SAM-COD: SAM-guided Unified Framework for Weakly-Supervised Camouflaged Object Detection

Huafeng Chen[1,3], Pengxu Wei[2,4], Guangqian Guo[1,3], and Shan Gao[1,3⋆]

[1] Northwestern Polytechnical University, [2] Sun Yat-Sen University, [3] National Key Laboratory of Unmanned Aerial Vehicle Technology, [4] Peng Cheng Laboratory, China
{{chf, guogq21}@mail, gaoshan@}.nwpu.edu.cn, weipx3@mail.sysu.edu.cn

Abstract. Most Camouflaged Object Detection (COD) methods heavily rely on mask annotations, which are time-consuming and labor-intensive to acquire. Existing weakly-supervised COD approaches exhibit significantly inferior performance compared to fully-supervised methods and struggle to simultaneously support all the existing types of camouflaged object labels, including scribbles, bounding boxes, and points. Even for Segment Anything Model (SAM), it is still problematic to handle the weakly-supervised COD and it typically encounters challenges of prompt compatibility of the scribble labels, extreme response, semantically erroneous response, and unstable feature representations, producing unsatisfactory results in camouflaged scenes. To mitigate these issues, we propose a unified COD framework in this paper, termed SAM-COD, which is capable of supporting arbitrary weakly-supervised labels. Our SAM-COD employs a prompt adapter to handle scribbles as prompts based on SAM. Meanwhile, we introduce response filter and semantic matcher modules to improve the quality of the masks obtained by SAM under COD prompts. To alleviate the negative impacts of inaccurate mask predictions, a new strategy of prompt-adaptive knowledge distillation is utilized to ensure a reliable feature representation. To validate the effectiveness of our approach, we have conducted extensive empirical experiments on three mainstream COD benchmarks. The results demonstrate the superiority of our method against state-of-the-art weakly-supervised and even fully-supervised methods.

Keywords: Weakly-Supervised Camouflaged Object Detection · SAM · Prompt Adapter · Prompt-Adaptive Knowledge Distillation

1 Introduction

Camouflaged Object Detection (COD) aims to detect concealed objects from various backgrounds [4, 8, 25, 32, 33], which have imperceptible visual appearances with extremely high similarity to the environment. It holds great promise for practical applications, *e.g.*, species discovery [9, 10, 12, 17, 18, 20], medical

⋆ Corresponding author

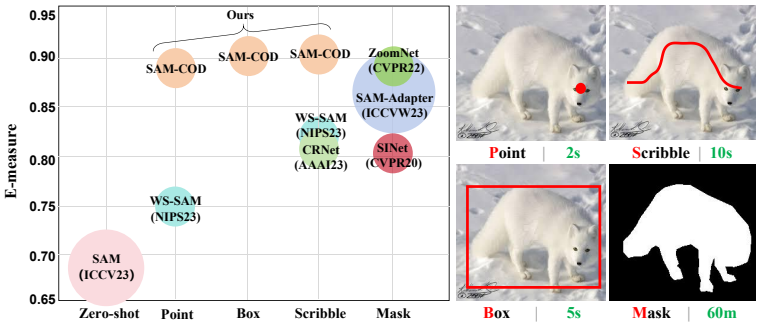

Fig. 1: Comparison of COD methods for different granularity labels. A larger circle denotes a higher-parameter model. SAM-COD is capable of handling three different labels for camouflaged objects. It achieves the highest performance under the weakly-supervised learning setting and even outperforms the fully supervised ZoomNet [25].

image segmentation, and animal tracking [7]. Considering that mask annotations as fully-supervised learning labels [8] are not always available for the time-consuming and laborious cost, *e.g.*, 60 minutes per image [8], weakly-supervised labels are promising as an attractive alternative, including scribble (∼10 seconds) [15], bounding box (∼5 seconds), point (∼2 seconds), *etc.*

However, few works explore how to employ weakly-supervised labels for COD. There are only two works, CRNet [15] utilizes scribble annotation, and WS-SAM [14] utilizes scribble and point annotation to address weakly-supervised COD. However, they exhibit a significant performance gap compared to fully supervised COD methods. Thus, in this paper, we make an early attempt to explore a unified resolution of weakly-supervised COD for different weakly-supervised labels, including *point*, *bounding box*, and *scribble*, achieving comparable performance to fully supervised COD methods, shown in Fig. 1.

Although Segment Anything Model (SAM) [19] directly provides candidates for WSCOD, it is not trivial to address WSCOD task with the aid of SAM. It mainly faces four typical challenges, *1) Prompt compatibility of scribble:* SAM mainly supports box, point, and text-type inputs, but does not support scribble inputs which are applicable for existing WSCOD [15], as shown in Fig. 2(a). Then, the direct use of point input does not always yield satisfactory results. It is desirable to explore how to make different types of annotations in WSCOD compatible with SAM. *2) Extreme response:* For COD, SAM is prone to producing erroneous responses in extremely small regions or the entire background area, as shown in Fig. 2(b). This is due to the protective features of camouflaged objects, such as various mimetic patterns, spots, and low-contrast surface textures. *3) Semantically erroneous response:* SAM is also prone to wrong semantic responses for camouflage objects, including Specifically, a) non-camouflaged object response: SAM lacks training on relevant data, and lacks an understanding of camouflage semantics, b) local response: SAM has a rich segmentation granularity, making it prone to generating local semantic responses, as shown in

