ST-SAM: SAM-Driven Self-Training Framework for Semi-Supervised Camouflaged Object Detection

Xihang Hu
Jilin University
Changchun, China
huxh24@mails.jlu.edu.cn

Fuming Sun
Dalian Minzu University
Dalian, China
sunfuming@dlnu.edu.cn

Jiazhe Liu
Jilin University
Changchun, China
liujz24@mails.jlu.edu.cn

Feilong Xu
Jilin University
Changchun, China
xufl23@mails.jlu.edu.cn

Xiaoli Zhang
Jilin University
Changchun, China
zhangxiaoli@jlu.edu.cn

ABSTRACT

Semi-supervised Camouflaged Object Detection (SSCOD) aims to reduce reliance on costly pixel-level annotations by leveraging limited annotated data and abundant unlabeled data. However, existing SSCOD methods based on Teacher-Student frameworks suffer from severe prediction bias and error propagation under scarce supervision, while their multi-network architectures incur high computational overhead and limited scalability. To overcome these limitations, we propose ST-SAM, a highly annotation-efficient yet concise framework that breaks away from conventional SSCOD constraints. Specifically, ST-SAM employs Self-Training strategy that dynamically filters and expands high-confidence pseudo-labels to enhance a single-model architecture, thereby fundamentally circumventing inter-model prediction bias. Furthermore, by transforming pseudo-labels into hybrid prompts containing domain-specific knowledge, ST-SAM effectively harnesses the Segment Anything Model's potential for specialized tasks to mitigate error accumulation in self-training. Experiments on COD benchmark datasets demonstrate that ST-SAM achieves state-of-the-art performance with only 1% labeled data, outperforming existing SSCOD methods and even matching fully supervised methods. Remarkably, ST-SAM requires training only a single network, without relying on specific models or loss functions. This work establishes a new paradigm for annotation-efficient SSCOD.

KEYWORDS

Semi-supervised, Self-Training, Camouflaged object detection, Segment Anything Model

1 INTRODUCTION

Camouflaged Object Detection (COD) aims to locate and segment camouflaged objects in complex scenes. The large size variations, complex contours, and similar textures of camouflaged objects render the COD task highly challenging and increase the need for pixel-level annotations. Nevertheless, the high cost of obtaining pixel-level annotated datasets significantly constrains the development and practical application of COD-related research.

To mitigate the strong reliance on large-scale annotated data, Semi-Supervised Learning (SSL) [16, 30, 40, 42, 47] has demonstrated outstanding performance in various fields. For instance, in Crowd Counting [45] and medical image segmentation [38], SSL has achieved remarkable success by leveraging limited annotated data and abundant unannotated data. However, integrating SSL with COD presents unique challenges due to the intrinsic feature-space entanglement between highly camouflaged foreground objects and their backgrounds. Current SSCOD approaches [10, 24] adopt Teacher-Student frameworks with multi-network co-training. These methods face two critical limitations when labeled data is scarce: (1) the complex multi-network architecture tends to overfit, resulting in poor generalization to unlabeled samples; and (2) increasing prediction discrepancies between networks lead to severe error propagation. These factors collectively create strong dependencies on annotation quantity, making it difficult to balance performance and annotation efficiency. Furthermore, the multi-network paradigm requires substantial design and training overhead, and its inherent inter-network compatibility requirements significantly constrain scalability. To overcome these limitations, we propose to transcend the conventional Teacher-Student framework and develop a more concise strategy that reduces reliance on annotated data.

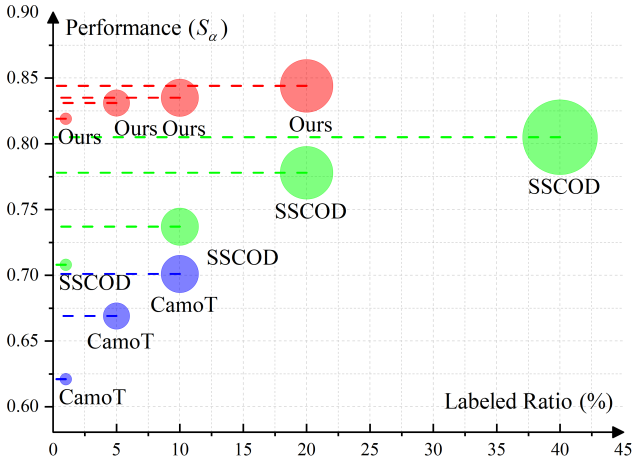

Figure 1: Performance comparison of semi-supervised COD methods on the CAMO dataset under different proportions of labeled samples.

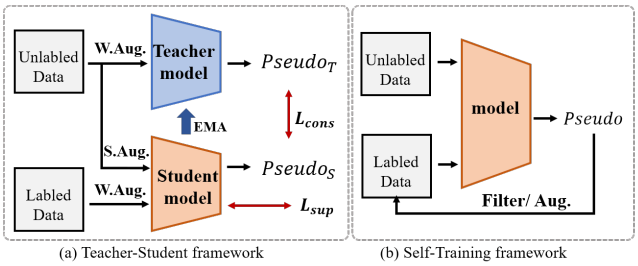

Figure 2: The comparison of typical SSCOD frameworks: (a) Teacher-Student framework; (b) Self-Training framework.

Unlike the teacher-student frameworks, Self-Training strategies [2, 37, 41] fundamentally expand the training set through pseudo-labels generated from unlabeled data to progressively refine the model (As shown in Fig. 2). This approach circumvents the severe prediction bias inherent in multi-network architectures when labeled data is scarce by reinforcing a single network. Moreover, its model-agnostic nature grants exceptional scalability, as it imposes no constraints on specific architectures or loss designs. However, conventional self-training often struggles with complex tasks. The inherent complexity of COD leads to noise-contaminated pseudo-labels during initial phases, where error accumulation becomes progressively amplified during subsequent training, ultimately causing performance collapse. To harness the advantages of self-training, we aim to effectively utilize these noisy pseudo-labels by enabling their self-correction, thereby enhancing the strategy's adaptability to complex tasks.

Recent advances in visual foundation models, particularly the Segment Anything Model (SAM) [23], have demonstrated remarkable segmentation capabilities across diverse scenarios, presenting a promising solution. Nevertheless, in specific domains like COD, the absence of domain-specific prior knowledge and the highly camouflaged nature of objects can easily mislead SAM. While fine-tuning SAM with camouflage-specific data could mitigate these issues, the prohibitive computational costs and substantial data requirements render this approach impractical. Contemporary research [6, 48] has made strides in adapting SAM for COD applications, though existing works predominantly focus on fully supervised [29, 43] and weakly supervised [4, 12] paradigms. Notably, the exploration of SAM for SSCOD is still blank. Therefore, this work bridges this critical gap by synergistically combining self-training with SAM's strengths, injecting domain knowledge into SAM without fine-tuning, enabling its adaptation to complex COD tasks. Simultaneously harnessing SAM's superior segmentation capability to rectify noisy pseudo-labels, effectively addressing error accumulation in self-training. Finally, we achieves efficient, low-sample-dependent SSCOD through an elegantly concise architecture.

