

**Qualitative Evaluation.** We visualize the predicted camouflaged object segmentation masks produced by our method and competing approaches under several challenging scenarios, including small objects, large objects, complex backgrounds, and blurry boundaries. As shown in Fig. 4, our method consistently achieves higher segmentation accuracy than existing UCOD and WSCOD methods, producing clearer object boundaries and more detailed representations of camouflaged objects.

**Computational Cost Analysis.** As shown in Tab. 1, although our model relies on foundation models, it requires fewer foundation models and has a smaller overall parameter footprint compared with other high-performing UCOD methods. Furthermore, unlike ProMac and GenSAM, which require six and twelve full iterations respectively, our method adopts a more targeted strategy by dynamically selecting objects for subsequent optimization based on the input sample, as shown in Tab. 2. These optimization targets constitute only a small subset of the dataset, and the additional optimization involves only two extra MLLM inferences. Since the foundation models do not require training, the memory footprint is approximately ∼11 GB, and inference on a single image takes only ∼4.14 s, which is substantially lower than the training cost of supervised methods (e.g., SAM-COD requires ∼21 GB of memory and ∼7 hours of training). Additionally, although each sample involves multiple category queries to SAM3, the image encoder is executed only once, and its features are reused across different queries. The additional computational overhead introduced by the text encoder and mask decoder is relatively small compared to that of the image encoder.

### 4.3 Ablation Study

To verify the effectiveness of our modules, we conduct ablation studies on the most challenging dataset, CAMO, where all methods achieve the lowest scores according to Tab. 1.

**Effectiveness of Fine-grained Category Query.** As shown in Tab. 3, compared with directly using a task prompt that prompts the MLLM to return a single category of camouflaged object, our fine-grained category queries significantly improve the final performance.

**Effectiveness of Semantic-Geometric Dual Confirmation.** The Semantic-Geometric Dual Confirmation (SGDC) module identifies low-quality segmentation masks and applies subsequent refinement to these cases, thereby improving the final segmentation performance. In the ablation study of this module, we disable the subsequent refinement operation (SGRI) for the identified masks. Instead, we use bounding boxes predicted by the MLLM as prompts for SAM3 to generate replacement masks. Even under this setting, the overall performance still improves significantly, as shown in Tab. 3.

**Effectiveness of Semantic-Geometric Reasoning Injection.** As shown in Tab. 3, applying Semantic-Geometric Rea-

| $\eta$ | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ | $\tau$ | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
|---|---|---|---|---|---|---|---|---|---|
| 0.20 | 0.058 | 0.837 | 0.889 | 0.806 | 0.20 | 0.057 | 0.838 | 0.890 | 0.807 |
| 0.30 | 0.056 | 0.842 | 0.895 | 0.810 | 0.25 | 0.057 | 0.839 | 0.892 | 0.808 |
| 0.40 | 0.057 | 0.839 | 0.891 | 0.808 | 0.30 | 0.056 | 0.842 | 0.895 | 0.810 |
| 0.50 | 0.059 | 0.832 | 0.886 | 0.803 | 0.35 | 0.058 | 0.834 | 0.887 | 0.806 |

Table 4: Parameter analysis on $\eta$ and $\tau$.

| Type | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
|---|---|---|---|---|
| Qwen2.5-$VL_{3B}$ [Bai *et al.*, 2025] | 0.056 | 0.842 | 0.895 | 0.810 |
| Qwen2.5-$VL_{7B}$ [Bai *et al.*, 2025] | 0.054 | 0.847 | 0.901 | 0.813 |
| Qwen2.5-$VL_{72B}$ [Bai *et al.*, 2025] | **0.053** | **0.851** | **0.904** | **0.815** |

Table 5: Ablation study of our proposed method with different MLLM scales.

| Type | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
|---|---|---|---|---|
| LLaVA-$1.5_{7B}$ [Liu *et al.*, 2023] | 0.061 | 0.827 | 0.874 | 0.790 |
| Qwen2.5-$VL_{3B}$ [Bai *et al.*, 2025] | **0.056** | **0.842** | **0.895** | **0.810** |

Table 6: Ablation study of our method using different MLLM backbones.

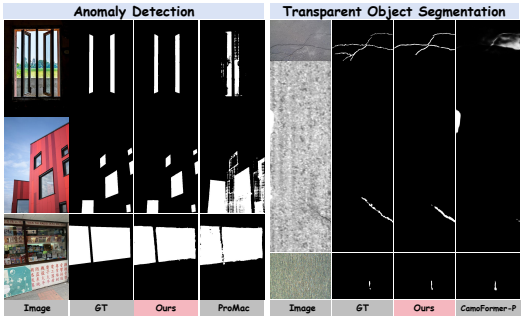


Figure 5: Qualitative results on transparent object segmentation and anomaly detection.

soning Injection to the low-quality masks identified by SGDC leads to a substantial improvement in the final performance.