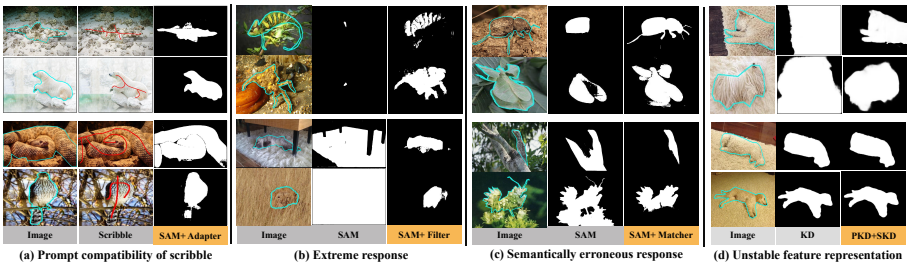

Fig. 2: Issues arising from SAM in COD, *i.e.*, a) prompt compatibility of scribble: SAM does not support the scribble input. b) extreme response: SAM produces extensive background responses (rows 3, 4) and minimal object responses (rows 1, 2). c) semantically erroneous response: SAM produces erroneous responses to non-camouflaged objects (rows 3, 4) and object-biased fine-grained semantic responses (rows 1, 2). d) unstable feature representation: SAM produces varied outcomes (1, 2 rows vs. 3, 4 rows) in similar scenarios. The contours of camouflaged objects are highlighted in blue.

Fig. 2(c). *4) Unstable feature representation:* The images of WSCOD task can exhibit completely different performance in very similar situations, as shown in Fig. 2(d). This is due to COD scenarios being challenging, and there is a significant gap in scale between the foundation model SAM and the student model. Direct distillation with limited supervision results in unstable learned features.

In this work, we propose a unified weakly-supervised COD framework, **SAM-COD**, supporting the input of arbitrary weakly-supervised label, *i.e.*, point, box or scribble by integrating the large visual model, SAM. We forgo the use of fully supervised labels for fine-tuning SAM and instead explore the use of weakly-supervised labels to prompt SAM. To mitigate the issues aforementioned, we first introduce the Prompt Adapter, which extracts the skeleton of the scribble label and then discretely samples it to points, making it compatible with SAM. Subsequently, we devise a Response Filter to filter out extreme responses from SAM by computing the ratio of the mask to the image size. Then, we construct a Semantic Matcher, which measures the semantic score of the mask by semantic entropy, which combines with the segmentation score of SAM to select masks that balance segmentation details and accurate semantics. We design a Prompt-Adaptive Knowledge Distillation according to different types of prompts, which enhances knowledge distillation by introducing prompt-guided knowledge for COD tasks, improving the quality of feature representation distilled from SAM.

Overall, our contributions are summarized as follows:

- We present a novel unified framework inheriting from SAM, integrating three supervision labels, *i.e.*, *scribble*, *bounding box*, and *point*, for weakly-supervised camouflaged object detection. To the best of our knowledge, this is the first WSCOD method to support all current weakly-supervised labels.
- We devise Response Filter and Semantic Matcher modules, addressing the issue that SAM is error-prone to producing unreliable extreme responses in COD scenarios, to obtain high-quality object masks.

- We propose a Prompt-adaptive Knowledge Distillation (PKD) for WSCOD. The distilled knowledge could be adaptively learned according to the three types of input prompts, which promotes knowledge distillation in WSCOD by focusing on distillation in high-value regions within the camouflage scene.
- We conduct extensive experiments on three widely-used COD datasets, demonstrating that our method achieves state-of-the-art performance. To the best of our knowledge, this is the first WSCOD method to outperform the state-of-the-art fully supervised methods under all the weakly-supervised labels. Moreover, when migrated to Salient Object Detection (SOD) and Polyp Segmentation tasks, our framework also achieves favorable results.

2 Related Work

Camouflaged Object Detection. COD focuses on detecting camouflaged objects within an image. SINet [8] proposes a COD dataset with 10K camouflaged images, which takes an average of around 60 minutes to annotate each image. [24, 27] attempt to mine inconspicuous features of camouflage objects from the background through meticulously designed feature exploration modules. ZoomNet [25] introduces a mixed-scale triplet network to address the challenges posed by COD. The aforementioned COD methods heavily rely on large-scale datasets with pixel-level annotations. However, the unclear boundaries make pixel-wise annotation of camouflaged objects a time-consuming and labor-intensive task. CRNet [15] was the first to introduce the S-COD dataset, which employs scribble annotations as weak supervision. WS-SAM [14] employs scribble and point annotations as weak supervision, but there is no dataset constructed with point annotation. Furthermore, box annotation has yet to be explored. So we propose box and point annotations to construct COD datasets. Furthermore, we propose the first model that simultaneously supports various weak supervision labels and outperforms fully supervised methods.

SAM in COD. SAM [19] excels in traditional segmentation tasks, achieving remarkable results, sometimes matching the performance of fully supervised methods, even in a zero-shot setting. [3, 28] indicate that while SAM shows promise in generic object segmentation, its performance on the COD task is constrained. SAM-Adapter [3] employs an adapter for efficient tuning instead of relying on traditional fine-tuning methods. This adaptation enables SAM to align with the data distribution in COD, reducing the cost of fine-tuning while simultaneously enhancing the performance of SAM in COD. WS-SAM [14] processes three augmented images through SAM and fuses the obtained masks to obtain the final pseudo-label. But the drawbacks of it are also obvious: 1) tripled SAM inference time 2) the full potential of SAM was not utilized, and only the highest scoring mask was used instead of the top-3 masks. We apply SAM to design a unified framework for point, box, and scribble annotations.