To address the aforementioned challenges in existing SSCOD methods, we proposes ST-SAM, a novel SAM-driven self-training framework for SSCOD. Our framework comprises two key components: Entropy-based Dynamic Filtering strategy and Domain Prompt-guided Mutual Correction strategy. Specifically, we first conduct preliminary training on a standard COD model using limited annotated data to acquire fundamental domain knowledge. Initial pseudo-labels generated from unlabeled data inevitably contain substantial noise due to task complexity and the model's early-stage limitations. To mitigate this, our entropy-based dynamic filtering strategy performs instance-level screening and ranking of pseudo-labels to identify relatively reliable predictions, preventing error propagation from challenging samples during early training phases. Subsequently, we compute pixel-level confidence scores for the filtered samples and implement weighted learning to minimize the impact of uncertain prediction regions. Concurrently, we leverage SAM's exceptional segmentation capability to address the severe noise in early-stage pseudo-labels. Inspired by SAM's interactive prompt-based segmentation, we convert the entropy-weighted pseudo-labels into specific hybrid prompts. This innovative approach injects camouflage domain knowledge into SAM without requiring fine-tuning, enabling accurate segmentation of camouflaged targets. The framework then establishes mutual correction between SAM-generated pseudo-labels and entropy-weighted pseudo-labels to obtain high-confidence labels, effectively circumventing error accumulation. Through iterative expansion of the training set using these refined high-confidence pseudo-labels, the COD model progressively enhances its capabilities. Remarkably, ST-SAM achieves efficient end-to-end semi-supervised COD while requiring minimal labeled samples, and training of only a single network, significantly reducing computational overhead compared to existing multi-network approaches.

Our contributions can be summarized as follows:

- This paper for the first time investigate the potential of self-training strategies for SSCOD, constructing ST-SAM by combining the advantages of both self-training and SAM. This effectively resolves the strong dependence on annotation data in existing methods while maintaining strong scalability.
- We introduce EDF, which performs instance-level filtering and ranking of pseudo-labels, enabling the model to learn samples from easy to hard. Additionally, pixel-level weighting is applied to mitigate the negative impact of uncertain regions.
- We propose DPC, which converts weighted pseudo-labels into hybrid prompt information to inject domain knowledge into SAM. Leveraging SAM's powerful segmentation capability facilitates mutual correction of pseudo-labels.
- Extensive experiments results on 4 benchmark datasets demonstrate that ST-SAM achieves remarkable annotation efficiency, outperforming existing SSCOD methods with minimal labeled data (only 1%) and even attaining performance comparable to fully supervised learning methods.

2 RELATED WORK

2.1 Camouflaged Object Detection

In recent years, with the emergence of open-source datasets [9, 25, 27, 33, 34, 46] and the rapid advancements in deep learning, significant progress has been made in COD research. For instance, SINet [27] decomposes COD into a two-stage search-and-identification process. ZoomNet [32] mimics the human observation

mechanism for blurred images by proposing a hybrid-scale triplet network to handle scale variations in camouflaged objects. FRINet [39] rethinks COD from a frequency-domain perspective, leveraging heterogeneous architectures to exploit spectral information. He et al. [15] introduce text prompts as semantic cues for effective segmentation, while Hao et al. [11] design a unified ViT framework for both COD and SOD with image reconstruction assistance.

Despite their impressive performance, existing methods share a critical limitation: heavy reliance on annotated data. Given the high annotation cost of COD datasets, large-scale data collection remains difficult, severely hindering further advancements in COD. To address this, weakly supervised and semi-supervised COD approaches have gained attention. He et al. [14] propose a weakly supervised COD framework using scribble annotations to refine boundary details. Chen et al. [3] construct the first weakly supervised COD dataset with point annotations. WS-SAM [12] transforms weak labels into prompts for SAM [23] to generate mask annotations. Niu et al. [31] introduce a Mutual Interaction Network to mitigate ambiguity in scribble-based boundary information. Chen et al. [4] further improve supervision by integrating multiple weak labels as prompts. In semi-supervised COD, research remains scarce. Lai et al. [24] establish a baseline using consistency regularization to reduce pseudo-label noise, while Fu et al. [10] propose an ensemble learning method to aggregate knowledge from different models and training stages. Additionally, Zhang et al. [44] introduce Weakly-Semi-Supervised COD (WSSCOD), combining weak labels with partial manual annotations to enhance performance.

Semi-supervised COD holds great promise in alleviating annotation burdens while leveraging limited labeled data. However, current research is still in its early stages. To bridge this gap, we propose a more efficient semi-supervised COD framework that achieves superior performance while drastically reducing the demand for labeled samples.

2.2 Application of SAM in COD

The Segment Anything Model (SAM) has demonstrated remarkable performance and surprising zero-shot generalization capabilities in traditional segmentation tasks. However, recent studies [21, 35] have revealed that when directly applied to domain-specific tasks like Camouflaged Object Detection (COD), SAM suffers from significant performance degradation due to task complexity and domain gap caused by the lack of domain prior knowledge. While fine-tuning techniques could potentially adapt SAM to specific domains, they typically require massive computational resources and training samples, making them impractical for COD. To address these challenges, researchers have conducted a series of investigations. For the issue of SAM's lack of domain-specific knowledge, MCA-SAM [48] and SAM-Adapter [6] employ adapter techniques to reduce fine-tuning costs while enhancing SAM's domain-specific capabilities. TSP-SAM [20] incorporates long-range spatiotemporal information into SAM for Video COD (VCOD). DSAM [43] introduces depth modality information to improve SAM's performance in COD tasks. On another front, SAM's prompt-based interactive segmentation capability offers new opportunities for weakly supervised COD. Several studies have explored utilizing diverse weak supervision signals as prompts to leverage SAM's potential in COD tasks [4, 12]. However, the exploration of SAM's potential in semi-supervised COD remains completely unexplored.

In this work, we present the first investigation into SAM's potential for semi-supervised COD. We design a Domain Prompt-guided Mutual Correction strategy that can inject domain knowledge into SAM without requiring parameter fine-tuning, thereby enabling SAM to effectively learn from unlabeled samples.

3 METHODOLOGY

3.1 Overview Architecture

For semi-supervised camouflaged object detection (COD), the training set D consists of a labeled sample set $D_L = \left\{x_i^l, y_i\right\}_{i=1}^{M}$ and an unlabeled sample set $D_U = \left\{x_i^u\right\}_{i=1}^{N}$, where $x_i^{l/u}$ represents the input image and y_i denotes the corresponding mask label. To ensure higher annotation efficiency, usually $M << N$.

Fig. 3 illustrates the overall architecture of proposed ST-SAM, which comprises three components: EDF, DPC, and a standard COD network (implemented as a basic encoder-decoder architecture [44], hereafter referred to as Net). The framework operates as follows: First, we pretrain Net on D_L to acquire preliminary domain knowledge about camouflage patterns. The pretrained Net then processes D_U to generate initial pseudo-labels P_i^I. Given the network's limited initial capability, these pseudo-labels contain numerous unreliable samples. We therefore apply EDF strategy to perform dual-level filtering: at the image level to select relatively reliable samples, followed by pixel-level weighting to filter out uncertain regions, ultimately producing entropy-weighted pseudo-labels P_i^E. This process is dynamic - the selection criteria gradually relax during training to incorporate more pseudo-labels to match the scale of the training set. In early training stages, both the scarcity of labeled data and network immaturity mean P_i^E still lacks sufficient accuracy. To prevent error accumulation at this critical phase, we design the DPC that converts P_i^E into domain-knowledge-enhanced hybrid prompts to guide SAM's segmentation. The SAM-generated masks P_i^S are then compared and fused with P_i^E to produce reliable co-corrected pseudo-labels P_i^C. Finally, we augment D_L by incorporating P_i^C with their corresponding images. This iterative process continues until all samples in D_U are progressively incorporated into D_L, achieving effective utilization of unlabeled data. Through ST-SAM, we significantly reduce dependence on labeled data while achieving high-performance, efficient semi-supervised COD.

