Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Multimodal Fusion and Pseudo-Labeling for Enhanced Weakly-Supervised Camouflaged Object Detection

Danyang Yang
Kunming University of Science and Technology
Zao Liu
Kunming University of Science and Technology
Zhihong Zeng
Guangdong Polytechnic Normal University
Bei Cheng
chengbei@kust.edu.cn

Kunming University of Science and Technology

Research Article

Keywords: Multimodal, Weak supervision, Camouflaged object detection

Posted Date: May 20th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9056417/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Version of Record: A version of this preprint was published at Multimedia Systems on July 16th, 2026.
See the published version at https://doi.org/10.1007/s00530-026-02505-5.

Multimodal Fusion and Pseudo-Labeling for Enhanced Weakly-Supervised Camouflaged Object Detection

First Danyang Yang[1], Second Zao Liu[1†], Third Zhihong Zeng[2], Fourth Bei Cheng[1*]

[1*]Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, 650500, China.
[2]Institute of Interdisciplinary Studies, Guangdong Polytechnic Normal University, Guangzhou, 510665, China.

*Corresponding author(s). E-mail(s): chengbei@kust.edu.cn;
Contributing authors: yangdanyang@stu.kust.edu.cn;
liuzao@stu.kust.edu.cn; azh@gpnu.edu.cn;
†These authors contributed equally to this work.

Abstract

Camouflaged object detection (COD) is a challenging task in computer vision, aiming to accurately identify and locate objects that blend into complex backgrounds. Traditional COD methods rely heavily on pixel-level annotations, which are labour-intensive and prone to subjective bias. To address this, we propose a weakly-supervised COD approach that leverages multimodal information, specifically RGB and depth data, to enhance detection accuracy. Our method introduces a cross-modal adaptive relationship fusion module to dynamically adjust the contribution of each modality, ensuring optimized feature integration. Additionally, we employ a cross-modal pseudo-supervised label mapping method to generate comprehensive pseudo-RGB-D labels for network supervision. To ensure complete supervision across all modalities, we introduce a modality-specific loss function called deep dual loss. Experimental results on four benchmark datasets demonstrate that our approach significantly outperforms existing state-of-the-art methods, establishing a new benchmark in weakly-supervised COD. Here, we show an average improvement of 3.2% in structure measure ($\boldsymbol{S_\alpha}$) and 2.5% in mean E-measure ($\boldsymbol{E_\xi^{mean}}$) across datasets, highlighting the effectiveness of our multimodal fusion and pseudo-labeling strategies.

Keywords: Multimodal, Weak supervision, Camouflaged object detection

1 Introduction

Camouflaged object detection (COD) is a challenging problem in the field of computer vision. Its main goal is to accurately identify and locate objects that blend into the background in complex or natural environments [1]. This task is of significant importance in various application areas, including military stealth technology [2], biodiversity conservation [3], medical image processing (such as tumor detection)[4, 5] and autonomous driving[6].

In Fig. 1, we present the scribble annotation dataset in the context of two distinct tasks: salient object detection (SOD) and camouflage object detection (COD). Both tasks involve distinguishing an object of interest from its background, where accurate differentiation between foreground and background plays a pivotal role in achieving successful detection. In both SOD and COD, scribble annotations provide a sparse, weakly supervised way to define the foreground, requiring models to learn to extrapolate the object's full form from minimal input.

However, despite these similarities, camouflaged object detection presents additional challenges compared to salient object detection [7, 8]. For SOD, RGB images typically exhibit clear differences between the foreground and background in terms of color, texture, and other attributes [9]. Moreover, salient objects generally have intact shapes and relatively simple boundaries, which makes it easier for most models to effectively use scribble annotations to predict the object accurately [10].

Conversely, for COD, the boundaries between the foreground and background tend to be ambiguous, as camouflaged objects are designed to blend into their surroundings. Foreground and background often have similar visual attributes, with partial occlusions and multiple overlapping objects, making it significantly harder for models to distinguish concealed classes. As a result, directly applying weakly supervised SOD methods to COD often results in poor performance.

Weakly supervised camouflaged object detection (WS-COD) aims to reduce reliance on large amounts of pixel-level labeled data, making it useful for detecting camouflaged objects in complex environments. Traditional COD methods require substantial high-quality annotations, which are labor-intensive and prone to subjective bias. Weakly supervised approaches use coarse annotations (e.g., scribble-level, bounding boxes) to reduce annotation costs [11, 12]. This weakly supervised learning paradigm has also shown great potential in mitigating annotation efforts in other computer vision tasks, such as human pose estimation [13]. However, the high similarity between camouflaged objects and backgrounds poses a significant challenge for weak supervision. Camouflaged objects blend seamlessly in terms of color, texture, and boundaries, making it difficult to capture subtle distinctions. Recent research has focused on improving weak supervision through adaptive mechanisms [14], multi-scale feature extraction [15, 16], and boundary guidance [17–20]. Other strategies include surpassing human vision [21, 22], multimodal fusion [23, 24], and joint modeling for



Fig. 1 A visual comparison between salient object detection datasets annotated with scribbles and camouflaged object detection datasets, both under weak supervision. The foreground is marked in red, while the background is marked in blue.

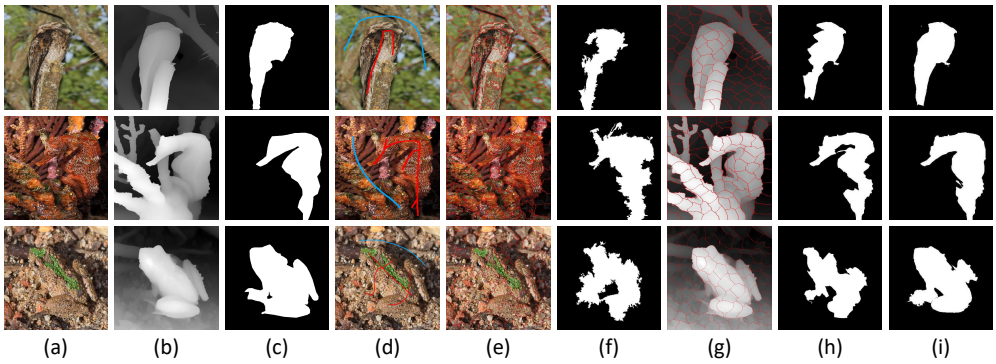


Fig. 2 The scribble supervised RGB-D camouflaged object detection dataset, along with its corresponding multimodal extended pseudo labels, comprises several components: (a) RGB image, (b) depth map, (c) ground truth, (d) scribble annotations image (where red indicates foreground and blue indicates background), (e) RGB superpixel image, (f) pseudo-RGB label, (g) depth superpixel image, (h) pseudo-depth label, (I) pseudo-RGB-D label.

multiple tasks [17, 25]. Despite progress, effectively utilizing high-level visual features for weakly supervised sparse data remains a challenge.

In this paper, we propose a novel algorithm for weakly supervised COD that introduces depth information as an additional clue to supplement sparse data. To address the issue of imbalanced fusion ratios between different modalities, we introduce a cross-modal adaptive relationship fusion (CARF) module. This module utilizes spatial view attention and channel view attention to generate cross-modal attention representations for different modalities, employing a two-stream encoder to extract RGB features and spatial position features separately. Subsequently, learnable parameters are introduced to fine-tune the input ratio of cross-modal attention. Through dynamic feature fusion, CARF promotes consistency in the structure and weight distribution between cross-modal information, allowing for dynamic adjustment of each modality's

contribution based on task requirements. This results in better handling of redundancy and conflicts between modalities, ultimately promoting complete segmentation under sparse information conditions.

To further address the challenge of sparse information in weakly supervised images, we propose a cross-modal pseudo-supervised label mapping (CPLM) module. In camouflaged object detection, pseudo-labels generated solely from RGB images often lack crucial structural information about the objects. While these labels may provide supplementary texture and detail information, they are limited in understanding three-dimensional space. In a multimodal setting, pseudo-labels generated from depth maps can better preserve the structural integrity of camouflaged objects and provide more accurate spatial localization. As illustrated in Fig.2, superpixel segmentation regions vary significantly across different modalities, as shown in Fig.2(e), (g). Consequently, the resulting pseudo-labels in Fig.2(f), (h) also differ accordingly. Although the extended pseudo-labels provide more accurate pixel-level information compared to scribble annotations, relying solely on pseudo-labels from a single modality can be inherently imprecise. To address this limitation, we propose a complementary approach that aggregates pseudo-labels generated from both modalities, as depicted in Fig.2(i). This strategy effectively mitigates the limitations of superpixel segmentation under a single modality. The CPLM module first expands the sparse annotations of RGB and depth superpixels, then aggregates the pseudo-labels using adaptive convolution iterations, and finally supervises using cross-modal multi-layer relationship fusion and deep feature branches.