Knowledge Distillation. Knowledge distillation (KD) [1, 16] has been primarily designed to train a small network to mimic the output of a larger network to compress models. DINO [2] has introduced a straightforward self-supervised

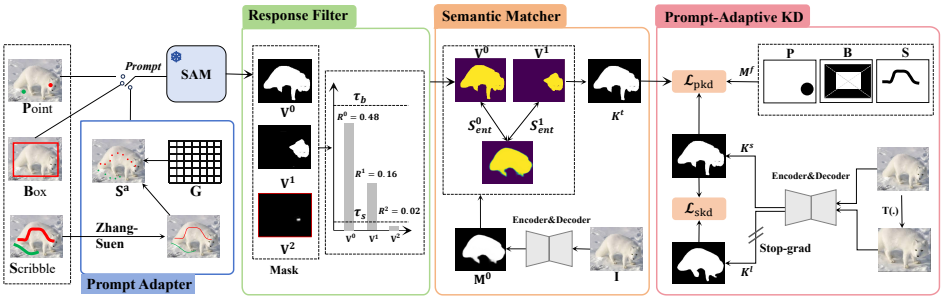

Fig. 3: The architecture of the proposed SAM-COD framework. Prompt Adapter supports scribbles to adapt the input prompt of SAM. Response Filter handles the extreme responses of SAM. Semantic Matcher is utilized to solve SAM's response issues arising from a lack of semantics in COD. Prompt-Adaptive Knowledge Distillation is designed for knowledge distillation in WSCOD.

method, which can be described as a label-less self-distillation model to optimize the representation learning. Distillation under WSCOD differs from traditional distillation, as 1) the COD scenario is challenging, and 2) there is little supervision. This makes traditional distillation methods unsuitable, and currently, there is a lack of exploration into distillation under the WSCOD task. So we design the prompt-adaptive knowledge distillation for the WSCOD task.

3 Approach

The overall architecture of the proposed framework is shown in Fig. 3. Prompt adapter is used to process scribbles to adapt SAM prompt input. Response filter is employed to handle extreme response situations of SAM under the prompt. Semantic matcher is utilized to improve SAM's response issues arising from a lack of COD-related semantics. Prompt-adaptive knowledge distillation is employed for the knowledge distillation in WSCOD.

3.1 Prompt Adapter

We use three kinds of weakly-supervised labels as prompts: point, box, and scribble. SAM directly supports types of point and box as input prompts. Unfortunately, SAM does not support scribble-type prompt. Therefore, we design a prompt adapter to convert scribbles into discrete points, making it compatible with SAM, as shown in Fig. 3.

Specifically, we first use the Zhang-Suen algorithm [35] to extract the skeleton of the scribble. Then, we perform discrete sampling on it. Specifically, we first create a grid G, where the grid points are uniformly distributed and the distance is $min(\alpha W, \alpha H)$, where H and W represent the length and width of the input image, respectively. α is the hyperparameter. Afterwards, we form a discrete

point set S^a by sampling points that coincide with both the scribble skeleton and grid lines. By now, we obtain the SAM prompt: $prt = \{P, B, S^a\}$, where P and B indicate Point and Box labels, respectively.

3.2 Response Filter

In COD, the camouflage objects usually exhibit excellent mimicry. So, SAM is prone to locate the extreme response under limited prompts, as shown in Fig. 2(b). To solve this, we design a response filter to prevent taking advantage of these evidently abnormal responses, as shown in Fig. 3.

Specifically, SAM outputs three valid masks and corresponding confidence scores given the prompt input:

$$\{V^i, S^i_{con} | i = 1, 2, 3\} = SAM(I, prt), \tag{1}$$

where V^i donate the *i-th* objects masks and S^i_{con} represents the corresponding segmentation confidence score. SAM defaults to using the mask with the maximum confidence score. Subsequently, we design a response filter to determine whether the mask exhibits extreme response by calculating the ratio of the mask size to the image size:

$$R^i = \mathbb{I}(\tau_s < \frac{A^i}{HW} < \tau_b), \tag{2}$$

where $\mathbb{I}(\cdot)$ is an indicator function. A^i is the area of the i-th mask V^i. τ_s and τ_b represent the maximum and minimum thresholds, respectively.

3.3 Semantic Matcher

SAM lacks the semantic knowledge, specifically the semantic understanding of camouflaged and overall granularity, leading to responses that do not align with the objects, as shown in Fig. 2(c). To solve it, we design a semantic matcher to measure the semantic score by the semantic entropy. It then selects masks with accurate semantics, as shown in Fig. 3.

Specifically, we first train the model on COD data to obtain the mask M^o:

$$M^o = D(E(I)), \tag{3}$$

where I donates input image, E and D are the encoder and decoder of the model, respectively. Although M^o may not rival the masks of SAM in segmentation details, training on COD data provides the model with a preliminary understanding of camouflage semantics.

Then, we design semantic entropy S^i_{ent} using M^o to measure the semantic score of the mask V^i:

$$S^i_{ent} = -\sum_j M^o_j log(V^i_j) + (1 - M^o_j) log(1 - V^i_j), \tag{4}$$

where j is the pixel index. Smaller values of S^i_{ent} indicate higher semantic score for V^i.

We select the mask with the highest product of S^i_{ent} and S^i_{con} scores, which balance segmentation details and accurate semantics, forming the optimal mask V^{opt} in V^i as:

$$opt = argmax(\frac{S^i_{con}}{S^i_{ent}}). \tag{5}$$

3.4 Prompt-Adaptive Knowledge Distillation

We employ knowledge distillation method to transfer the knowledge from large visual model SAM to a smaller model, thereby reducing the data cost and model size. However, COD task is challenging and the weak supervision makes knowledge distillation more difficult. Specifically, the proposed framework distillates the optimal mask V^{opt} from SAM as **t**eacher knowledge K^t to the **s**tudent knowledge K^s in our model. Moreover, we leverage the prior knowledge of different prompts to enhance the distillation quality,

Prompt-adaptive Mask Generation. The input prompts (scribble, box, and point) contain the structure, boundary, and discriminative region of camouflaged objects, respectively. These have been confirmed to be crucial for COD tasks [15, 33]. Therefore, we construct a prompt-adaptive mask M^f for the knowledge distillation according to the input prompt. The key distillation regions in M^f are marked as 0 (black areas). Specifically, 1) Scribble label, retaining the labeled foreground while discarding the background yields the corresponding M^f. 2) Point label, an inscribed circle of K^t with point label as the center. 3) Box labels, represented by *bold boxes*, with edge width and height being one-fourth of the length and width of box label, respectively.