3.2 Entropy-based Dynamic Filtering strategy

Considering the difference in camouflage intensity among samples, we posit that a portion of the initial pseudo-labels P_i^I is relatively reliable, with this proportion progressively increasing as training advances and model capability improves. Nevertheless, due to the limited ability of Net during early training stages, even relatively reliable samples inevitably contain uncertain regions. To address this, we design EDF to obtain more trustworthy pseudo-labels.

Specifically, for the initial pseudo-labels $P_i^I, i = 1, ..., N$ obtained from the test set D_U, we first compute the local mean for each pixel using a window of size $\omega \times \omega$:

$$p_f = UF(Norm(P_i^I), \omega \times \omega). \quad (1)$$

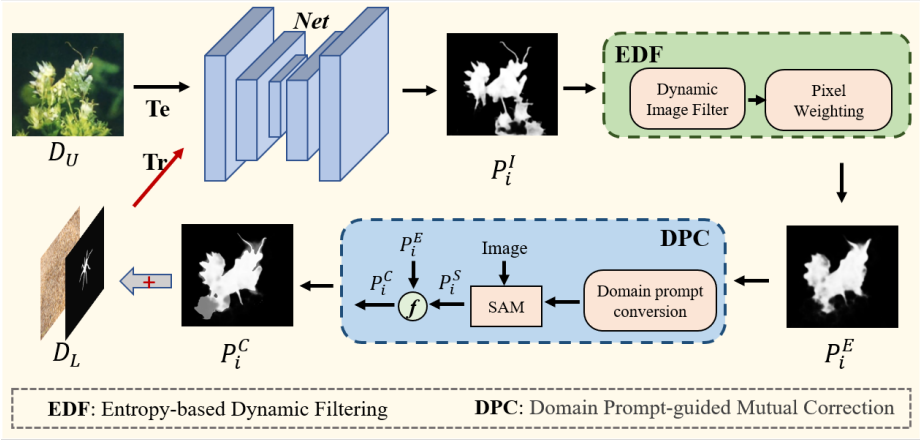

Figure 3: The overall flowchart of ST-SAM. Tr: The Net is preliminarily trained using labeled data D_L; Te: The unlabeled data D_U is processed through Net to obtain initial pseudo-labels P_i^I, which are then filtered using EDF to produce P_i^E. Finally, P_i^E serves as domain prompts for DPC to make predictions and perform mutual correction, yielding high-confidence pseudo-labels P_i^C for expansion.

Here, p_f represents the local foreground probability, $p_b = 1 - p_f$, $Norm(\cdot)$ denotes normalization, and $UF(\cdot)$ represents mean filtering, $\omega = 7$. Subsequently, the local entropy E_{local} is computed to quantify the uncertainty in the neighborhood of each pixel:

$$E_{local} = -p_f \log(p_f) - p_b \log(p_b). \tag{2}$$

Similarly, we compute the global foreground probability $\tilde{p_f} = \frac{1}{HW} \sum Norm(P_i^I)$ for the entire mask to derive the global entropy E_{global}, which quantifies the uncertainty of the mask prediction:

$$E_{global} = -\tilde{p_f} \log(\tilde{p_f}) - \tilde{p_b} \log(\tilde{p_b}). \tag{3}$$

Subsequently, an uncertainty metric u_α is defined to assess the reliability of the pseudo-labels:

$$u_\alpha = \frac{1}{N} \sum \mathbb{I}(E_{local} > E_{global} \times 0.5), \tag{4}$$

where N is the total number of pixels, $\mathbb{I}(\cdot)$ is the indicator function. The sample retention condition is:

$$Retain\ Sample\ If\ u_\alpha < \tau_\alpha,\ Else\ Discard. \tag{5}$$

where uncertainty threshold $\tau_\alpha = 0.3$.

Subsequently, for the retained masks, an entropy weight map is generated based on E_{local} to weight them, thereby further mitigating the negative impact of uncertain regions. This process ultimately yields the entropy-weighted pseudo-labels P_i^E, as described below:

$$P_i^E = P_i^I \cdot (0.5 + 0.5(1 - E_{local})^k). \tag{6}$$

Here, the entropy weight decay coefficient $k = 1$.

Finally, considering the limited performance of the Net in the early stages of training and the sample size $M << N$, it is necessary to dynamically expand the entropy-weighted pseudo-labels P_i^E. This allows the Net to learn samples from easy to difficult, progressively enhancing its ability to distinguish camouflaged objects and ultimately make accurate predictions for challenging samples. Specifically, the local entropy mean $\bar{E}_{local}$ of P_i^E is computed to rank the samples. Assuming the current training set size is x, the top x low-entropy samples $P_i^E, i = 1, ..., x$ are selected as candidates for expansion. As the training epochs increase, x will gradually increase, thereby achieving dynamic expansion.

3.3 Domain Prompt-guided Mutual Correction strategy

Due to the complex contours of camouflaged objects, obtaining accurate boundary predictions is highly challenging. Furthermore, limited by the scarcity of learning samples, the P_i^E obtained in the early stages of training often fail to provide sufficient supervisory information, which represents another difficulty in self-training strategies. While SAM has demonstrated robust capabilities in segmentation tasks, it struggles to adapt to specific domains. Therefore, we designed DPC to effectively leverage the potential of SAM in weakly supervised self-training.

Inspired by SAM's ability to perform interactive segmentation by incorporating user-provided prompts, we transform P_i^E into domain-specific prompt to guide SAM in locating and segmenting camouflaged objects. Common prompt types supported by SAM include points, boxes, and scribbles. However, to avoid introducing additional manual annotations, we primarily consider points and boxes as prompts.

An intuitive approach is to sample the geometric center of each contour in P_i^E to obtain multiple point prompts. This can effectively assist in locating multiple camouflaged objects. However, the reliability of this method depends on the completeness of P_i^E. As shown in the 1st row of Fig. 4, when the quality of P_i^E is poor and incomplete, the mask of a single object is divided into multiple disconnected regions, resulting in multiple point prompts. This misleads SAM into treating a single object as multiple objects, leading to incorrect segmentation. Similarly, generating multiple box prompts based on each region of the mask suffers from the same issue. Additionally, when the object is irregular, the sampled points

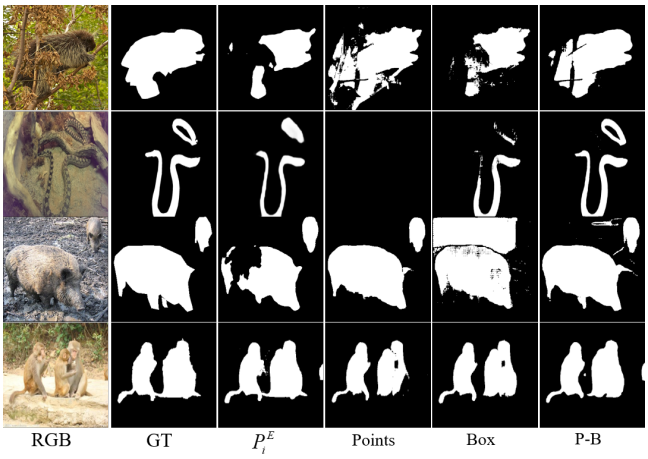

Figure 4: Pseudo-labels generated by SAM based on different prompts.

may fall outside the target region, providing erroneous guidance and thus failing to achieve segmentation (2nd row of Fig. 4). Calculating the minimum bounding rectangle for all regions of the mask as a single box prompt can maximize the integrity of the target. However, when multiple objects are sparsely distributed, the resulting prompt box may cover an excessively large area, thereby failing to provide effective localization guidance (3rd row of Fig. 4).