To make effective use of the pseudo-labels generated across different modalities, we introduce a modality-specific loss function, referred to as deep dual loss, to ensure comprehensive supervision for all modalities. A unified structural loss is used to supervise the output of the main branch, while the deep dual loss supervises the output of the auxiliary branch, ensuring that the internal structure of the object remains smooth and consistent in depth map predictions, resulting in a clearer and more accurate depth representation.

Our main contributions are summarized as follows:

- To the best of our knowledge, this work represents the first integration of cross-modal information into weakly supervised COD, achieving fusion at both the feature and pseudo-label levels.
- We propose the CARF module, which utilizes learnable parameters to adaptively fine-tune cross-modal attention ratios. This mechanism enables deep optimization and adaptive fusion of cross-modal features.
- We design the CPLM module to extend multimodal superpixel generation. It incorporates dynamically weighted aggregation and a bidirectional supervision mechanism, effectively overcoming the limitations of single-modality superpixel segmentation.

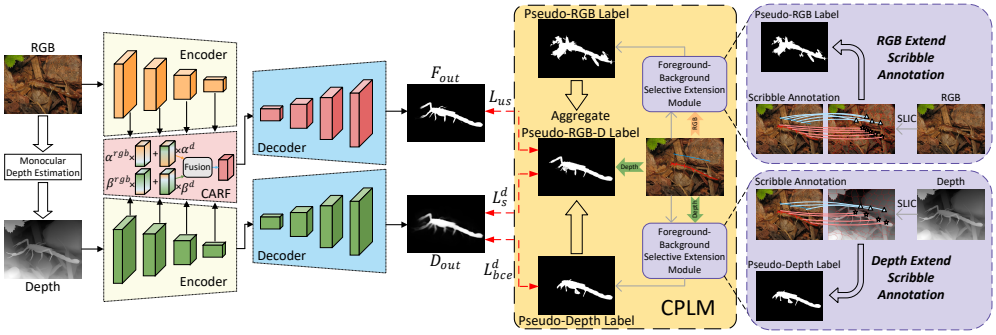


Fig. 3 Overall architecture of the proposed SPMCNet. It comprises three key components: a Transformer-based encoder, cross-modal adaptive relationship fusion (CARF) and cross-modal pseudo-supervised label mapping (CPLM).

2 Related Work

2.1 Weakly Supervised Camouflaged Detection

Pixel-level annotation in COD is highly time-consuming. To mitigate this burden and reduce annotation costs, recent years have witnessed the exploration of weakly-supervised COD approaches. He et al.[12] pioneered the use of scribbled labels for model training. He et al.[26] leveraging the Segment Anything Model to generate segmentation masks by treating scribbled annotations as prompts. Chen et al.[27] adopted a novel cue-adaptive knowledge distillation strategy to ensure reliable feature representation and mitigate the negative impact of inaccurate mask prediction. Zhang et al.[28] explored the potential of using bounding boxes as cues in camouflage scenes and proposed the first weakly semi-supervised COD method. Niu et al.[29] used extracted regional features as prior knowledge to produce discriminative edge maps to refine object boundaries in an iterative and multi-level manner. Subsequently, Chen et al.[30] introduced point labels for the first time in training, further streamlining the annotation process. Nevertheless, both methods rely solely on a single type of label for supervised training. We propose incorporating superpixel segmentation to generate pseudo labels for scribbled annotated data. By utilizing pseudo labels derived from superpixel segmentation mappings across different modalities for supervised training, we significantly enhance the model's robustness and accuracy. Furthermore, managing the inherent label noise generated during the pseudo-labeling process is crucial, a challenge that has also inspired robust learning strategies in other visual domains, such as 3D point cloud segmentation under label noise [31].

2.2 Multimodal Fusion Strategy

For RGB-D/T SOD and COD, avoiding sole reliance on RGB or depth/temp images, mitigating low-quality maps' impact, and enhancing model robustness is crucial [32]. Efficiently fusing cross-modal complementary info is a pivotal step. Exploring fusion methods significantly improves SOD and COD quality. For instance, advanced

transformer-based attention fusion mechanisms have demonstrated remarkable efficacy in capturing structural details for low-quality image enhancement [33]. In the field of RGB-D/T, Tu et al.[34] apply graph learning to cross-modal fusion, segmenting inputs and constructing graphs to capture relationships and saliency cues. Fu et al.[35] leverage a Siamese network for cross-modal knowledge sharing, treating depth maps as color images for feature extraction. Zhang et al.[36] captures contextual features through multi-scale and multi-modal fusion, integrating complementary information to improve robustness to illumination and occlusion. Li et al.[37] introduced the mutual information upper bound and mutual information minimization regularization term based on the principle of disentangled representation learning to promote the disentangled representation of each modality in salient object detection. Zhan et al.[38] designed a cross-modal fusion Mamba to perform feature enhancement of specific modalities and simulate the inter-modal correlation between RGB and depth features. Additionally, recent advancements have proposed integrating multi-scale refinement with cross-modalities fusion networks to deeply mine RGB and depth cues for salient object detection [39].

There are limited studies on RGB-D COD, which led us to analyze existing research. Xiang et al.[40] first incorporated depth as an auxiliary branch, but its lightweight design failed to effectively control depth quality during fusion, impacting detection performance. Wu et al.[11] adopted a three-segment network comprising a deep generation network, an object pop-up network, and a segmentation network to predict segmentation. Yet, even with prior knowledge, their approach was found to be insufficient in distinguishing objects from backgrounds when fusing different modalities. Unlike these methods, our approach dynamically adjusts fusion ratios with learnable weights, optimally combining RGB and depth features, and uses pseudo-labels for supervision to enhance accuracy.

3 Methodology

The structure of our model, illustrated in Fig. 3, is named the SuperPixel-Multimodal Collaboration Network (SPMCNet). Given an input RGB image I with size $I \in \mathbb{R}^{3\times H\times W}$ where H and W represent the spatial resolution of the image, our objective is to leverage the depth map from multimodal data to facilitate the modeling of pseudo-depth labels while integrating spatial information for more precise segmentation of camouflaged objects. Initially, the input image I is fed into a monocular depth estimation network to produce a highly accurate depth map D. Subsequently, I and D are individually processed through a two-stream encoder to extract RGB features and spatial features, respectively. Following this, we fuse the RGB and spatial features and decode them as the primary branch output F_{out} of the network, aiming to extract boundary information of the camouflaged object. In parallel, we decode the spatial features separately as the auxiliary branch output D_{out} to capture the location information of the camouflaged object. Ultimately, the pseudo-RGB labels and pseudo-depth labels generated by the joint action of scribbled annotation and I, D are aggregated to generate pseudo-RGB-D labels. These pseudo-RGB-D labels serve

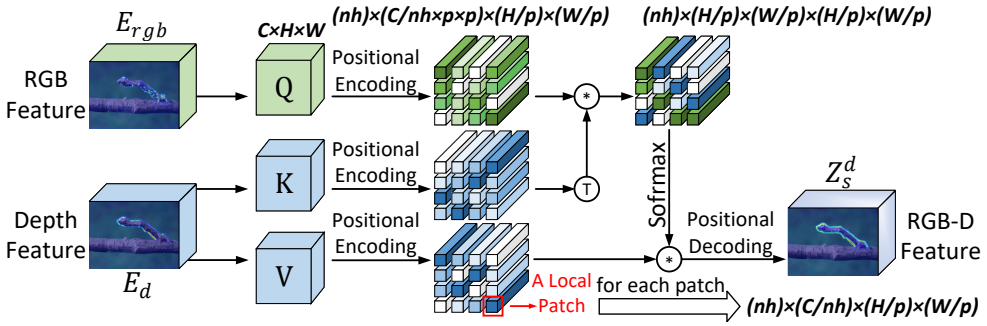


Fig. 4 Illustrations of the cross-modal spatial view attention. The category activation map shows that our attention module yields more precise and comprehensive segmentation, enhancing representation.

to supervise the entire network, with the pseudo-depth labels also providing auxiliary supervision for improved performance.