Then, the prompt-adaptive knowledge distillation loss is defined as:

$$\mathcal{L}_{pkd} = -\sum_j M^F(K^t_j log(K^s_j) + (1 - K^t_j) log(1 - K^s_j)), \tag{6}$$

where K^s is the prediction mask and j is the pixel index. M^F=1+$\mathbb{I}$(M^f=0) and $\mathbb{I}(\cdot)$ is an indicator function. M^F as a coefficient in the distillation loss to allocate weight to prompt-guided regions, guides the distillation process to focus on learning key distillation regions.

Self-Knowledge Distillation. The learned feature representation of the model may not be robust enough, as shown in Fig. 2(d). Inspired by Self-Knowledge Distillation (SKD), we design a student model to enhance the representation learning. Specifically, for image I, we adopt visual transformations T, selecting from scale, colorjitter, etc. These visual transformations are able to change the appearance of images, as $I^t = T(I)$.

Then we encode and decode the augment images I^t, and transform them into two prediction maps K^s and K^l, denoting as:

$$K^s = D(E(I)), K^l = D(E(I^t)), \tag{7}$$

Our objective is to minimize the distance between two prediction maps:

$$\min \mathcal{D}(K^s, K^l) = \sum_i |K^s_i - K^l_i|, \tag{8}$$

where i is the pixel index, when a transformation T (e.g., scale, crop) is applied to the image I, this transformation T should be applied to K^s to be aligned with K^l. Here we follow the design of SKD, *i.e.*, stopping the gradient (*stopgrad*) update at one end, so the SKD loss function is defined as

$$\mathcal{L}_{skd} = \mathcal{D}(K^s, stopgrad(K^l)). \tag{9}$$

A robust feature representation could be learned from the teacher model to the student one by minimizing the above loss.

3.5 Network

Encoder&Decoder. Encoder and decoder designs can be flexibly replaced with existing models. In this work, we employ PVT [29] as the encoder, which obtains multi-scale features (F_{eat_1}, F_{eat_2}, F_{eat_3}, F_{eat_4}). The decoder consists of four 3x3 convolutional layers to reduce the channel dimension of F_{eat_i} to 64, followed by upsampling these F_{eat_i} to the same size. Subsequently, they are combined through concatenation, and finally, a 3×3 convolutional layer is used to obtain the final mask. In our method, all encoders and decoders refer to the same model.
Training Details. Our training process consists of two main steps. In Training Step 1, we train the encoder and decoder in the semantic matcher to obtain the distillation source K^t at the end. In Training Step 2, we use K^t for knowledge distillation to retrain the encoder and decoder. Further details are in the S.M.
Loss. Compared to other weakly-supervised methods [15, 31, 34], we have only two losses. The final loss $\mathcal{L}$ includes $\mathcal{L}_{pkd}$ and $\mathcal{L}_{skd}$ defined as:

$$\mathcal{L} = \mathcal{L}_{pkd} + \mathcal{L}_{skd}. \tag{10}$$

4 Experiments

4.1 Experimental Setup

Datasets. Our experiments are conducted on three COD benchmarks, CAMO [20], COD10K [8], and NC4K [22]. In order to evaluate our method, we first train our network on scribble annotated dataset S-COD [15]. Subsequently, we re-annotated 4040 images (3040 from COD10K, 1000 from CAMO) to create point-supervised dataset (P-COD) and bounding box-supervised dataset (B-COD) for training, while the remaining images are used for testing.
Evaluation Metrics. We adopt four evaluation metrics: Mean Absolute Error (MAE), S-measure (S_m) [5], E-measure (E_m) [6], weighted F-measure (F_β^w) [23].
Implementation Details. We implement our method with PyTorch and conduct experiments on one GeForce RTX4090 GPU and use ViT-H version of SAM. We chose PVT-B4 [29] as the encoder. We use the stochastic gradient descent optimizer with a momentum of 0.9, a weight decay of $5e$-4, and triangle learning rate schedule with maximum learning rate $1e$-3. The batch size is 8, and the training epoch is 60. Input images are resized to 512×512. We adopt the offline distillation, and the forward computation is performed only once, taking only 7 hours in training.

Table 1: Quantitative comparison with state-of-the-arts on three benchmarks. "F", "U", "S", "P", "Mix" denote fully-supervised label, unsupervised, scribble, point, and mixed random selection of one of three weakly-supervised labels, respectively. "–" is not available. Red and blue represent the first and second best performance, respectively.

Methods	Label	CAMO				COD10K				NC4K			
		MAE ↓	S_m ↑	E_m ↑	F_β^w ↑	MAE ↓	S_m ↑	E_m ↑	F_β^w ↑	MAE ↓	S_m ↑	E_m ↑	F_β^w ↑
SINet [8]	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
MGL-R [32]	F	0.088	0.775	0.812	0.673	0.035	0.814	0.851	0.666	0.052	0.833	0.867	0.740
PFNet [24]	F	0.085	0.782	0.841	0.695	0.040	0.800	0.877	0.660	0.053	0.829	0.887	0.745
UGTR [30]	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
UJSC [21]	F	0.073	0.800	0.859	0.728	0.035	0.809	0.884	0.684	**0.047**	**0.842**	**0.898**	**0.771**
ZoomNet [25]	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**
SAM-Ada. [3]	F	**0.070**	**0.847**	**0.873**	**0.765**	**0.025**	**0.883**	**0.918**	**0.801**	-	-	-	-
SAM [19]	-	0.132	0.684	0.687	0.606	0.050	0.783	0.798	0.701	0.078	0.767	0.776	0.696
SCSOD [31]	S	0.102	0.713	0.795	0.618	0.055	0.710	0.805	0.546	-	-	-	-
CRNet [15]	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
SAM-S [19]	S	0.105	0.731	0.774	-	0.046	0.772	0.828	-	0.071	0.763	0.832	-
WS-SAM [14]	S	0.092	0.759	0.818	-	0.038	0.803	0.878	-	0.052	0.829	0.886	-
SAM-P [19]	P	0.123	0.677	0.693	-	0.069	0.765	0.796	-	0.082	0.776	0.786	-
WS-SAM [14]	P	0.102	0.718	0.757	-	0.039	0.790	0.856	-	0.057	0.813	0.859	-
SAM-COD	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
SAM-COD	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**
SAM-COD	P	0.066	0.820	0.885	0.760	0.031	0.831	0.901	0.725	0.041	0.858	**0.918**	**0.802**
SAM-COD	Mix	**0.058**	**0.839**	**0.907**	**0.784**	0.031	**0.833**	0.903	0.725	**0.039**	**0.862**	0.912	0.798