Therefore, we design a hybrid prompt strategy to transform P_i^E into reliable prompt information, effectively addressing the aforementioned issues. Specifically, for the multiple contour masks $Mask_i$ in P^E, we first filter out extremely small contour regions to prevent noise interference, resulting in $Mask_i^{'}, i = 1, ..., n$. Then, we extract the minimum bounding rectangle of $Mask_i^{'}$ as the box prompt $Prompt_B$:

$$Prompt_B = Mask_i^{'}[x_{\min}, y_{\min}, x_{\max}, y_{\max}]. \tag{7}$$

Next, we calculate the geometric center point $c_i(x_i, y_i)$ of $Mask_i^{'}$ and verify whether $c_i(x_i, y_i)$ lies inside $Mask_i^{'}$. If c_i falls outside $Mask_i^{'}$, we employ an axial search strategy $AxialS(\cdot)$ to search along the positive and negative directions of the major axis from c_i for the nearest point located inside $Mask_i^{'}$, which is designated as the safe center point $c_i^{'} = AxialS(c_i)$. This yields the point prompt set $Prompt_P$:

$$Prompt_P = \{c_i, IsInside(c_i) = 1\} \bigcup \left\{c_i^{'}, IsInside(c_i) = 0\right\}. \tag{8}$$

Then, $Prompt_B$ is combined with $Prompt_P$ to obtain the hybrid prompt $Prompt_{P-B}$. Through meticulously designed prompts, P_i^E excels in providing domain-specific knowledge to guide SAM, even under conditions such as incomplete masks, multiple camouflaged objects, and irregular contours, resulting in the SAM pseudo-labels P_i^S. Finally, P_i^E and P_i^S are fused in equal proportion, leveraging the shared effective information to mutually correct potential errors in the pseudo-labels, thereby obtaining reliable corrected pseudo-labels P_i^C, which are dynamically expanded into the training set:

$$P_i^C = fuse(P_i^E, P_i^S), P_i^S = SAM(Image, Prompt_{P-B}). \tag{9}$$

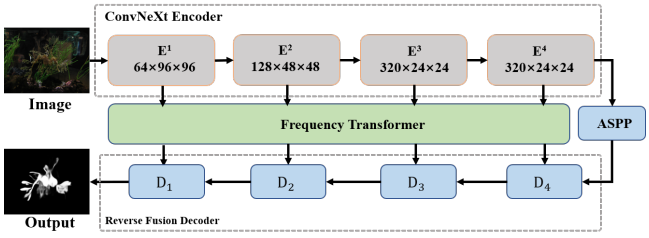

Figure 5: The overall architecture of COD network, consisting of an encoder, Frequency Transformer, ASPP, and decoder.

3.4 COD Model and Loss Function

The ST-SAM framework proposed in this paper is a flexible and scalable semi-supervised COD framework, in which the COD network is replaceable and does not constitute the primary contribution of this work. In our experiments, we adopted a classic U-Net COD network proposed in [44]. As shown in Fig. 5, the network employs ConvNeXt [26] as the encoder to extract multi-scale features $\{f_i^e\}_{i=1}^4$ from the input. Subsequently, the Frequency Transformer FT is utilized to capture fine details and deep semantics, yielding frequency-domain features $\{f_i^F\}_{i=1}^4$. For the high-level feature f_4^e, the ASPP [5] module is applied to obtain deep semantic representation feature f_4^E. Then, in the reverse fusion decoder, the reverse mask block amplifies the differences between the background and foreground, enabling the convergence of multi-level features. Following the U-Net architecture, f_4^E and f_i^F are fused progressively to produce the camouflaged object prediction.

In this paper, a hybrid loss function is employed to supervise the predictions of the camouflaged object detection network. The total loss $\mathcal{L}_{total}$ consists of three components: the structural loss $\mathcal{L}_S$, the Dice loss $\mathcal{L}_{dice}$, and the uncertainty-aware loss $\mathcal{L}_{UAL}$. Specifically, $\mathcal{L}_S = wbce + wiou$ is used to balance the global structure and local details of the predictions. $\mathcal{L}_{dice}$ is employed to learn the boundary information of the target, defined as follows:

$$\mathcal{L}_{dice} = 1 - \frac{2 \cdot \sum (Pred^p \cdot Target^p) + S}{\sum Pred^p + \sum + Target^p + S}, \tag{10}$$

where $p = 2$, and S is a smoothing term. $\mathcal{L}_{UAL}$ is designed to address noise and ambiguous boundaries, enhancing the robustness of the model. It is defined as:

$$\mathcal{L}_{UAL} = \lambda_{ual} \cdot \sum (Pred \cdot \log(Pred) + (1 - Pred) \cdot \log(1 - Pred)), \tag{11}$$

where λ_{ual} is a dynamic weighting coefficient that decreases with the learning rate. The overall loss $\mathcal{L}_{total}$ of the model is defined as:

$$\mathcal{L}_{total} = \mathcal{L}_s + \alpha \cdot \mathcal{L}_{dice} + \beta \cdot \mathcal{L}_{UAL}. \tag{12}$$

where $\alpha = 4$, $\beta = 2$.

4 EXPERIMENT

4.1 Experimental Setup

Datasets. In this paper, the proposed ST-SAM is evaluated on 4 COD benchmark datasets: CAMO [25], CHAMELEON [34], COD10K [9], and NC4K [27]. CAMO consists of 1,250 challenging

Table 1: Quantitative comparison between our method and other 11 SOTA methods on 4 benchmark datasets. '-' indicates the code or result is not available.The optimal and suboptimal results are represented in Red and Blue.