3.1 Generation of Detailed Depth Maps from RGB Images

In actual environments, acquiring depth maps is often prohibitively expensive, requiring significant manpower and material resources. To address this issue, we employ a pre-existing method for generating depth maps, enabling multimodal input. Thanks to the rapid advancements in depth estimation, we have selected the state-of-the-art Depth Anything V2 [41] in the field of monocular estimation, which provides us with precise depth map. As depicted in Fig. 2(a), (b), it becomes evident that for the task of identifying camouflaged objects, distinguishing the object under an RGB image poses a challenge due to its similar color and texture information to the surrounding environment. In contrast, the depth map, which represents distance information through pixel grayscale values, facilitates the identification of camouflaged objects. Consequently, the spatial position information encoded within the depth map is invaluable in locating camouflaged objects.

3.2 RGB and Depth Map Fusion

We input I and D into the dual-stream encoder to extract RGB features and spatial position features, respectively.

We draw inspiration from Pang et al.[42] and introduce CARF. This approach utilizes spatial view attention and channel view attention tailored to different modalities to derive cross-modal attention representations. Ultimately, we introduce parameters to fine-tune the input ratio of cross-modal attention, thereby achieving a fully integrated and optimized fusion result.

3.2.1 Cross-modal Spatial View Attention

We first convolve the input RGB feature map E_{rgb} and the deep feature map E_d to obtain Q, K and V, where the information sources for Q, K and V stem from distinct

modalities. This transformation facilitates the establishment of interactions between these diverse information sources. Subsequently, position encoding is applied to achieve block-level patch re-embedding. By reorganizing the inputs Q, K and V, the attention mechanism operates at the block level, enabling each block to independently perform attention calculations. Finally, we leverage the multi-head self-attention mechanism within the transformer [43] to compute attention scores between each feature encoding block and perform position decoding. This approach captures the spatial relationships between image blocks in a multi-modal context. Fig. 4 illustrates the entire process of this module. The category activation map we selected represents the input and output from the shallow layer. It becomes evident that, after the interaction of different information sources, the attention is more pronounced around the edges of the camouflaged object. This can be expressed as:

$$Z_s^d = [\text{Attn}(Q_h^{rgb}, K_h^{d\top}, V_h^d, e)]_{h=1}^{N_h} \mathbf{W}_s \tag{1}$$

where Q_h, K_h and V_h are a single head of query, key and value, respectively. N_h is the number of heads, D denote the number of pixel tokens and the embedding dimension. $e = \sqrt{D/N_h}$ is a scaling factor. $[\ldots]_{h=1}^{N_h}$ represents the concatenation operation for all heads and $\mathbf{W}_s$ is output projection matrices corresponding to spatial views to fuse these heads.

3.2.2 Cross-modal Channel View Attention

Cross-modal channel view attention is akin to cross-modal spatial view attention, with the key distinction lying in the omission of position encoding and the shift in the focus of pairwise similarity calculations. Instead of comparing spatial positions within the feature sequence, this approach calculates similarities between independent feature channels, thereby capturing the intricate relationships existing among channels within the multimodal feature map. Mathematically, this can be expressed as:

$$Z_c^d = [\text{Attn}(Q_h^{rgb\top}, K_h^d, V_h^{d\top}, n)]_{h=1}^{N_h} \mathbf{W}_c \tag{2}$$

where $N = H \times W$. $n = \sqrt{N}$ is a scaling factor. $\mathbf{W}_c$ is output projection matrices corresponding to channel views to fuse these heads.

3.2.3 Cross-modal Adaptive Relationship Fusion

The output features Z_s^d and Z_c^d from the cross-modal spatial view attention branch and the cross-modal channel view attention branch are integrated through learnable weights $\alpha^d \in [0, 1]$ and $\beta^d \in [0, 1]$, respectively, yielding the cross-modal attention output Z_{sc}^d, which represents the propagation of RGB information into the deep features. However, merely obtaining information flow from one modality to another is insufficient. We also aim to capture the propagation of deep information back to the RGB features, denoted as Z_{sc}^{rgb}. Consequently, we adjust the variables in the aforementioned formula accordingly: $Q^{rgb} \to Q^d$, $K^d \to K^{rgb}$, $V^d \to V^{rgb}$, and introduce additional learnable weights $\alpha^{rgb} \in [0, 1]$ and $\beta^{rgb} \in [0, 1]$ for this purpose. Ultimately,

we augment each original feature map with its respective cross-modal attention output, and then concatenate these enhanced feature maps to produce the final output Z_{sc}^{rgbd}, which can be expressed as follows:

$$Z_{sc}^{d} = \alpha^{d} Z_{s}^{d} + \beta^{d} Z_{c}^{d} \tag{3}$$

$$Z_{sc}^{rgb} = \alpha^{rgb} Z_{s}^{rgb} + \beta^{rgb} Z_{c}^{rgb} \tag{4}$$

$$Z_{sc}^{rgbd} = \mathrm{Cat}(Z_{sc}^{d} + E_{d}, Z_{sc}^{rgb} + E_{rgb}) \tag{5}$$

Algorithm 1 Foreground-Background Selective Extension.

Input: (1) An RGB or depth image M;
(2) Scribble annotation $Y = \{y_i\}_{i=1}^{N}$, where $y_i = 1/2/0$ indicates foreground, background, and unknown pixels, respectively, and N is the number of pixels;
Output: Extended pseudo labels $\hat{Y}$;
/*Step1: Use SLIC algorithm to segment the image M, generating the superpixel image mask set $\mathbb{P} = \{P^k\}_{k=1}^{T}$, where T is the number of superpixels; Each P^k only highlights the current superpixel.*/
$\mathbb{P} \leftarrow SLIC(M)$;
/*Step2: If the degree of overlap between a superpixel and the foreground scribble $y_i = 1$ surpasses that with the background scribble $y_i = 2$, and additionally, the foreground overlap exceeds 1, then the superpixel is labeled as foreground.*/
Initialize $\hat{Y} \leftarrow zeros(N)$;
for each P^k in $\mathbb{P}$ **do**
    if $\sum(P^k \cdot (y_i = 1)) > \sum(P^k \cdot (y_i = 2))$ **and** $\sum(P^k \cdot (y_i = 1)) > 1$ **then**
        $\hat{Y} \leftarrow \hat{Y} + P^k$
    end if
end for
    return $\hat{Y}$.

3.3 Cross-modal Pseudo-supervised Label Mapping

Existing methods [44, 45] employ scribble annotation on RGB images to generate pseudo labels through superpixel segmentation. However, for the detection of camouflaged objects, as exemplified in Fig. 2(f), (h), the pseudo labels derived solely from RGB images frequently lack substantial structural information about the camouflaged objects. They may also have limitations in comprehending three-dimensional space, though they offer certain textural and detailed information as supplements. Conversely, we have observed that the pseudo labels generated from depth maps, in a multimodal setting, can better preserve the structural integrity of camouflaged objects, with relatively accurate general positions. However, they lack the richness of texture details found in pseudo-RGB labels and struggle to capture the shapes and edges of small objects.

Given these insights, we aim to supervise the network's output by fusing pseudo-RGB labels with pseudo-depth labels to create comprehensive pseudo-RGB-D labels. Additionally, considering the unique challenges posed by camouflaged objects, we incorporate pseudo-depth labels to refine the final prediction results from the depth branch, thereby achieving more precise localization of camouflaged objects. Our approach comprises three key steps: extension, aggregation and supervision.

3.3.1 Extension

We introduce a foreground-background selective extension module to generate pseudo labels for two modalities. Initially, we utilize simple linear iterative clustering (SLIC) [46] to segment the image into multiple superpixel regions, which simultaneously considers both spatial and color distances for clustering. Subsequently, we perform algorithmic processing on the scribble annotated image. For camouflaged objects, if the number of superpixel regions is insufficient, it becomes challenging to distinguish between foreground and background pixels within the same superpixel, as they may coexist. Conversely, overly large superpixel regions increase computational complexity. To tackle this issue, we adopt a compromise algorithm strategy to address the challenge of fine-grained superpixel segmentation. Specifically, we first segment the image into appropriately sized superpixel regions. If the number of foreground pixels within a superpixel exceeds the number of background pixels, and the count of foreground pixels is greater than one, we label that superpixel as foreground. Otherwise, it is labeled as background. Consequently, the scribble annotation Y is expanded through RGB and depth superpixels to generate corresponding pseudo-RGB label $\hat{Y}^r$ and pseudo-depth label $\hat{Y}^d$, striking a balance between efficiency and performance. The algorithm's process is outlined in Alg 1.