4.2 Compare with State-of-the-art Methods

Quantitative Comparison. Being the first WSCOD method to incorporate point, scribble and box supervision, the proposed approach primarily leverages scribble supervision and full (mask) supervision as baselines. As demonstrated in Tab. 1, our method achieves substantial improvements, we averaged the results under three weakly-supervised labels, with an average enhancement of 26.8% for MAE, 6.1% for S_m, and 5.5% for E_m compared to the state-of-the-art weakly-supervised COD method, WS-SAM [14]. In particular, our approach performs exceptionally well under point and box supervision. It highlights our capability to achieve better performance with fewer annotations. Our approach even outperforms the state-of-the-art fully supervised method, ZoomNet [25]. To verify the advantages of our method over simply using of SAM, we compare with SAM-S and SAM-P, which fine-tune the mask decoder of SAM with scribble and point supervisions, respectively, by the partial cross-entropy loss. When testing, SAM-S and SAM-P use the automatic prompt generation strategy and report the results with the highest IoU scores. We do see performance gains after finetuning SAM with point (SAM-P) and scribble (SAM-S) supervision, but the results are still far below our method. This demonstrates the superiority of our method, which utilizes SAM prompt-adaptive knowledge distillation for small models. To

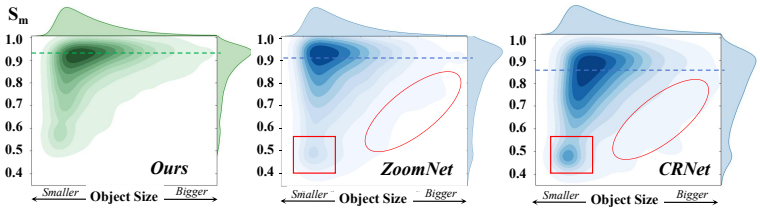

Fig. 4: Density distribution map about S_m and object size. Box and ellipse respectively represent challenging small and big objects, which have poor performance.

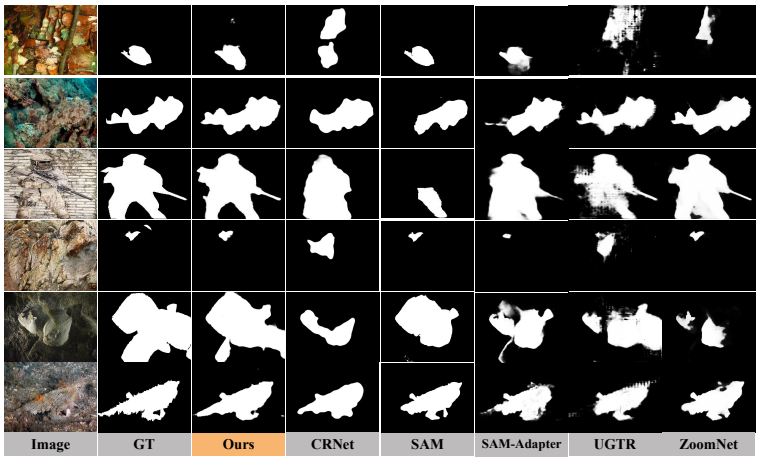

Fig. 5: Visual comparison with some representative state-of-the-art fully-supervised and scribble-supervised models.

further analyze the segmentation quality, we draw the density distribution map about S_m and object size on the test dataset in Fig. 4. It can be observed that the proposed method achieves an overall improvement and more stable performance on arbitrary sized objects compared to CRNet and ZoomNet. Especially for challenging small and large objects, our model has a significant improvement compared to CRNet and ZoomNet. Specifically, we design the "Mix" training method, i.e., randomly assigning one type of weakly-supervised label to each image in training. It is found that the performance is close to that of box-supervised method, particularly demonstrating a significant performance advantage on the CAMO dataset. The diversity of training introduced by mixing different labels is beneficial to learn more complex and rich feature representations comprehensively, capturing the feature at various levels.

Qualitative Evaluation. Our method produces prediction maps characterized by clearer and more complete object regions, along with sharper contours, significantly outperforming state-of-the-art weakly-supervised COD method CRNet [15] and fully supervised COD method ZoomNet [25], as shown in Fig. 5. Our method performs well in various challenging scenarios, including scenarios

Table 2: Comparison of parameters and MACs. "W" denotes the average of three weakly-supervised labels. All metrics are averages of the three datasets.

Methods	Label	Para.	MACs	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
ZoomNet	F	**32.38**	95.50	0.046	0.837	0.899	0.755
Ours	W	62.64	**52.63**	**0.044**	**0.843**	**0.907**	**0.770**

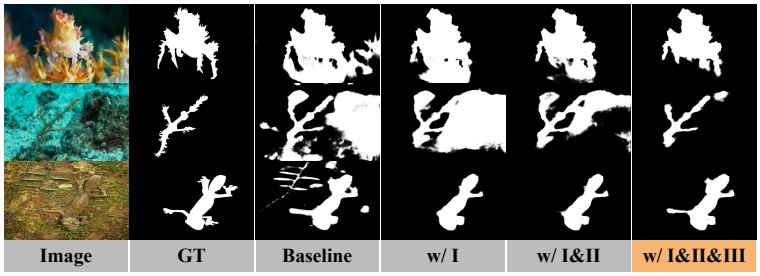

Fig. 6: Visualization of the component ablation study. I, II, and III represent prompt adapter, response filter, and semantic matcher, respectively.

with tiny objects (row 3), huge objects (row 4), high intrinsic similarities (row 2), indefinable boundaries (row 2 and 3), and complex backgrounds (row 1).
Parameter Complexity. Under similar parameter complexity and computational cost overhead, our model outperforms fully-supervised method ZoomNet [25], as shown in Tab. 2.