### 4.4 Further Analysis

**Hyperparameter sensitivity.** Two hyperparameters are involved in our model, namely $\tau$ and $\eta$. We evaluate different values of these hyperparameters, as reported in Tab. 4. Based on the results, we set $\tau = 0.3$ and $\eta = 0.3$ in all experiments. Notably, the model performance is relatively insensitive to the specific values of these two hyperparameters.

**Comparison across Different MLLMs.** As shown in Tab. 5, we evaluate the performance of our method with MLLMs of different scales. While the performance generally improves as the model scale increases, the gains are relatively modest. In addition, we examine the performance across different types of MLLMs. As shown in Tab. 6, our method still achieves strong results when LLaVA is used as the MLLM.

### 4.5 Transferability Studies

To evaluate the transferability of our proposed method to related downstream tasks, we conduct experiments on Anomaly Detection (AD) and Transparent Object Segmentation (TOS).

**Extension to Transparent Object Segmentation.** To demonstrate the strong generalization capabilities, we evaluated our model on the transparent object segmentation (TOS) task. Specifically, we evaluate our proposed model on the GSD [Lin *et al.*, 2021] dataset for TOS. As shown in Tab. 7 and Fig. 5, our method also performs well on the TOS task.

| Methods | GSD [Lin *et al.*, 2021] | | | |
|---|---|---|---|---|
| | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta$↑ |
| *Weakly-supervised Methods* | | | | |
| SAM-P [Kirillov *et al.*, 2023] | 0.266 | 0.514 | 0.501 | 0.473 |
| WSSA [Zhang *et al.*, 2020] | 0.185 | 0.650 | 0.712 | 0.661 |
| CRNet [He *et al.*, 2023b] | 0.154 | 0.710 | 0.751 | 0.743 |
| *Unsupervised Methods* | | | | |
| GenSAM [Hu *et al.*, 2024a] | 0.155 | 0.559 | 0.700 | 0.394 |
| ProMac [Hu *et al.*, 2024b] | 0.147 | 0.569 | 0.723 | 0.409 |
| Ours | **0.103** | **0.786** | **0.842** | **0.796** |

Table 7: Performance comparison on the Transparent Object Segmentation task.

| Methods | CDS2K [Fan *et al.*, 2023] | | | |
|---|---|---|---|---|
| | MAE↓ | $S_m$↑ | $E_m$↑ | $F_\beta^w$↑ |
| SINetV2 [Fan *et al.*, 2021] | 0.102 | 0.551 | 0.567 | 0.215 |
| HitNet [Hu *et al.*, 2023] | 0.118 | 0.563 | 0.564 | 0.276 |
| DGNet [Ji *et al.*, 2023] | 0.089 | 0.578 | 0.569 | 0.258 |
| CamoFormer-P [Yin *et al.*, 2024] | 0.100 | 0.589 | 0.588 | 0.298 |
| Ours | **0.047** | **0.627** | **0.603** | **0.340** |

Table 8: Performance comparison on the Anomaly Detection task.

**Extension to Anomaly Detection.** For the anomaly detection task, we conduct experiments on the CDS2K benchmark dataset [Fan *et al.*, 2023], following the experimental settings of [Fan *et al.*, 2023]. As shown in Tab. 8 and Fig. 5, our method also achieves competitive performance on anomaly detection.

## 5 Conclusion

In this work, we presented a novel training-free framework for unsupervised camouflaged object detection that leverages SAM3 via category-level interaction with MLLMs. By introducing Fine-grained Category Queries, Text Prompt Augmentation, and Semantic–Geometric Dual Confirmation with Reasoning Injection, our method effectively mitigates the limitations of spatial prompts and improves segmentation robustness in challenging camouflaged scenarios. Extensive experiments demonstrate that our method consistently outperforms existing UCOD approaches and achieves performance comparable to weakly supervised methods, all without requiring any training. Beyond its empirical success, our method provides a new paradigm for integrating foundation models in UCOD, highlighting the potential of robust category-level interactions for tackling highly ambiguous visual tasks.



## Acknowledgements

This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (grant no. JYB2025XDXM902), the National Natural Science Foundation of China (grant no. 62502200), the Jiangsu Provincial Science and Technology Major Project (grant no. BG2024042), and the Natural Science Foundation of Jiangsu Province (grant no. BK20251203).

## References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Bai *et al.*, 2025] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

[Carion *et al.*, 2025] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. Sam 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025.

[Chen *et al.*, 2023] Tianrun Chen, Lanyun Zhu, Chaotao Deng, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *ICCV*, pages 3367–3375, 2023.