3.3.2 Aggregation

The extended pseudo-labels, obtained through SLIC which considers local similarity, inherently contain some noise at the object boundaries, as illustrated in Fig. 2(f), (h). To mitigate this, we employ the PAR method [47] to smooth and refine both pseudo-labels, ultimately generating the aggregated pseudo-RGB-D label $\hat{Y}^{rd}$. The PAR method defines a convolution kernel that incorporates RGB and spatial information and computes it based on the pixel's neighboring positions. This kernel is iteratively applied to the average values using adaptive convolution [48], effectively correcting the pseudo-labels.

$$\hat{Y}^{rd} = PAR(Avg(\hat{Y}^r, \hat{Y}^d), D) \tag{6}$$

Here, we utilize the depth map D instead of the RGB image I to refine the average, as the depth map offers richer distance information that is more effective in refining camouflaged objects compared to the RGB map. As evident in Fig. 2(i), the aggregated pseudo-RGB-D label yields superior results.

3.3.3 Supervision

We employ the decoded output from the CARF as the main branch prediction, while utilizing the decoded output of the deep feature branch as an auxiliary prediction for supervision. Notably, we apply a prediction head and perform upsampling on four decoded outputs in the main branch for supervision, leveraging deep supervision to concurrently oversee the outputs of the initial layers. In contrast, the auxiliary branch utilizes only the final layer's decoded output for supervision, as the low-level features in the depth map, rich in details but also noise, may lead to unstable predictions if directly decoded. Consequently, we rely on the model's ability to filter out noise and irrelevant information during the initial layers' extraction and fusion process.

4 Loss Function

To optimally harness the pseudo-labels generated across different modalities, we introduce a modality-specific loss function, thereby ensuring comprehensive supervision of the entire model.

4.1 Unified Structural Loss

We employ the unified structural loss to supervise the main branch output. This loss is specifically designed to optimize pixel-level accuracy, region overlap accuracy, and structural similarity of the image simultaneously, utilizing binary cross entropy (BCE), intersection over union (IoU) and structural similarity (SSIM) losses. By doing so, it comprehensively highlights the camouflaged object from various perspectives. It can be formulated as:

$$\begin{aligned} L_{us} = & \sum_{i=1}^{4} (L_{bce}(F_{out}^{i}, \hat{Y}^{rd}) + L_{iou}(F_{out}^{i}, \hat{Y}^{rd}) + \\ & L_{ssim}(F_{out}^{i}, \hat{Y}^{rd})) \end{aligned} \tag{7}$$

4.2 Depth Dual Loss

We use depth dual loss to supervise the auxiliary branch, emphasizing the structural distinctiveness of camouflaged objects. The depth dual loss comprises two parts: (1) BCE loss (L_{bce}^{d}) to differentiate depth values between the camouflaged object and its surroundings using pseudo-depth labels; (2) Depth smoothness loss (L_{s}^{d}), which multiplies the pseudo-RGB-D label with the depth prediction ($D_{obj} = D_{out} \otimes \hat{Y}^{rd}$) to mask background pixels, ensuring smooth internal object structure and reducing high-frequency noise. Let ∇_x and ∇_y be the Sobel operations.

$$L_{bce}^{d} = L_{bce}(D_{out}, \hat{Y}^{d}) \tag{8}$$

$$\overrightarrow{n(p)} = (-\nabla_x(D_{obj}(p)); -\nabla_y(D_{obj}(p)); 1) \tag{9}$$

$$L_{s}^{d} = \sum_{p} \sum_{q \in \mathcal{N}(p)} 1 - cosine(\overrightarrow{n(p)}; \overrightarrow{n(q)}) \tag{10}$$

$$L_{dd} = L_{bce}^{d} + L_{s}^{d} \tag{11}$$

where $\overrightarrow{n}$ stands for the normal, p denotes the pixels within the object region, $\mathcal{N}(p)$ is the neighboring pixels, and cosine is the cosine similarity between two vectors.

Therefore, the final loss function L_{total} is recorded as:

$$L_{total} = L_{us} + L_{dd} \tag{12}$$

5 Experiments

5.1 Experiment Settings

Datasets. The training set we utilize is S-COD [12], introduced by He *et al.*, marking the first dataset tailored specifically for the COD task under a weakly-supervised framework. Fig. 2(d) visually represents the annotations of this dataset. For the test set, we employ four widely recognized datasets: CHAMELEON [49], CAMO [50], COD10K [51] and NC4K [52].

Evaluation Metrics. We use four common metrics for evaluation, including structure measure (S_α) [53], mean E-measure (E_ξ^{mean}) [54], adaptive F-measure (F_β^w) [55] and mean absolute error ($\mathcal{M}$).

Implementation Details. Our proposed model is implemented in PyTorch and accelerated using an NVIDIA GTX 4060 Ti GPU. We use PVT-v2-B2 [56] as the backbone, initialized with pre-trained weights, with independent backbones for RGB images and depth maps. Under this setting, the overall computational complexity is 182.13G FLOPs, and the total number of parameters is 53.33M. Here, for a comprehensive comparison, we also provide one variants of our SPMCNet with backbones of ResNet50. Input images are resized to 352×352 pixels, and we set the number of superpixels to 100. To reduce overfitting, we apply random data augmentation, including scaling, flipping, rotation, and color enhancement. The model is trained using the Nadam [57] optimizer for 50 epochs, with a batch size of 8. The initial learning rate is $1e^{-4}$, decayed to 1/10 every 30 epochs.

5.2 Compare with State-of-the-arts

Quantitative Comparison. We selected a total of 14 methods for comparison, comprising 8 fully-supervised , 2 unsupervised and 6 weakly-supervised methods. Among these, PopNet and RISNet are distinct for utilizing RGB-D input, while PCOD is a weakly-supervised method that uses point-annotated data for training, in contrast to the other methods that rely on scribble-annotated data.

Tab. 1 presents the quantitative analysis results, demonstrating that our model exhibits robust performance across all four datasets, consistently outperforming the unsupervised and weakly-supervised baselines. Notably, our model achieves the most significant improvement on the CAMO dataset, surpassing even most of fully-supervised methods. This is attributed to the camouflaged objects in the CAMO dataset being more visually similar to the background, resulting in lower $\mathcal{M}$ values

Table 1 Quantitative comparison with state-of-the-arts on four benchmarks. “F”, “U”, “S”, “P” denote fully-supervised, unsupervised, scribble-supervised and point-supervised methods, respectively. The best result among scribble-supervised and point-supervised methods is in red, and the second-best result is in blue.