Methods	CAMO					CHAMELEON					COD10K					NC4K				
	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$S_\alpha \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
Fully-Supervised Methods																				
DSAM [43] MM '24	0.906	0.832	0.794	0.824	0.061	-	-	-	-	-	0.913	0.846	0.760	0.761	0.033	0.928	0.871	0.826	0.845	0.040
TJNet [36] AI '24	0.890	0.841	0.779	-	0.064	0.958	0.913	0.859	-	0.024	0.907	0.844	0.738	-	0.030	-	-	-	-	-
ICEG [13] ICLR '24	0.879	0.810	-	0.789	0.068	0.950	0.899	-	0.858	0.027	0.906	0.826	-	0.747	0.030	0.908	0.849	-	0.814	0.044
DINet [49] TMM '24	0.883	0.821	-	0.790	0.068	-	-	-	-	-	0.901	0.832	-	0.744	0.031	0.910	0.856	-	0.820	0.043
Weakly-Supervised Methods																				
WS-SAM [12] NeurIPS '23	0.818	0.759	-	0.742	0.092	0.897	0.824	-	0.777	0.046	0.878	0.803	-	0.719	0.038	0.886	0.829	-	0.802	0.052
PSCOD [3] ECCV '24	0.872	0.798	0.727	-	0.074	-	-	-	-	-	0.859	0.784	0.650	-	0.042	0.889	0.822	0.748	-	0.051
CRNet [14] AAAI '23	0.815	0.735	0.641	-	0.092	0.897	0.818	0.744	-	0.046	0.832	0.733	0.576	-	0.049	-	-	-	-	-
ProMaC [18] NeurIPS '24	0.846	0.767	-	0.725	0.090	0.899	0.833	-	0.790	0.044	0.876	0.805	-	0.716	0.042	-	-	-	-	-
GenSAM [17] AAAI '24	0.775	0.719	-	0.659	0.113	0.807	0.764	-	0.680	0.090	0.838	0.775	-	0.681	0.067	-	-	-	-	-
Semi-Supervised Methods																				
CamoTeacher [24] -1%	0.669	0.621	0.456	0.545	0.136	0.714	0.652	0.476	0.558	0.093	0.788	0.699	0.517	0.582	0.062	0.779	0.718	0.599	0.675	0.090
CamoTeacher -5%	0.711	0.669	0.523	0.601	0.122	0.785	0.729	0.587	0.656	0.070	0.827	0.745	0.583	0.644	0.050	0.834	0.777	0.677	0.739	0.071
CamoTeacher -10% ECCV '24	0.742	0.701	0.560	0.635	0.112	0.813	0.756	0.617	0.684	0.065	0.836	0.759	0.594	0.652	0.049	0.842	0.791	0.687	0.746	0.068
SSCOD [10] -1%	0.804	0.708	0.583	0.653	0.110	0.771	0.683	0.574	0.629	0.063	0.805	0.725	0.537	0.578	0.057	0.844	0.767	0.652	0.700	0.073
SSCOD -10%	0.806	0.737	0.638	0.708	0.094	0.878	0.805	0.707	0.751	0.047	0.852	0.779	0.639	0.676	0.042	0.868	0.808	0.729	0.775	0.059
SSCOD -20% MM '24	0.844	0.778	0.704	0.767	0.078	0.906	0.834	0.761	0.802	0.042	0.864	0.791	0.662	0.699	0.039	0.882	0.821	0.750	0.795	0.055
Ours -1%	0.879	0.819	0.753	0.804	0.070	0.920	0.848	0.778	0.815	0.038	0.884	0.824	0.713	0.737	0.032	0.907	0.855	0.795	0.830	0.043
Ours -5%	0.886	0.831	0.761	0.807	0.066	0.921	0.855	0.776	0.798	0.036	0.871	0.834	0.709	0.716	0.031	0.909	0.870	0.803	0.823	0.038
Ours -10%	0.886	0.835	0.778	0.818	0.063	0.919	0.850	0.770	0.792	0.036	0.882	0.837	0.723	0.729	0.029	0.911	0.868	0.807	0.830	0.038
Ours -20%	0.890	0.844	0.779	0.809	0.058	0.926	0.876	0.804	0.818	0.032	0.874	0.837	0.713	0.717	0.030	0.911	0.874	0.807	0.824	0.037

camouflaged images, with 1,000 images for training and 250 images for testing. CHAMELEON contains 76 images of camouflaged animals. COD10K includes 5,066 camouflaged images, with 3,040 images for training and 2,026 images for testing. NC4K comprises 4,121 camouflaged images collected from the web. Following the same dataset partitioning as in other studies [10, 24], we construct the training set using 1,000 samples from CAMO and 3,040 samples from COD10K. Then, the entire NC4K and CHAMELEON datasets, along with the remaining samples from CAMO and COD10K, are used as the test set to evaluate the performance of ST-SAM and competing models. For the training set, we adhere to the semi-supervised learning partitioning strategy [10, 24], randomly sampling 1%, 5%, 10%, and 20% of the labeled data from the training set as the labeled sample set D_L, with the remaining portion serving as the unlabeled sample set D_U.

Evaluation Metrics. Following previous works [10, 24], we adopt five widely used evaluation metrics to assess the performance of our model. These include E-measure (E_ξ) [8], S-measure (S_α) [7], Weighted F-score (F_β^ω) [28], F-measure (F_β) [1], and Mean Absolute Error (MAE). Here, E_ξ and F_β represent adaptive values.

Implementation Details. To enhance the robustness, we employ data augmentation strategies such as random flipping, rotation, and boundary cropping on the training images to prevent overfitting. During both the training and testing phases, the input image size is resized to 384×384. The COD network initializes the encoder parameters using ConvNeXt-B [26], while SAM [23] adopts the parameters and settings of vit-h. We utilize the Adam optimizer [22] to train our network, with a batch size set to 8. The initial learning rate is set to 1e-7, linearly warmed up to 1e-4 within 10 epochs, and then cosine annealed to 1e-7 over 150 epochs to complete the preliminary training of the COD network. Subsequently, the learning rate is reset to 1e-4 and dynamically expanded and trained using the proposed two strategies every 20 epochs, with cosine annealing to 1e-7 over 300 epochs to complete the training.

4.2 Comparison with state-of-the-art methods

To validate the effectiveness of ST-SAM, we compared it with 11 SOTA methods, including semi-supervised methods: CamoTeacher [24], SSCOD [10]; weakly supervised methods: WS-SAM [12], PSCOD [3], CRNet [14], ProMaC [18], GenSAM [17]; and fully supervised methods: DSAM [43], TJNet [36], DINet [49], ICEG [13]. To ensure a fair comparison, the results of these methods were either obtained from publicly available data or generated by training their source code. The comparison results are as follows:

Quantitative Comparison. The quantitative comparison results of the proposed algorithm and the 11 other methods on the four datasets are shown in Table 1. From the results, it can be observed that, compared to the existing two semi-supervised methods, ST-SAM achieves the best performance under the same amount of labeled samples. Additionally, under the condition of scarce labeled samples (1%), the performance of the other two algorithms significantly declines, while ST-SAM maintains excellent performance, even outperforming the versions of existing methods with 20% labeled samples. This demonstrates that ST-SAM effectively addresses the strong dependency on labeled samples in semi-supervised COD. Furthermore, we compared ST-SAM with weakly supervised and fully supervised methods. It can be seen that ST-SAM outperforms SOTA weakly supervised COD methods and even achieves competitive performance compared to fully supervised methods. These experiments validate the effectiveness of ST-SAM and highlight the great potential of semi-supervised COD strategies.

Qualitative Comparison. To visually demonstrate the capabilities of our method, Fig. 6 presents a visual comparison of ST-SAM

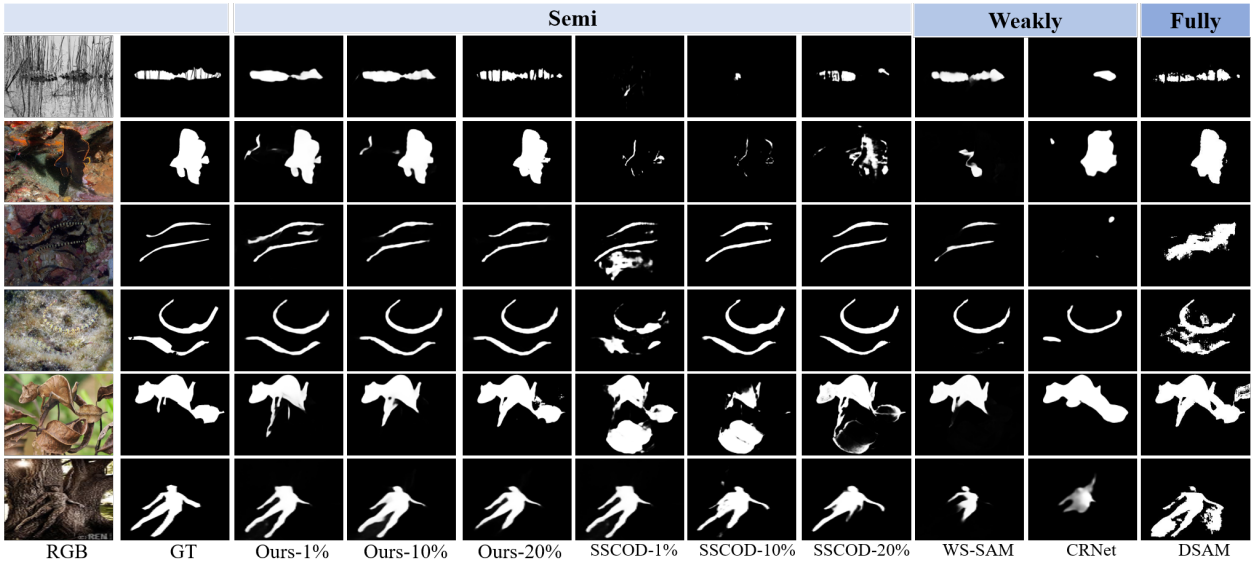

Figure 6: Qualitative comparison results between ST-SAM and SOTA COD models.