4.3 Ablation Study

As COD10K is the most representative dataset, all following ablation experiments are performed on it. Unless specifically indicated, all results are the averages of three different prompts (point, box, and scribble).
Effectiveness of Prompt Adapter. The ablation results of prompt adapter are presented in Tab. 4. Adapter has a large influence on the performance for scribble prompt. In addition, compared with baseline, a more accurate prediction map can be obtained by using the adapter, as shown in Fig. 6. In addition, adapter has a hyperparameter α to control the degree of discrete sampling, as shown in Tab. 5, with optimal effects achieved for suitable discrete sampling.
Effectiveness of Response Filter. As shown in Tab. 3, the results are significantly improved using response filter. Fig. 6 intuitively illustrates that the response filter enhances the precision of prediction maps. In addition, response filter has two hyperparameters τ_s and τ_b to control effects, as shown in Tab. 5.
Effectiveness of Semantic Matcher. We conduct ablation experiments for the semantic matcher, as shown in Tab. 3. In addition, a more complete visualization of the prediction map can be obtained by using semantic matcher, as shown in Fig. 6.

Table 3: Ablations study of SAM-COD.

Settings					Box				Point				Scribble			
SAM	SKD	Filt.	Match.	PKD	MAE↓	S_m ↑	E_m ↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	F_β^w ↑	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
✓					0.039	0.792	0.866	0.680	0.056	0.793	0.849	0.663	0.041	0.801	0.864	0.696
✓	✓				0.037	0.801	0.874	0.685	0.053	0.800	0.865	0.680	0.038	0.814	0.874	0.701
✓	✓	✓			0.035	0.817	0.883	0.698	0.036	0.821	0.890	0.708	0.035	0.823	0.890	0.711
✓	✓	✓	✓		0.031	0.831	0.903	0.725	0.032	0.829	0.899	0.720	0.032	0.827	0.901	0.722
✓	✓	✓	✓	✓	**0.028**	**0.842**	**0.914**	**0.745**	**0.028**	**0.831**	**0.901**	**0.725**	**0.029**	**0.831**	**0.904**	**0.728**

Table 4: Effect of the operation in prompt adapter. Discret. represents discrete sampling in the prompt adapter.

Zhang-suan	Discret.	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
×	×	0.189	0.591	0.592	0.364
✓	×	0.093	0.712	0.751	0.519
✓	✓	**0.038**	**0.814**	**0.874**	**0.701**

Table 5: The impact of α, τ_s, and τ_b in model.

α	MAE↓	τ_s	MAE↓	τ_b	MAE↓
0.025	0.031	0.001	0.038	0.5	0.032
0.050	0.030	0.003	0.036	**0.6**	**0.029**
0.075	**0.029**	**0.005**	**0.029**	0.7	0.029
0.100	0.038	0.010	0.039	0.8	0.033

Effectiveness of Prompt-Adaptive KD. We test the effect of prompt-adaptive KD compared with traditional KD. Tab. 3 shows that our PDK has a better performance. Additionally, using PKD also enhances the precision of prediction maps and able to continuously optimize representation and separate entangled object from background, making the model eventually learn robust representations, as shown in Fig. 7. Tab. 7 shows that CE loss performs best in PKD.

Effectiveness of Self-Knowledge Distillation. We conduct ablation experiments for SKD. Firstly, we separately test the performance of models with and without SKD, our proposed self-knowledge distillation obtains a significant improvement, as shown in Tab. 3. In addition, we conduct exhaustive experiments for data augmentation, which is an important operation for SKD, as shown in Tab. 6. We test different types of knowledge distillation losses and find that L1 loss is performing best, as shown in Tab. 7.

4.4 Extension to SOD

Our method excels not only in COD but also demonstrates remarkable performance in SOD. Specifically, we train on the SOD dataset using the labels of point, scribble, and box, respectively, and the results obtained are shown in Tab. 8. We attribute this success to our exploration of the potential of SAM and improvements in knowledge distillation. which contributes to our strong performance in WSSOD.

Table 6: The ablation study for different augmentations in Self-Knowledge Distillation. "S", "C", "T", "F", "G" are Scale, Crop, Translate, Flip, Guassblur, respectively.

Augmentations					MAE↓	S_m ↑	E_m ↑	F_β^w ↑
S	C	T	F	G				
					0.033	0.826	0.897	0.714
✓					0.032	0.829	0.899	0.721
✓	✓				0.032	0.828	0.898	0.719
✓	✓	✓			0.032	0.831	0.901	0.723
✓	✓	✓	✓		0.031	0.830	0.901	0.723
✓	✓	✓	✓	✓	**0.029**	**0.835**	**0.906**	**0.732**

Table 7: Ablation study on knowledge distillation losses. MSE, L1, CE mean Mean Square Error, L1, and Cross Entropy loss, respectively.

SKD	PKD	MAE↓	S_m ↑	E_m ↑	F_β^w ↑
w/o	w/o	0.039	0.801	0.873	0.684
w/MSE	w/o	0.038	0.801	0.878	0.693
w/CE	w/o	0.039	0.798	0.875	0.689
w/L1	w/o	**0.037**	**0.807**	**0.882**	**0.701**
w/o	w/ MSE	0.035	0.803	0.840	0.669
w/o	w/ L1	**0.032**	0.788	0.845	0.678
w/o	w/ CE	0.034	**0.821**	**0.891**	**0.707**

Table 8: Comparison with state-of-the-art WSSOD methods in SOD task.