Methods	Backbone	Sup.	Input	CHAMELEON				CAMO				COD10K				NC4K			
				$S_\alpha \uparrow$	$E_\xi^{mean} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{mean} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{mean} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{mean} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
$SINet_{21}$ [1]	ResNet50	F	RGB	0.872	0.936	0.806	0.034	0.745	0.829	0.644	0.092	0.776	0.864	0.631	0.043	0.808	0.883	0.723	0.058
LSR_{21} [52]	ResNet50	F	RGB	0.893	0.938	0.823	0.033	0.793	0.826	0.696	0.085	0.793	0.868	0.673	0.041	0.839	0.883	0.766	0.053
$ERRNet_{22}$ [58]	ResNet50	F	RGB	0.877	0.927	0.696	0.036	0.779	0.842	0.612	0.085	0.786	0.867	0.523	0.043	0.827	0.887	0.643	0.054
$FDNet_{22}$ [23]	Res2Net50	F	RGB	0.898	0.949	0.837	0.027	0.844	0.898	0.778	0.062	0.837	0.918	0.731	0.030	0.834	0.893	0.750	0.052
$FEDER_{23}$ [24]	ResNet50	F	RGB	0.894	0.947	0.855	0.028	0.807	0.873	0.785	0.069	0.823	0.900	0.740	0.032	0.823	0.905	0.817	0.045
$DPSNet_{25}$ [10]	PVTv2	F	RGB	0.917	0.965	0.882	0.020	0.888	0.893	0.866	0.044	0.875	0.938	0.806	0.022	0.891	0.941	0.866	0.030
$PopNet_{23}$ [11]	Res2Net50	F	RGB-D	0.910	0.962	0.875	0.022	0.806	0.869	0.744	0.073	0.827	0.897	0.757	0.031	0.852	0.908	0.802	0.043
$RISNet_{24}$ [59]	PVT-v2-B2	F	RGB-D	-	-	-	-	0.870	0.922	0.827	0.050	0.873	0.931	0.799	0.025	0.882	0.925	0.834	0.037
$DUSD_{18}$ [60]	ResNet101	U	RGB	0.578	0.634	0.316	0.129	0.551	0.594	0.308	0.166	0.580	0.646	0.276	0.107	-	-	-	-
$USPS_{19}$ [61]	ResNet101	U	RGB	0.573	0.631	0.380	0.188	0.568	0.641	0.399	0.207	0.519	0.536	0.265	0.196	-	-	-	-
SS_{20} [62]	VGG16	S	RGB	0.782	0.860	0.654	0.067	0.696	0.786	0.562	0.118	0.684	0.770	0.461	0.071	0.761	0.779	-	0.091
$SCWS_{21}$ [63]	ResNet50	S	RGB	0.792	0.881	0.714	0.053	0.713	0.795	0.618	0.102	0.710	0.805	0.546	0.055	0.779	0.853	-	0.064
$CRNet_{23}$ [12]	ResNet50	S	RGB	0.818	0.897	0.744	0.046	0.735	0.815	0.641	0.092	0.733	0.832	0.576	0.049	0.775	0.855	0.688	0.063
$WS\text{-}SAM_{23}$ [26]	ResNet50	S	RGB	0.824	0.897	**0.757**	0.046	0.759	0.818	0.667	0.092	0.803	**0.878**	0.680	0.038	0.829	0.886	0.757	0.052
$PCOD_{24}$ [30]	PVT-v2-B4	P	RGB	**0.828**	**0.902**	0.753	**0.044**	0.798	**0.872**	0.727	0.074	0.784	0.859	0.650	0.042	0.822	0.889	0.748	0.051
$SAM\text{-}COD{+}_{24}$ [27]	PVT-v2-B4	S	RGB	-	-	-	-	**0.836**	**0.903**	**0.779**	**0.060**	**0.833**	**0.904**	**0.728**	**0.029**	**0.859**	**0.912**	**0.795**	**0.039**
Ours	ResNet50	S	RGB-D	0.853	0.896	0.770	0.042	0.846	0.891	0.770	0.060	0.791	0.843	0.635	0.043	0.830	0.876	0.738	0.055
Ours	PVT-v2-B2	S	RGB-D	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	0.867	**0.682**	**0.036**	**0.846**	**0.893**	**0.771**	**0.049**

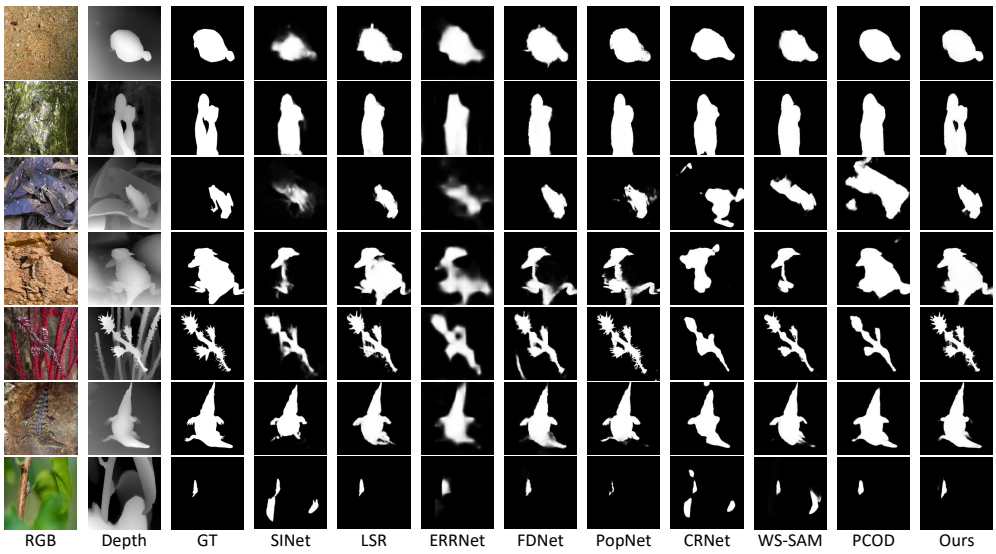


Fig. 5 Qualitative comparison of our method with state-of-the-arts fully-supervised and weakly-supervised methods in challenging scenarios.

for other methods. However, the integration of depth maps significantly enhances SPMCNet's ability to segment camouflaged objects.

Compared to SAM-COD+, the second-best performing weakly-supervised method in CAMO dataset, our model achieves improvements of 2.3% on S_α, 2% on F_β^w, and a reduction of 1% on $\mathcal{M}$. Furthermore, when compared to other scribble annotation based models using the same annotated data, our approach demonstrates superior performance. In several metrics across multiple datasets, our method even surpasses fully-supervised methods, highlighting its effectiveness in detecting camouflaged objects.

Qualitative Evaluation. Fig. 5 presents the outcomes of our qualitative analysis, demonstrating that our method effectively tackles various challenging scenarios. Notably, the incorporation of depth information significantly aids in pinpointing camouflaged objects. These challenging scenarios encompass high-similarity objects (rows 1-4), irregularly shaped objects (row 5), large-sized objects (row 6), and minute objects (row 7).

5.3 Ablation Study

To further validate the effectiveness of our proposed method, we conducted experiments on the CHAM, CAMO and COD10K datasets.

5.3.1 Exploring the Impact of Multimodality

To delve deeper into the impact of integrating depth maps on camouflaged object detection, we modified the two-stream encoder to generate experimental variants,

Table 2 Ablation on input of different modalities. The best results are highlighted in **bold**.

Input Type	CHAM				CAMO				COD10K			
	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
RGB+RGB	0.851	0.900	0.764	0.042	0.810	0.862	0.722	0.077	0.791	0.843	0.637	0.043
Depth+Depth	0.864	0.910	0.770	0.038	0.848	0.888	0.780	0.060	0.798	0.850	0.649	0.042
RGB+Depth	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	**0.867**	**0.682**	**0.036**

Table 3 Ablation on CARF and aggregated pseudo-RGB-D labels. The best results are highlighted in **bold**.

NO.	Settings	Pseudo Labels	CHAM				CAMO				COD10K			
			$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
I	w/o CARF	RGB-D+Depth	0.846	0.898	0.757	0.042	0.808	0.864	0.720	0.076	0.791	0.846	0.637	0.043
II	w/o CARF	RGB+Depth	0.794	0.837	0.640	0.067	0.769	0.815	0.636	0.098	0.742	0.786	0.542	0.060
III	w/ CARF	RGB+Depth	0.811	0.853	0.666	0.060	0.804	0.843	0.685	0.081	0.751	0.799	0.560	0.055
IV	w/ CARF	RGB-D+Depth	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	**0.867**	**0.682**	**0.036**

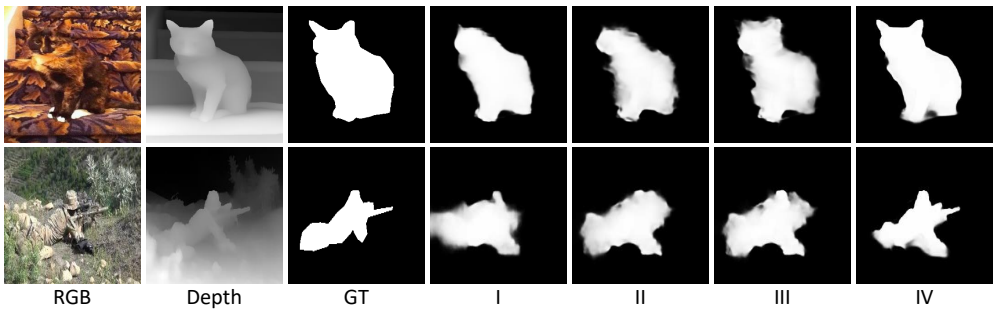


Fig. 6 A visual example of the component ablation study.

where either RGB images or depth maps served as the sole input. The outcomes of these experiments with varying modality inputs are presented in Tab. 2. Our analysis reveals that using multi-modality inputs outperforms single-modality inputs, and the data suggests that incorporating different modalities enables the model to learn cues that facilitate the refinement of COD. Additionally, we observed that when only RGB is used as input, the performance is similar to that achieved with a ResNet50 backbone, indicating that our method remains highly competitive even without relying on precise depth maps.