with other semi-supervised, weakly supervised, and fully supervised SOTA algorithms in some representative camouflaged scenarios. From the results, it is evident that ST-SAM, with only 1% labeled samples, can effectively detect and segment camouflaged objects in most scenarios, even outperforming semi-supervised methods with 20% labeled samples and weakly supervised methods. When the labeled samples are increased to 20%, the performance of ST-SAM further improves, enabling it to more effectively handle complex scenarios such as occlusions (Row 1) and small objects (Rows 3, 4), significantly reducing misjudgments (Row 2) and producing more complete segmentation results (Rows 5, 6). ST-SAM can match or even surpass fully supervised methods.

4.3 Ablation Study

We conducted experiments to demonstrate the effectiveness of ST-SAM. All experiments were performed with 1% labeled samples. Due to space limitations, the analysis of complexity, failure cases, ablation of hyper-parameters, and some details are provided in supplementary materials.

Table 2: Ablation experiments on the effectiveness of each component, the best results are marked in Red.

ID	Variants	COD10K				NC4K			
		$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
1	Baseline	0.793	0.517	0.569	0.054	0.834	0.611	0.693	0.077
2	Self-Training	0.807	0.535	0.585	0.057	0.830	0.628	0.688	0.078
3	+EDF	0.868	0.676	0.708	0.037	0.883	0.748	0.804	0.056
4	+DPC	0.856	0.655	0.684	0.038	0.879	0.740	0.778	0.052
5	+EDF+DPC	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043

Ablation on EDF and DPC. To validate the effectiveness of the proposed EDF and DPC, the following experiments were conducted: (1) Learning only D_L as the baseline; (2) A classic self-training strategy, where pseudo-labels are obtained from D_U using the COD network to directly expand D_L; (3) Incorporating the EDF strategy into the COD network; (4) Incorporating the DPC strategy into the COD network; (5) Combining both the EDF and DPC strategies in the COD network, i.e., the ST-SAM framework.

Taking F_β^ω of COD10K as an example, Table 2 shows that due to the complexity of camouflaged scenes, directly applying the self-training strategy fails to unlock the potential of D_U, providing almost no performance gain. The EDF strategy reduces error accumulation through global and local filtering as well as dynamic expansion, improving performance by 15.9% compared to the baseline. The DPC strategy compensates for the shortcomings of the initial model by leveraging domain knowledge to drive SAM, improving performance by 13.8% compared to the baseline, thus proving their effectiveness. Furthermore, the combination of both strategies further enhances performance to 19.6%, demonstrating the rationality of the ST-SAM framework design.

Table 3: Ablation experiments on the effectiveness of EDF, the results marked in Red indicate the best performance.

ID	Variants	COD10K				NC4K			
		$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
1	One-shot	0.817	0.621	0.649	0.047	0.872	0.737	0.775	0.055
2	Equal Ratio	0.850	0.682	0.703	0.036	0.891	0.778	0.812	0.045
3	Epoch-Dynamic	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043
4	H→L Entropy	0.788	0.525	0.590	0.059	0.813	0.612	0.685	0.078
5	Random Select	0.883	0.710	0.737	0.032	0.902	0.789	0.830	0.044
6	L→H Entropy	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043

Ablation of EDF. We conducted the following experiments to validate the rationality of the EDF, with the results shown in Table 3, using F_β^ω of COD10K as an example for comparison.

Regarding the dynamic expansion strategy for pseudo-labels: (1) Expanding all qualified samples at once for learning; (2) Expanding an equal proportion (20%) of samples at each step for learning; (3) Dynamically expanding the proportion of samples as training epochs progress. From the results, it can be observed that due to

the limited ability of the early model to learn difficult samples, expanding all samples at once yields the worst performance at only 62.1%. Expanding an equal proportion effectively mitigates this issue, improving performance by 6.1%. The dynamic expansion strategy gradually increases the proportion of samples as the model's capability improves, achieving a 9.2% improvement.

Regarding the sample learning order: (4) Learning high-entropy samples first, followed by low-entropy samples; (5) Randomly selecting samples from the qualified candidates; (6) Learning low-entropy samples first, followed by high-entropy samples. From the results, it can be seen that although for many semi-supervised self-training tasks, prioritizing high-entropy samples can better learn complex patterns and enhance generalization, making it a more common choice, for COD tasks, the complexity of camouflaged scenes leads to performance collapse (only 52.5%) when learning high-entropy samples too early without sufficient reliable information, which is far lower than random selection. Gradually enhancing the model by expanding from low-entropy to high-entropy samples brings further improvement compared to random selection.

Table 4: Ablation experiments on the effectiveness of DPC, the results marked in Red indicate the best performance.

ID	Variants	COD10K				NC4K			
		$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
1	Points	0.830	0.645	0.681	0.050	0.872	0.742	0.794	0.061
2	Box	0.876	0.701	0.731	0.035	0.905	0.795	0.827	0.044
3	P-B	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043
4	Intersect	0.821	0.581	0.649	0.049	0.826	0.639	0.740	0.077
5	Union	0.800	0.622	0.626	0.048	0.871	0.752	0.762	0.049
6	Ratio	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043

Ablation of DPC. We conducted the following experiments to validate the rationality of the DPC, with the results shown in Table 4, using F_β^ω of COD10K as an example for comparison.

Regarding different prompt types: (1) Multiple point prompts; (2) A single box prompt; (3) A hybrid prompt combining multiple points and a single box. From the results, it can be observed that due to the complex morphology of camouflaged objects, converting limited-quality pseudo-labels into point prompts provides misleading guidance, resulting in incomplete object segmentation and poor performance at only 64.5%. In contrast, the box prompt, leveraging region information, effectively complements missing areas, achieving a result of 70.1%. The hybrid prompt adopted in this paper combines the advantages of both, effectively addressing scenarios with multiple objects and incomplete regions, further improving performance to 71.3%.

Regarding the fusion method of P_i^E and P_i^S: (4) Taking the intersection of the two labels; (5) Taking the union of the two labels; (6) Fusing the two labels in equal proportion. From the results, it can be seen that taking the intersection of pseudo-labels filters out erroneous regions but leads to false negatives. For complex COD scenes, the lack of sufficient guidance can easily cause performance collapse, achieving only 58.1%. Taking the union of pseudo-labels ensures completeness but introduces false positives, bringing in many errors, achieving 62.2%. In contrast, proportionally fusing the pseudo-labels maximizes information retention while suppressing potential errors, achieving the optimal performance of 71.3%.

Table 5: Validation of the scalability of ST-SAM, the results marked in Red indicate the best performance.