[Chen *et al.*, 2024a] Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. Just a hint: Point-supervised camouflaged object detection. In *ECCV*, pages 332–348. Springer, 2024.

[Chen *et al.*, 2024b] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. Sam-cod: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *ECCV*, pages 315–331. Springer, 2024.

[Du *et al.*, 2025] 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.

[Fan *et al.*, 2017] 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.

[Fan *et al.*, 2018] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 698–704, 2018.

[Fan *et al.*, 2020a] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, pages 2777–2787, 2020.

[Fan *et al.*, 2020b] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, pages 263–273. Springer, 2020.

[Fan *et al.*, 2020c] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8):2626–2637, 2020.

[Fan *et al.*, 2021] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 44(10):6024–6042, 2021.

[Fan *et al.*, 2023] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):16, 2023.

[He *et al.*, 2023a] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *CVPR*, pages 22046–22055, 2023.

[He *et al.*, 2023b] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson WH Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *AAAI*, volume 37, pages 781–789, 2023.

[He *et al.*, 2025a] Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Youwei Pang, Longxiang Tang, Chengyu Fang, Yulun Zhang, Linghe Kong, Xiu Li, et al. Segment concealed objects with incomplete supervision. *IEEE TPAMI*, 2025.

[He *et al.*, 2025b] 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. *arXiv preprint arXiv:2501.18783*, 2025.

[Hu *et al.*, 2023] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *AAAI*, volume 37, pages 881–889, 2023.

[Hu *et al.*, 2024a] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in sam: A single generic prompt for segmenting camouflaged objects. In *AAAI*, volume 38, pages 12511–12518, 2024.

[Hu *et al.*, 2024b] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. *Advances in Neural Information Processing Systems*, 37:107171–107197, 2024.

[Ji *et al.*, 2023] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *MIR*, 20(1):92–108, 2023.

[Khan *et al.*, 2024] 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.

[Kirillov *et al.*, 2023] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura

Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023.

[Le *et al.*, 2019] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer vision and image understanding*, 184:45–56, 2019.

[Lin *et al.*, 2021] Jiaying Lin, Zebang He, and Rynson WH Lau. Rich context aggregation with reflection prior for glass surface detection. In *CVPR*, pages 13415–13424, 2021.

[Liu *et al.*, 2023] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[Liu *et al.*, 2024] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *CVPR*, pages 26296–26306, 2024.

[Liu *et al.*, 2025] Jiaming Liu, Linghe Kong, and Guihai Chen. Improving sam for camouflaged object detection via dual stream adapters. In *ICCV*, pages 21906–21916, 2025.

[Luo *et al.*, 2024] 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 *CVPR*, pages 17169–17180, 2024.

[Lv *et al.*, 2021] 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 *CVPR*, pages 11591–11601, 2021.

[Margolin *et al.*, 2014] 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.

[Niu *et al.*, 2024] Yuzhen Niu, Lifen Yang, Rui Xu, Yuezhou Li, and Yuzhong Chen. Minet: Weakly-supervised camouflaged object detection through mutual interaction between region and edge cues. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 6316–6325, 2024.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

[Pang *et al.*, 2022] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, pages 2160–2170, 2022.

[Radford *et al.*, 2021] 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.

[Shou *et al.*, 2025] Peiyao Shou, Yixiu Liu, Wei Wang, Yaoqi Sun, Zhigao Zheng, Shangdong Zhu, and Chenggang Yan. Sdalsnet: Self-distilled attention localization and shift network for unsupervised camouflaged object detection. In *AAAI*, volume 39, pages 6914–6921, 2025.

[Siméoni *et al.*, 2023] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobeckỳ, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *CVPR*, pages 3176–3186, 2023.

[Tang *et al.*, 2024] Lv Tang, Peng-Tao Jiang, Zhi-Hao Shen, Hao Zhang, Jin-Wei Chen, and Bo Li. Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In *ACMMM*, pages 8805–8814, 2024.

[Yan *et al.*, 2025] Weiqi Yan, Lvhai Chen, Huaijia Kou, Shengchuan Zhang, Yan Zhang, and Liujuan Cao. Ucoddpl: Unsupervised camouflaged object detection via dynamic pseudo-label learning. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 30365–30375, 2025.

[Yang *et al.*, 2021] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *CVPR*, pages 4146–4155, 2021.

[Yin *et al.*, 2024] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. Camoformer: Masked separable attention for camouflaged object detection. *IEEE TPAMI*, 2024.

[Zhang and Wu, 2023] Yi Zhang and Chengyi Wu. Unsupervised camouflaged object segmentation as domain adaptation. In *ICCV*, pages 4334–4344, 2023.

[Zhang *et al.*, 2020] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, pages 12546–12555, 2020.