5.3.2 Analyze the Effectiveness of CARF and Aggregated Pseudo-RGB-D Labels

To assess the impact of our proposed CARF and aggregated pseudo-RGB-D labels, we conducted two experiments. In the first, we removed the cross-modal attention module, supervising only the RGB and depth branches independently. In the second, we used only pseudo-RGB or pseudo-depth labels instead of aggregated pseudo-RGB-D labels. The comparative results are summarized in Tab. 3.

We observed a significant performance drop without the cross-modal attention module, highlighting the importance of effective information exchange between

Table 4 Ablation on selection of auxiliary branches. The best results are highlighted in **bold**.

Auxiliary Branches	CHAM				CAMO				COD10K			
	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
RGB	0.867	0.896	0.765	0.042	0.847	0.887	0.769	0.063	0.811	0.851	0.653	0.041
Depth	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	**0.867**	**0.682**	**0.036**

Table 5 Ablation on loss function. The best results are highlighted in **bold**.

Loss	Unified Structural			Depth Dual		CHAM				CAMO				COD10K			
	BCE	IoU	SSIM	BCE	DS	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$	$S_\alpha \uparrow$	$E_\xi^{\text{mean}} \uparrow$	$F_\beta^w \uparrow$	$\mathcal{M} \downarrow$
L_{us}	✓					0.843	0.864	0.726	0.058	0.843	0.840	0.708	0.077	0.784	0.821	0.647	0.054
L_{us}	✓	✓				0.848	0.873	0.735	0.051	0.850	0.881	0.760	0.065	0.791	0.834	0.653	0.049
L_{us}	✓	✓	✓			0.861	0.884	0.751	0.043	0.853	0.892	0.781	0.059	0.802	0.851	0.668	0.045
L_{us}+L_{dd}	✓	✓	✓	✓		0.864	0.905	0.778	0.038	0.858	0.896	0.786	0.057	0.812	0.861	0.674	0.039
L_{us}+L_{dd}	✓	✓	✓		✓	0.869	0.910	0.781	0.036	0.857	0.901	0.794	0.056	0.807	0.858	0.671	0.041
L_{us}+L_{dd}	✓	✓	✓	✓	✓	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	**0.867**	**0.682**	**0.036**

modalities. Using aggregated pseudo-RGB-D labels also mitigated inaccuracies from single-modality pseudo-labels. Fig. 6 corresponds to the quantitative results reported in Tab. 3 visually demonstrates that both CARF and aggregated pseudo-RGB-D labels effectively enhance our approach. However, when using only pseudo-RGB labels for prediction, the resulting boundaries become blurred and indistinct.

5.3.3 Study the Selection of Auxiliary Branches

In our model, we employ the prediction results derived from depth features as an auxiliary branch to enhance network supervision. To investigate the effectiveness of alternative modalities, we also considered using RGB features as an auxiliary branch for supervision. The experimental results, presented in Tab. 4, reveal that utilizing RGB prediction results as auxiliary branches is less effective compared to leveraging depth prediction results. This suggests that depth cues are more suitable for auxiliary supervision when segmenting camouflaged objects. Consequently, we conclude that additional spatial information, as provided by depth features, is more effective in accurately segmenting camouflaged objects.

5.3.4 Investigating the Role of Loss Functions

To investigate the contribution of each component in our proposed loss function, we conducted a comparative analysis presented in Tab. 5. Our findings indicate that as the number of supervised loss functions increases, the model's prediction accuracy improves significantly, validating the correctness and effectiveness of our loss function design. Furthermore, the introduction of Depth Dual Loss, which involves smoothing the depth map for auxiliary supervision, has a positive influence on the model's final prediction, underscoring its importance in enhancing performance.

5.3.5 Discussion on the Sensitivity to Superpixel Granularity and Algorithm Design

We set the parameters for superpixel segmentation to 50 and 100, respectively. Additionally, for Alg 1, we designed a variant in which pseudo-labels are generated directly

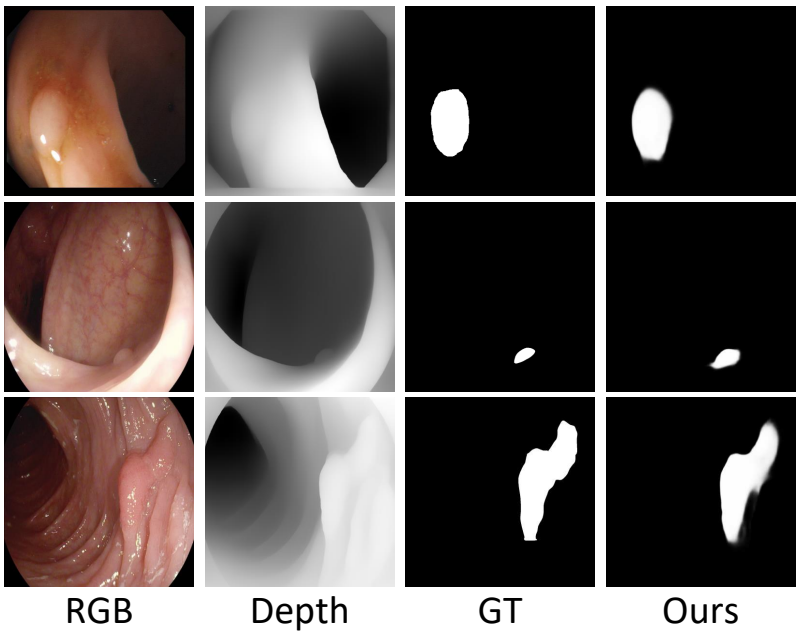


Fig. 7 Qualitative comparison on CVC-ColonDB, ETIS and KVASIR datasets.

Table 6 Quantitative comparison on CVC-ColonDB, ETIS, and Kvasir datasets.

Methods	Backbone	Sup.	Input	CVC-ColonDB				ETIS				Kvasir			
				$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$
PraNet$_{20}$ [64]	Res2Net50	F	RGB	0.820	0.847	0.699	0.043	0.794	0.808	0.600	0.031	0.915	0.944	0.885	0.030
C2FNet$_{22}$ [65]	Res2Net50	F	RGB	0.826	0.854	0.705	0.044	0.827	0.860	0.668	0.022	0.905	0.929	0.869	0.036
M2SNet$_{23}$ [66]	Res2Net50	F	RGB	0.838	0.872	0.736	0.041	0.845	0.875	0.677	0.020	0.923	0.947	0.892	0.028
SS$_{20}$ [62]	ResNet50	S	RGB	0.713	0.732	0.645	0.127	0.762	0.733	0.647	0.123	0.828	0.852	0.822	0.082
SCWS$_{21}$ [63]	ResNet50	S	RGB	0.787	0.758	0.674	0.082	0.731	0.768	0.646	0.085	0.831	0.860	0.837	0.078
CRNet$_{23}$ [12]	ResNet50	S	RGB	0.802	0.795	0.691	0.077	0.766	0.802	0.664	0.071	0.836	0.877	0.853	0.071
SAM$_{23}$ [67]	ViT	-	RGB	0.427	0.419	0.343	0.479	0.503	0.512	0.439	0.429	0.582	0.564	0.545	0.320
WS-SAM$_{23}$ [26]	ResNet50	S	RGB	0.816	0.839	-	0.043	0.797	0.849	-	0.037	0.877	0.917	-	0.046
SAM-COD+$_{24}$ [27]	Vit-H	S	RGB	0.839	0.848	0.737	0.036	0.841	0.849	0.651	0.020	0.861	0.930	0.902	0.029
PPLNet$_{25}$ [68]	ResNet50	S	RGB	0.812	0.801	0.697	0.067	0.764	0.807	0.672	0.069	0.838	0.880	0.862	0.074
Ours	PVT-v2-B2	S	RGB-D	0.810	0.849	0.651	0.038	0.803	0.803	0.554	0.026	0.875	0.907	0.810	0.043

Table 7 Ablation on sensitivity to superpixel granularity and algorithm design. The best results are highlighted in **bold**. "F": Foreground-based pseudo-labels; "B": Background-based pseudo-labels.