Methods	Label	ECSSD			DUT-O			HKU-IS			DUTS-TE		
		MAE ↓	S_m ↑	F_β^{max} ↑	MAE ↓	S_m ↑	F_β^{max} ↑	MAE ↓	S_m ↑	F_β^{max} ↑	MAE ↓	S_m ↑	F_β^{max} ↑
AFNet [11]	F	0.042	0.913	0.935	0.057	0.826	0.797	0.036	0.905	0.923	0.046	0.867	0.863
BASNet [26]	F	0.037	0.916	0.943	0.057	0.836	0.805	0.032	0.909	0.928	0.048	0.866	0.859
SCSOD [31]	S	0.049	0.881	0.914	0.060	0.811	0.782	0.038	0.882	0.908	0.049	0.853	0.858
PSOD [13]	P	0.036	0.913	0.935	0.064	0.824	0.808	0.033	0.901	0.923	0.045	0.853	0.858
SAM-COD	P	0.033	0.925	0.947	0.051	0.844	0.826	0.024	0.946	0.941	0.034	0.892	0.898
SAM-COD	S	0.034	0.921	0.944	**0.050**	**0.846**	**0.829**	0.024	0.947	0.944	**0.033**	0.898	0.901
SAM-COD	B	**0.031**	**0.929**	**0.952**	0.051	0.844	0.828	**0.023**	**0.952**	**0.949**	**0.033**	**0.899**	**0.903**

4.5 Discussion

Is the prompt-adaptive KD from SAM important?

1) Data efficiency. We also evaluate the performance of our model and CRNet in few-shot setting, as shown in Fig. 8(a). Specifically, our model is trained only using the COD10K-Train dataset, which contains categories, and tested on the COD10K-Test dataset. Compared to CRNet, our model achieves promising results with much fewer training data. Especially in the extreme scenario, our method only uses one training image in each category, performance significantly surpasses that of CRNet training on the complete dataset. Fig. 8(a) verifies the effectiveness and efficiency of the proposed method. Through prompt-adaptive knowledge distillation, we transfer the knowledge from SAM to our model, only requiring a small amount of data.

2) Training efficiency. We visualize the curves of various metrics during the training, as shown in Fig. 8(b), where CRNet and our model share the same implementation details, including the optimizer, learning rate, epochs, and other relevant parameters. It is observed that our model demonstrates extremely fast convergence speed. To achieve the same performance, our model only needs one epoch of training, while CRNet typically requires more than 10 epochs. Because

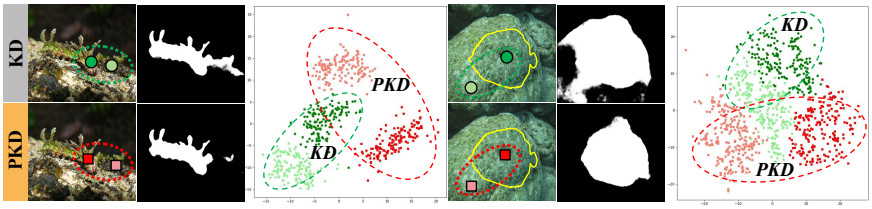

Fig. 7: Visualization of the feature. Entangled features from foregrounds and backgrounds are well separated by our prompt-adaptive KD. (visualized by t-SNE). Green and red colors represent features acquired under KD and PKD, respectively.

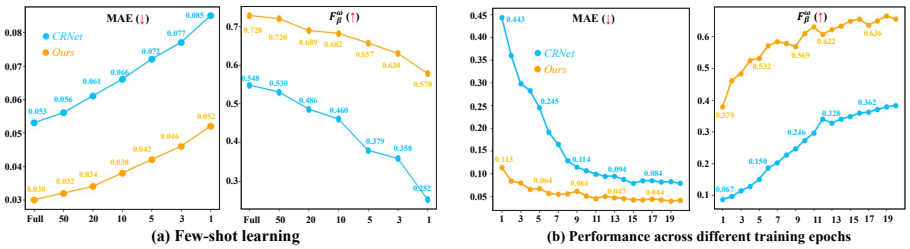

Fig. 8: The benefits of prompt-adaptive distilling knowledge from SAM. (a) Data efficiency: Few-shot performance. For each k-shot setting, we repeat the experiment 5 times to randomly select k images as training data. The average results are shown in the curve. (b) Training efficiency: Performance across different training epochs with the same training setting.

our model transfers the teacher knowledge from SAM to our small model through prompt-Adaptive knowledge distillation, which is much faster than learning a model from scratch.

5 Conclusion

In this paper, we propose a SAM-guided unified framework for weakly-supervised camouflaged object detection (WSCOD), named SAM-COD. It integrates all the existing labels for camouflaged objects (*i.e.*, scribbles, bounding boxes, and points), and achieves remarkable performance against the state-of-the-art weakly-supervised methods and even fully-supervised methods. The proposed SAM-COD typically aims to address the issues of SAM in the WSCOD task, *i.e.*, prompt compatibility of the scribble labels, extreme response, semantically erroneous response, and unstable feature representations. Specifically, in SAM-COD, we devise a prompt adapter to handle different labels and employ response filter and semantic matcher to mitigate the issue of imperfect outputs of SAM for camouflaged objects. Moreover, a prompt-adaptive knowledge distillation is proposed for reliable feature representations. We have conducted extensive experiments on camouflaged object datasets, demonstrating the effectiveness of the proposed method, which improves SAM to be more applicable to WSCOD.