ID	Variants	COD10K				NC4K			
		$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$	$E_\xi \uparrow$	$F_\beta^\omega \uparrow$	$F_\beta \uparrow$	$\mathcal{M} \downarrow$
1	ST-SAM(Light)	0.852	0.653	0.694	0.040	0.884	0.741	0.796	0.051
2	ST-SAM(BCE-IoU)	0.864	0.698	0.717	0.034	0.904	0.798	0.826	0.040
3	SSCOD	0.805	0.537	0.578	0.057	0.844	0.652	0.700	0.073
4	ST-SAM	0.884	0.713	0.737	0.032	0.907	0.795	0.830	0.043

Scalability Verification. We conducted the following experiments to verify that ST-SAM does not rely on a specific network or loss function, with the results shown in Table 5, using F_β^ω of COD10K as an example for comparison.

(1) Replacing the COD network with PRNet [19], a lightweight network designed for fully supervised COD; (2) Replacing the loss function with BCE-IoU loss. From the results, it can be observed that replacing the network with a lightweight one leads to a performance loss of 6%, but still achieves an 11.6% improvement compared to SSCOD. Meanwhile, replacing the loss function with BCE-IoU results in only a slight performance loss of 1.5%, while maintaining a 16.1% improvement over SSCOD. In conclusion, ST-SAM retains its low manual annotations dependency when different COD networks and loss functions are replaced, demonstrating its flexibility to adapt to various application scenarios and strong scalability.

5 LIMITATION AND PROSPECT

Although the Self-Training SSCOD successfully circumvents dependency on large-scale annotated data, its iterative expansion approach inevitably leads to noise accumulation. This phenomenon tends to obscure crucial information contained in the annotated samples, ultimately resulting in performance plateaus. Our analysis suggests this issue shares fundamental similarities with the catastrophic forgetting problem observed in continual learning tasks, where models struggle to retain previously learned knowledge when acquiring new tasks. To address these challenges, our future work will focus on incorporating regularization-based continual learning and dynamic replay strategies. These approaches aim to better preserve essential features from annotated data while progressively integrating new knowledge from pseudo-labels, thereby further unlocking the potential of SSCOD frameworks.

6 CONCLUSION

In this paper, we explore the potential of SAM for self-training semi-supervised camouflaged object detection for the first time and propose ST-SAM. To move beyond the limitations of complex model design details and achieve stronger scalability and generalization, we introduce EDF to prevent error accumulation during self-training. To meet the high demand for initial models in self-training, we propose DPC to enable SAM to unleash its potential on specific tasks through hybrid prompts containing domain knowledge. Experimental results on four datasets demonstrate that ST-SAM achieves performance comparable to fully supervised and semi-supervised methods while significantly reducing the requirement for labeled samples to 1%. We believe that ST-SAM will inject new vitality into research on SSCOD.

REFERENCES

[1] Radhakrishna Achanta, Sheila S. Hemami, Francisco J. Estrada, and Sabine Süsstrunk. 2009. Frequency-tuned salient region detection. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 1597–1604. https://doi.org/10.1109/CVPR.2009.5206596

[2] Massih-Reza Amini, Vasilii Feofanov, Loïc Pauletto, Lies Hadjadj, Emilie Devijver, and Yury Maximov. 2025. Self-training: A survey. *Neurocomputing* 616 (2025), 128904. https://doi.org/10.1016/J.NEUCOM.2024.128904

[3] Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. 2024. Just a Hint: Point-Supervised Camouflaged Object Detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXV (Lecture Notes in Computer Science, Vol. 15093)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 332–348. https://doi.org/10.1007/978-3-031-72761-0_19

[4] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. 2024. SAM-COD: SAM-Guided Unified Framework for Weakly-Supervised Camouflaged Object Detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XXXV (Lecture Notes in Computer Science, Vol. 15093)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 315–331. https://doi.org/10.1007/978-3-031-72761-0_18

[5] Liang-Chieh Chen, George Papandreou, Florian Schroff, and Hartwig Adam. 2017. Rethinking Atrous Convolution for Semantic Image Segmentation. *CoRR* abs/1706.05587 (2017). arXiv:1706.05587 http://arxiv.org/abs/1706.05587

[6] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. 2023. SAM-Adapter: Adapting Segment Anything in Underperformed Scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*. IEEE, 3359–3367. https://doi.org/10.1109/ICCVW60793.2023.00361

[7] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-Measure: A New Way to Evaluate Foreground Maps. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*. IEEE Computer Society, 4558–4567. https://doi.org/10.1109/ICCV.2017.487

[8] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jérôme Lang (Ed.). ijcai.org, 698–704. https://doi.org/10.24963/IJCAI.2018/97

[9] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. 2020. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2777–2787.

[10] Yuanbin Fu, Jie Ying, Houlei Lv, and Xiaojie Guo. 2024. Semi-supervised Camouflaged Object Detection from Noisy Data. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 4766–4775. https://doi.org/10.1145/3664647.3680645

[11] Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jing-Yu Yang. 2025. A Simple Yet Effective Network Based on Vision Transformer for Camouflaged Object and Salient Object Detection. *IEEE Trans. Image Process.* 34 (2025), 608–622. https://doi.org/10.1109/TIP.2025.3528347

[12] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. 2023. Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. *Advances in Neural Information Processing Systems* 36 (2023), 30726–30737.

[13] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Chenyu You, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. 2024. Strategic Preys Make Acute Predators: Enhancing Camouflaged Object Detectors by Generating Camouflaged Objects. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net. https://openreview.net/forum?id=hywpSoHwgX

[14] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W. H. Lau. 2023. Weakly-Supervised Camouflaged Object Detection with Scribble Annotations. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, Brian Williams, Yiling Chen, and Jennifer Neville (Eds.). AAAI Press, 781–789. https://doi.org/10.1609/AAAI.V37I1.25156

[15] Zhentao He, Changqun Xia, Shengye Qiao, and Jia Li. 2024. Text-prompt Camouflaged Instance Segmentation with Graduated Camouflage Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 5584–5593. https://doi.org/10.1145/3664647.3681132

[16] Jie Hu, Chen Chen, Liujuan Cao, Shengchuan Zhang, Annan Shu, Guannan Jiang, and Rongrong Ji. 2023. Pseudo-label Alignment for Semi-supervised Instance Segmentation. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*. IEEE, 16291–16301. https://doi.org/10.1109/ICCV51070.2023.01497

[17] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. 2024. Relax Image-Specific Prompt Requirement in SAM: A Single Generic Prompt for Segmenting Camouflaged Objects. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 12511–12518. https://doi.org/10.1609/AAAI.V38I11.29144

[18] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. 2024. Leveraging Hallucinations to Reduce Manual Prompt Dependency in Promptable Segmentation. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (Eds.). http://papers.nips.cc/paper_files/paper/2024/hash/c1e1ad233411e25b54bb5df3a0576c2c-Abstract-Conference.html

[19] Xihang Hu, Xiaoli Zhang, Fasheng Wang, Jing Sun, and Fuming Sun. 2024. Efficient Camouflaged Object Detection Network Based on Global Localization Perception and Local Guidance Refinement. *IEEE Trans. Circuits Syst. Video Technol.* 34, 7 (2024), 5452–5465. https://doi.org/10.1109/TCSVT.2023.3349209

[20] Wenjun Hui, Zhenfeng Zhu, Shuai Zheng, and Yao Zhao. 2024. Endow SAM with Keen Eyes: Temporal-Spatial Prompt Learning for Video Camouflaged Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*. IEEE, 19058–19067. https://doi.org/10.1109/CVPR52733.2024.01803

[21] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Bowen Zhou, Ming-Ming Cheng, and Luc Van Gool. 2023. SAM struggles in concealed scenes - empirical study on "Segment Anything". *Sci. China Inf. Sci.* 66, 12 (2023). https://doi.org/10.1007/S11432-023-3881-X

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1412.6980

[23] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF international conference on computer vision*. 4015–4026.