Superpixels	Algorithm	CHAM				CAMO				COD10K			
		$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$S_\alpha\uparrow$	$E_\xi^{\text{mean}}\uparrow$	$F_\beta^w\uparrow$	$\mathcal{M}\downarrow$
50	F	0.568	0.526	0.316	0.262	0.584	0.550	0.373	0.266	0.546	0.499	0.253	0.213
50	F+B	0.837	0.871	0.712	0.050	0.835	0.877	0.735	0.065	0.771	0.805	0.585	0.052
100	F	0.643	0.641	0.397	0.172	0.658	0.666	0.455	0.183	0.607	0.583	0.313	0.147
100	F+B	**0.871**	**0.914**	**0.786**	**0.034**	**0.859**	**0.903**	**0.799**	**0.054**	**0.815**	**0.867**	**0.682**	**0.036**

for the superpixel regions corresponding to the foreground in the labels, without considering background annotations, as compared in Table. 7. We observe that while increasing the number of superpixel blocks improves performance, the gain is limited. Increasing the number of superpixels further improves performance again, but the

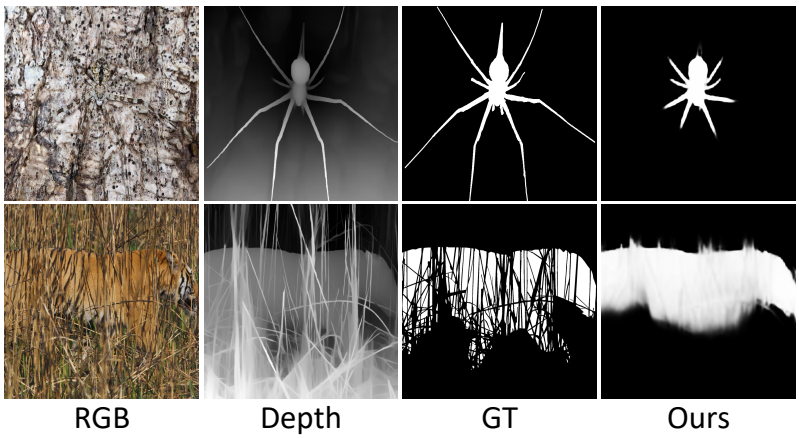


Fig. 8 Two examples of our method's failures that inaccurately predict complex boundaries.

improvement is diminishing. Moreover, incorporating background annotation information in the algorithm generates more accurate pseudo-labels, preventing the model from learning incorrect information. Therefore, the proposed Alg 1 is highly effective.

6 Discussion

Our method generally performs well, but it does have some limitations. Specifically, due to the imprecise scribble annotation of the foreground in the dataset, our method struggles in scenes with intricate boundary details. Fig. 8 presents two examples of where our method failed. The first row features a spider with slender legs; our method is only able to identify some of the leg joints adjacent to the spider's body and fails to capture the entire leg. The second row shows a tiger partially obscured by dense weeds. While our method successfully predicts the tiger's main body, it is unable to segment the weeds accurately. In future research, we will focus more on local details, particularly in scenarios with complex boundaries. We aim to achieve more precise segmentation of camouflaged objects under weakly supervised learning conditions.

7 Conclusion

We introduce SPMCNet, a weakly-supervised COD model leveraging scribble annotation. It comprises two core components: multi-modal feature extraction and pseudo-label mapping. For pseudo-label mapping, we harness superpixel segmentation and a tailored algorithm for camouflaged objects to create pseudo-labels. These pseudo-labels, aggregated across multiple modalities, yield reliable training data for our segmentation model. In terms of multi-modal feature extraction, we first employ a depth estimation model to obtain a depth map, then leverage multi-modal fusion to enhance the utilization of spatial features under camouflaged objects. Additionally, we design a specialized loss function to supervise both the main and auxiliary branches. Extensive experiments on the weakly-supervised COD task demonstrate that

our SPMCNet surpasses existing methods, even outperforming many fully-supervised approaches.

Declarations

- Funding: This work was supported in part by the Yunnan Province Fundamental Research Projects under Grant 202401AU070154.
- Conflict of interest: Conflict of interest The authors declare no competing interests.
- Data availability: The data can be obtained upon request from the authors.
- Code availability: The source code is publicly available at https://github.com/YDsheep/SPMCNet.git. The permanent DOI for the code repository is https://doi.org/10.5281/zenodo.18872858.

References

[1] 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)

[2] Liu, M., Di, X.: Extraordinary mhnet: Military high-level camouflage object detection network and dataset. Neurocomputing **549**, 126466 (2023)

[3] Liu, Y., Che, S., Ai, L., Song, C., Zhang, Z., Zhou, Y., Yang, X., Xian, C.: Camouflage detection: Optimization-based computer vision for alligator sinensis with low detectability in complex wild environments. Ecological Informatics **83**, 102802 (2024)

[4] Chen, T., Zhu, L., Ding, C., Cao, R., Wang, Y., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam fails to segment anything?–sam-adapter: Adapting sam in underperformed scenes: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2304.09148 (2023)

[5] Chen, T., Lu, A., Zhu, L., Ding, C., Yu, C., Ji, D., Li, Z., Sun, L., Mao, P., Zang, Y.: Sam2-adapter: Evaluating & adapting segment anything 2 in downstream tasks: Camouflage, shadow, medical image segmentation, and more. arXiv preprint arXiv:2408.04579 (2024)

[6] Zheng, J., Lin, C., Sun, J., Zhao, Z., Li, Q., Shen, C.: Physical 3d adversarial attacks against monocular depth estimation in autonomous driving. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 24452–24461 (2024)

[7] Gupta, A.K., Seal, A., Khanna, P., Herrera-Viedma, E., Krejcar, O.: Almnet: Adjacent layer driven multiscale features for salient object detection. IEEE Transactions on Instrumentation and Measurement **70**, 1–14 (2021)

[8] Wan, B., Zhou, X., Zheng, B., Yin, H., Zhu, Z., Wang, H., Sun, Y., Zhang, J., Yan, C.: Lfrnet: Localizing, focus, and refinement network for salient object detection of surface defects. IEEE Transactions on Instrumentation and Measurement **72**, 1–12 (2023)

[9] Jiang, X., Cai, W., Ding, Y., Wang, X., Hong, D., Yang, Z., Gao, W.: Camouflaged object segmentation based on joint salient object for contrastive learning. IEEE Transactions on Instrumentation and Measurement **72**, 1–16 (2023)

[10] Li, X., Long, S., Yang, J., Lei, J., Li, S., Zhang, J., Cohen, L.D.: Dpsnet: A detail perception synergistic network for camouflaged object detection. IEEE Transactions on Instrumentation and Measurement **74**, 1–14 (2025)

[11] Wu, Z., Paudel, D.P., Fan, D.-P., Wang, J., Wang, S., Demonceaux, C., Timofte, R., Van Gool, L.: Source-free depth for object pop-out. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 1032–1042 (2023)

[12] 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, vol. 37, pp. 781–789 (2023)

[13] Hu, X., Bao, X., Wei, G., Li, Z.: Human-pose estimation based on weak supervision. Virtual Reality & Intelligent Hardware **5**(4), 366–377 (2023)

[14] Zhang, Y., Wu, C.: Unsupervised camouflaged object segmentation as domain adaptation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4334–4344 (2023)

[15] 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)

[16] Sun, Y., Chen, G., Zhou, T., Zhang, Y., Liu, N.: Context-aware cross-level fusion network for camouflaged object detection. arXiv preprint arXiv:2105.12555 (2021)

[17] Sun, Y., Wang, S., Chen, C., Xiang, T.-Z.: Boundary-guided camouflaged object detection. arXiv preprint arXiv:2207.00794 (2022)

[18] Zhou, T., Zhou, Y., Gong, C., Yang, J., Zhang, Y.: Feature aggregation and propagation network for camouflaged object detection. IEEE Transactions on Image Processing **31**, 7036–7047 (2022)

[19] Zhu, H., Li, P., Xie, H., Yan, X., Liang, D., Chen, D., Wei, M., Qin, J.: I can find you! boundary-guided separated attention network for camouflaged object detection. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol.