References

1. Buciluǎ, C., Caruana, R., Niculescu-Mizil, A.: Model compression. In: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 535–541 (2006)
2. Caron, M., Touvron, H., Misra, I., Jégou, H., Mairal, J., Bojanowski, P., Joulin, A.: Emerging properties in self-supervised vision transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 9650–9660 (2021)
3. Chen, T., Zhu, L., Ding, C., Cao, R., Zhang, S., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, and more. arXiv preprint arXiv:2304.09148 (2023)
4. Chen, X., He, K.: Exploring simple siamese representation learning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 15750–15758 (2021)
5. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision. pp. 4548–4557 (2017)
6. Fan, D.P., Gong, C., Cao, Y., Ren, B., Cheng, M.M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18. pp. 698–704 (2018)
7. Fan, D.P., Ji, G.P., Cheng, M.M., Shao, L.: Concealed object detection. IEEE transactions on pattern analysis and machine intelligence **44**(10), 6024–6042 (2021)
8. Fan, D.P., Ji, G.P., Sun, G., Cheng, M.M., Shen, J., Shao, L.: Camouflaged object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2777–2787 (2020)
9. Fan, D.P., Ji, G.P., Zhou, T., Chen, G., Fu, H., Shen, J., Shao, L.: Pranet: Parallel reverse attention network for polyp segmentation. In: International conference on medical image computing and computer-assisted intervention. pp. 263–273. Springer (2020)
10. Fan, D.P., Zhou, T., Ji, G.P., Zhou, Y., Chen, G., Fu, H., Shen, J., Shao, L.: Inf-net: Automatic covid-19 lung infection segmentation from ct images. IEEE transactions on medical imaging **39**(8), 2626–2637 (2020)
11. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1623–1632 (2019)
12. Pérez-de la Fuente, R., Delclòs, X., Peñalver, E., Speranza, M., Wierzchos, J., Ascaso, C., Engel, M.S.: Early evolution and ecology of camouflage in insects. Proceedings of the National Academy of Sciences **109**(52), 21414–21419 (2012)
13. Gao, S., Zhang, W., Wang, Y., Guo, Q., Zhang, C., He, Y., Zhang, W.: Weakly-supervised salient object detection using point supervision. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 670–678 (2022)
14. He, C., Li, K., Zhang, Y., Xu, G., Tang, L., Zhang, Y., Guo, Z., Li, X.: Weakly-supervised concealed object segmentation with sam-based pseudo labeling and multi-scale feature grouping. Advances in Neural Information Processing Systems **36** (2024)
15. He, R., Dong, Q., Lin, J., Lau, R.W.: Weakly-supervised camouflaged object detection with scribble annotations. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 781–789 (2023)

16. Hinton, G., Vinyals, O., Dean, J.: Distilling the knowledge in a neural network. stat **1050**, 9 (2015)
17. Ji, G.P., Chou, Y.C., Fan, D.P., Chen, G., Fu, H., Jha, D., Shao, L.: Progressively normalized self-attention network for video polyp segmentation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 142–152. Springer (2021)
18. Ji, W., Yu, S., Wu, J., Ma, K., Bian, C., Bi, Q., Li, J., Liu, H., Cheng, L., Zheng, Y.: Learning calibrated medical image segmentation via multi-rater agreement modeling. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12341–12351 (2021)
19. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., Dollár, P., Girshick, R.B.: Segment anything. 2023 IEEE/CVF International Conference on Computer Vision (2023)
20. Le, T.N., Nguyen, T.V., Nie, Z., Tran, M.T., Sugimoto, A.: Anabranch network for camouflaged object segmentation. Computer vision and image understanding **184**, 45–56 (2019)
21. Li, A., Zhang, J., Lv, Y., Liu, B., Zhang, T., Dai, Y.: Uncertainty-aware joint salient object and camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10071–10081 (2021)
22. Lv, Y., Zhang, J., Dai, Y., Li, A., Liu, B., Barnes, N., Fan, D.P.: Simultaneously localize, segment and rank the camouflaged objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11591–11601 (2021)
23. Margolin, R., Zelnik-Manor, L., Tal, A.: How to evaluate foreground maps? In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 248–255 (2014)
24. Mei, H., Ji, G.P., Wei, Z., Yang, X., Wei, X., Fan, D.P.: Camouflaged object segmentation with distraction mining. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 8772–8781 (2021)
25. Pang, Y., Zhao, X., Xiang, T.Z., Zhang, L., Lu, H.: Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition. pp. 2160–2170 (2022)
26. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 7479–7489 (2019)
27. Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N.: Context-aware cross-level fusion network for camouflaged object detection. In: Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21. pp. 1025–1031 (2021)
28. Tang, L., Xiao, H., Li, B.: Can sam segment anything? when sam meets camouflaged object detection. arXiv preprint arXiv:2304.04709 (2023)
29. Wang, W., Xie, E., Li, X., Fan, D.P., Song, K., Liang, D., Lu, T., Luo, P., Shao, L.: Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 568–578 (2021)
30. Yang, F., Zhai, Q., Li, X., Huang, R., Luo, A., Cheng, H., Fan, D.P.: Uncertainty-guided transformer reasoning for camouflaged object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4146–4155 (2021)

31. Yu, S., Zhang, B., Xiao, J., Lim, E.G.: Structure-consistent weakly supervised salient object detection with local saliency coherence. In: Proceedings of the AAAI conference on artificial intelligence. pp. 3234–3242 (2021)
32. Zhai, Q., Li, X., Yang, F., Chen, C., Cheng, H., Fan, D.P.: Mutual graph learning for camouflaged object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12997–13007 (2021)
33. ZHAN, C., WANG, A., WANG, M.: Camouflage object segmentation method based on channel attention and edge fusion. Journal of Computer Applications **43**(7), 2166 (2023)
34. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 12546–12555 (2020)
35. Zhang, T.Y., Suen, C.Y.: A fast parallel algorithm for thinning digital patterns. Communications of the ACM **27**(3), 236–239 (1984)