[24] Xunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songan Zhang, and Rongrong Ji. 2024. CamoTeacher: Dual-Rotation Consistency Learning for Semi-supervised Camouflaged Object Detection. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part XLV (Lecture Notes in Computer Science, Vol. 15103)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 438–455. https://doi.org/10.1007/978-3-031-72995-9_25

[25] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. 2019. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.* 184 (2019), 45–56. https://doi.org/10.1016/J.CVIU.2019.04.006

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. 2022. A ConvNet for the 2020s. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 11966–11976. https://doi.org/10.1109/CVPR52688.2022.01167

[27] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. 2021. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 11591–11601. https://doi.org/10.1109/CVPR46437.2021.01142

[28] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to Evaluate Foreground Maps. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*. IEEE Computer Society, 248–255. https://doi.org/10.1109/CVPR.2014.39

[29] Muhammad Nawfal Meeran, Gokul Adethya T, and Bhanu Pratyush Mantha. 2024. SAM-PM: Enhancing Video Camouflaged Object Detection using Spatio-Temporal Attention. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024 - Workshops, Seattle, WA, USA, June 17-18, 2024*. IEEE, 1857–1866. https://doi.org/10.1109/CVPRW63382.2024.00192

[30] Peng Mi, Jianghang Lin, Yiyi Zhou, Yunhang Shen, Gen Luo, Xiaoshuai Sun, Liujuan Cao, Rongrong Fu, Qiang Xu, and Rongrong Ji. 2022. Active Teacher for Semi-Supervised Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 14462–14471. https://doi.org/10.1109/CVPR52688.2022.01408

[31] Yuzhen Niu, Lifen Yang, Rui Xu, Yuezhou Li, and Yuzhong Chen. 2024. MiNet: Weakly-Supervised Camouflaged Object Detection through Mutual Interaction between Region and Edge Cues. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 6316–6325. https://doi.org/10.1145/3664647.3680891

[32] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. 2022. Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2150–2160. https://doi.org/10.1109/CVPR52688.2022.00220

[33] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. 2025. Open-Vocabulary Camouflaged Object Segmentation. In *Computer Vision – ECCV 2024*, Aleš Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer Nature Switzerland, Cham, 476–495.

[34] Przemysław Skurowski, Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. 2018. Animal camouflage analysis: Chameleon database. *Unpublished manuscript* 2, 6 (2018), 7.

[35] Lv Tang, Haoke Xiao, and Bo Li. 2023. Can SAM Segment Anything? When SAM Meets Camouflaged Object Detection. *CoRR* abs/2304.04709 (2023). https://doi.org/10.48550/ARXIV.2304.04709 arXiv:2304.04709

[36] Zhe Tang, Jing Tang, Dengpeng Zou, Junyi Rao, and Fang Qi. 2024. Two guidance joint network based on coarse map and edge map for camouflaged object detection. *Appl. Intell.* 54, 15-16 (2024), 7531–7544. https://doi.org/10.1007/S10489-024-05559-Y

[37] Bingyang Wang, Tanlin Li, Jiannan Wu, Yi Jiang, Huchuan Lu, and You He. 2023. A Simple Baseline for Open-World Tracking via Self-training. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 2765–2774. https://doi.org/10.1145/3581783.3611695

[38] Song Wu, Xiaoyu Wei, Xinyue Chen, Yazhou Ren, Jing He, and Xiaorong Pu. 2024. Cross-View Mutual Learning for Semi-Supervised Medical Image Segmentation. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 9253–9261. https://doi.org/10.1145/3664647.3680699

[39] Chenxi Xie, Changqun Xia, Tianshu Yu, and Jia Li. 2023. Frequency Representation Integration for Camouflaged Object Detection. In *Proceedings of the 31st ACM International Conference on Multimedia, MM 2023, Ottawa, ON, Canada, 29 October 2023- 3 November 2023*, Abdulmotaleb El-Saddik, Tao Mei, Rita Cucchiara, Marco Bertini, Diana Patricia Tobon Vallejo, Pradeep K. Atrey, and M. Shamim Hossain (Eds.). ACM, 1789–1797. https://doi.org/10.1145/3581783.3611773

[40] Lihe Yang, Zhen Zhao, and Hengshuang Zhao. 2025. UniMatch V2: Pushing the Limit of Semi-Supervised Semantic Segmentation. *IEEE Trans. Pattern Anal. Mach. Intell.* 47, 4 (2025), 3031–3048. https://doi.org/10.1109/TPAMI.2025.3528453

[41] Lihe Yang, Wei Zhuo, Lei Qi, Yinghuan Shi, and Yang Gao. 2022. ST++: Make Self-trainingWork Better for Semi-supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 4258–4267. https://doi.org/10.1109/CVPR52688.2022.00423

[42] Mingzhao Yang, Shangchao Su, Bin Li, and Xiangyang Xue. 2024. Exploring One-Shot Semi-supervised Federated Learning with Pre-trained Diffusion Models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan (Eds.). AAAI Press, 16325–16333. https://doi.org/10.1609/AAAI.V38I15.29568

[43] Zhenni Yu, Xiaoqin Zhang, Li Zhao, Yi Bin, and Guobao Xiao. 2024. Exploring Deeper! Segment Anything Model with Depth Perception for Camouflaged Object Detection. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 4322–4330. https://doi.org/10.1145/3664647.3681119

[44] Jin Zhang, Ruiheng Zhang, Yanjiao Shi, Zhe Cao, Nian Liu, and Fahad Shahbaz Khan. 2024. Learning Camouflaged Object Detection from Noisy Pseudo Label. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I (Lecture Notes in Computer Science, Vol. 15059)*, Ales Leonardis, Elisa Ricci, Stefan Roth, Olga Russakovsky, Torsten Sattler, and Gül Varol (Eds.). Springer, 158–174. https://doi.org/10.1007/978-3-031-73232-4_9

[45] Shiwei Zhang, Wei Ke, Shuai Liu, Xiaopeng Hong, and Tong Zhang. 2024. Boosting Semi-supervised Crowd Counting with Scale-based Active Learning. In *Proceedings of the 32nd ACM International Conference on Multimedia, MM 2024, Melbourne, VIC, Australia, 28 October 2024 - 1 November 2024*, Jianfei Cai, Mohan S. Kankanhalli, Balakrishnan Prabhakaran, Susanne Boll, Ramanathan Subramanian, Liang Zheng, Vivek K. Singh, Pablo César, Lexing Xie, and Dong Xu (Eds.). ACM, 8681–8690. https://doi.org/10.1145/3664647.3680976

[46] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. 2025. Referring Camouflaged Object Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 47, 5 (2025), 3597–3610. https://doi.org/10.1109/TPAMI.2025.3532440

[47] Zhen Zhao, Lihe Yang, Sifan Long, Jimin Pi, Luping Zhou, and Jingdong Wang. 2023. Augmentation Matters: A Simple-Yet-Effective Approach to Semi-Supervised Semantic Segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*. IEEE, 11350–11359. https://doi.org/10.1109/CVPR52729.2023.01092

[48] Ke Zhou, Zhongwei Qiu, and Dongmei Fu. 2024. Multi-scale contrastive adaptor learning for segmenting anything in underperformed scenes. *Neurocomputing* 606 (2024), 128395. https://doi.org/10.1016/J.NEUCOM.2024.128395

[49] Xiaofei Zhou, Zhicong Wu, and Runmin Cong. 2024. Decoupling and Integration Network for Camouflaged Object Detection. *IEEE Trans. Multim.* 26 (2024), 7114–7129. https://doi.org/10.1109/TMM.2024.3360710