36, pp. 3608–3616 (2022)

[20] Ye, Q., Li, Q., Huo, G., Liu, Y., Zhou, Y.: Boundary-guided multi-scale refinement network for camouflaged object detection. The Visual Computer **41**(8), 6271–6297 (2025)

[21] Zhang, M., Xu, S., Piao, Y., Shi, D., Lin, S., Lu, H.: Preynet: Preying on camouflaged objects. In: Proceedings of the 30th ACM International Conference on Multimedia, pp. 5323–5332 (2022)

[22] 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)

[23] Zhong, Y., Li, B., Tang, L., Kuang, S., Wu, S., Ding, S.: Detecting camouflaged object in frequency domain. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 4504–4513 (2022)

[24] He, C., Li, K., Zhang, Y., Tang, L., Zhang, Y., Guo, Z., Li, X.: Camouflaged object detection with feature decomposition and edge reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 22046–22055 (2023)

[25] Zhai, W., Cao, Y., Zhang, J., Zha, Z.-J.: Exploring figure-ground assignment mechanism in perceptual organization. Advances in Neural Information Processing Systems **35**, 17030–17042 (2022)

[26] 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)

[27] Chen, H., Wei, P., Guo, G., Gao, S.: Sam-cod+: Sam-guided unified framework for weakly-supervised camouflaged object detection. IEEE Transactions on Circuits and Systems for Video Technology (2024)

[28] Zhang, J., Zhang, R., Shi, Y., Cao, Z., Liu, N., Khan, F.S.: Learning camouflaged object detection from noisy pseudo label. In: European Conference on Computer Vision, pp. 158–174 (2024). Springer

[29] Niu, Y., Yang, L., Xu, R., Li, Y., Chen, Y.: Minet: Weakly-supervised camouflaged object detection through mutual interaction between region and edge cues. In: Proceedings of the 32nd ACM International Conference on Multimedia, pp. 6316–6325 (2024)

[30] Chen, H., Shao, D., Guo, G., Gao, S.: Just a hint: Point-supervised camouflaged object detection. In: European Conference on Computer Vision, pp. 332–348

(2025). Springer

[31] Zhang, M., Zhou, J., Miao, T., Zhao, Y., Si, X., Zhang, J.: Joint-learning: A robust segmentation method for 3d point clouds under label noise. Computer Animation and Virtual Worlds **36**(3), 70038 (2025)

[32] Luo, Y., Shao, F., Xie, Z., Wang, H., Chen, H., Mu, B., Jiang, Q.: Hfmdnet: Hierarchical fusion and multi-level decoder network for rgb-d salient object detection. IEEE Transactions on Instrumentation and Measurement (2024)

[33] Wen, Y., Luo, B., Shi, W., Ji, J., Cao, W., Yang, X., Sheng, B.: Sat-net: structure-aware transformer-based attention fusion network for low-quality retinal fundus images enhancement. IEEE Transactions on Multimedia (2025)

[34] Tu, Z., Xia, T., Li, C., Wang, X., Ma, Y., Tang, J.: Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia **22**(1), 160–173 (2019)

[35] Fu, K., Fan, D.-P., Ji, G.-P., Zhao, Q.: Jl-dcf: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 3052–3062 (2020)

[36] Zhang, Q., Xiao, T., Huang, N., Zhang, D., Han, J.: Revisiting feature fusion for rgb-t salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **31**(5), 1804–1818 (2020)

[37] Li, A., Mao, Y., Zhang, J., Dai, Y.: Mutual information regularization for weakly-supervised rgb-d salient object detection. IEEE Transactions on Circuits and Systems for Video Technology **34**(1), 397–410 (2023)

[38] Zhan, Y., Zeng, Z., Liu, H., Tan, X., Tian, Y.: Mambasod: Dual mamba-driven cross-modal fusion network for rgb-d salient object detection. Neurocomputing, 129718 (2025)

[39] Chen, K., Zhou, Z., Li, K., Su, T., Zhang, Z., Liu, J., Ying, C.: Red green blue-depth salient object detection based on multi-scale refinement and cross-modalities fusion network. The Visual Computer **41**(10), 7617–7640 (2025)

[40] Xiang, M., Zhang, J., Lv, Y., Li, A., Zhong, Y., Dai, Y.: Exploring depth contribution for camouflaged object detection. arXiv preprint arXiv:2106.13217 (2021)

[41] Yang, L., Kang, B., Huang, Z., Zhao, Z., Xu, X., Feng, J., Zhao, H.: Depth anything v2. arXiv preprint arXiv:2406.09414 (2024)

[42] Pang, Y., Zhao, X., Zhang, L., Lu, H.: Caver: Cross-modal view-mixed transformer for bi-modal salient object detection. IEEE Transactions on Image Processing **32**, 892–904 (2023)

[43] Vaswani, A.: Attention is all you need. Advances in Neural Information Processing Systems (2017)

[44] Wei, Y., Ji, S.: Scribble-based weakly supervised deep learning for road surface extraction from remote sensing images. IEEE Transactions on Geoscience and Remote Sensing **60**, 1–12 (2021)

[45] Chen, Q., Hong, Y.: Scribble2d5: Weakly-supervised volumetric image segmentation via scribble annotations. In: International Conference on Medical Image Computing and Computer-Assisted Intervention, pp. 234–243 (2022). Springer

[46] Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(11), 2274–2282 (2012)

[47] Ru, L., Zhan, Y., Yu, B., Du, B.: Learning affinity from attention: End-to-end weakly-supervised semantic segmentation with transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16846–16855 (2022)

[48] Su, H., Jampani, V., Sun, D., Gallo, O., Learned-Miller, E., Kautz, J.: Pixel-adaptive convolutional neural networks. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11166–11175 (2019)

[49] Skurowski, P., Abdulameer, H., Błaszczyk, J., Depta, T., Kornacki, A., Kozieł, P.: Animal camouflage analysis: Chameleon database. Unpublished manuscript **2**(6), 7 (2018)

[50] 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)

[51] 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)

[52] 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)

[53] 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)

[54] Fan, D.-P., Ji, G.-P., Qin, X., Cheng, M.-M.: Cognitive vision inspired object segmentation metric and loss function. Scientia Sinica Informationis **6**(6), 5 (2021)

[55] 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)

[56] 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)

[57] Dozat, T.: Incorporating nesterov momentum into adam. In: International Conference on Learning Representations (2016)

[58] Ji, G.-P., Zhu, L., Zhuge, M., Fu, K.: Fast camouflaged object detection via edge-based reversible re-calibration network. Pattern Recognition **123**, 108414 (2022)

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

[60] Zhang, J., Zhang, T., Dai, Y., Harandi, M., Hartley, R.: Deep unsupervised saliency detection: A multiple noisy labeling perspective. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 9029–9038 (2018)

[61] Nguyen, T., Dax, M., Mummadi, C.K., Ngo, N., Nguyen, T.H.P., Lou, Z., Brox, T.: Deepusps: Deep robust unsupervised saliency prediction via self-supervision. Advances in Neural Information Processing Systems **32** (2019)

[62] 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)

[63] 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, vol. 35, pp. 3234–3242 (2021)

[64] 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 (2020). Springer

[65] Chen, G., Liu, S.-J., Sun, Y.-J., Ji, G.-P., Wu, Y.-F., Zhou, T.: Camouflaged object detection via context-aware cross-level fusion. IEEE Transactions on Circuits and Systems for Video Technology **32**(10), 6981–6993 (2022)

[66] Zhao, X., Jia, H., Pang, Y., Lv, L., Tian, F., Zhang, L., Sun, W., Lu, H.: M2snet: Multi-scale in multi-scale subtraction network for medical image segmentation. arXiv preprint arXiv:2303.10894 (2023)

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

[68] Han, T., Wang, X., Dong, Q., Huang, M., Jia, J., Zhang, F.: Weakly supervised camouflaged object detection as progressive perception learning. Knowledge-Based Systems, 113993 (2025)