RA-COD: Retrieval-Augmented Camouflaged Object Detection

Ji Du, Jiesheng Wu, Desheng Kong, Fangwei Hao, Jing Xu, *Member, IEEE*, and Ping Li, *Member, IEEE*

***Abstract*—Camouflaged Object Detection (COD) is pivotal for segmenting objects that seamlessly blend into their surroundings. While prior endeavors demonstrate impressive performance through training on predefined labels, they heavily rely on labor-intensive data annotation and struggle to adapt to open-world scenarios. In this light, we propose RA-COD, a training-free paradigm that enables COD by retrieving the most similar samples from the prototype repository. The efficacy of RA-COD hinges on 1) *capturing the nuanced resemblance between objects and their environments* and 2) *excelling in dense prediction tasks*. To achieve (1), the crux lies in ensuring diversity and discriminability within the prototype repository. In this context, we propose GenPro, an automated pipeline for crafting Generative Prototypes. GenPro integrates a range of foundation models, including the Diffusion Model, Vision-Language Model, Segment Anything Model (SAM), and DINOv2, in a complementary manner that synergistically generates diverse and distinguishable prototype samples. To achieve (2), we propose C2F to retrieve camouflaged objects in a Coarse-to-Fine regime. We commence with pixel-level retrieval in the feature space, which generates a coarse mask that effectively captures class discrimination and object localization. Further refinement is achieved by extracting bounding boxes from this coarse mask to prompt SAM in generating mask proposals for region-level retrieval. Evaluations on four benchmarks showcase that RA-COD achieves state-of-the-art performance compared to existing training-free methods.**

***Index Terms*—COD, retrieval-augmented, diffusion models, VLMs, SAM.**

I. Introduction

CAMOUFLAGED object detection (COD) has garnered growing research interest ascribed to its more challenging characteristics compared to general segmentation tasks [1]. The predominant focus of research lies in fully supervised learning on annotated datasets. These endeavors have significantly pushed the boundaries of COD research by devising elaborate modules [2], [3], [4], [5], [6], [7], incorporating extra information [8], [9], [10], [11], [12], [13], [14], or adopting novel learning paradigms [15], [16], [17]. Despite the progress, these methods suffer from reliance on time-consuming and labor-intensive data annotation processes, particularly for COD datasets. To remedy this, weakly supervised COD capitalizes on sparse labels, e.g., scribbles [18] or pseudo masks [19], to segment camouflaged objects. However, they still require labels and exhibit significant performance degradation compared to fully supervised methods.

With the remarkable success achieved by vision and vision-language foundation models in downstream tasks, there is a growing interest in leveraging more powerful external models and knowledge to implement COD without training. This line of research primarily focuses on leveraging SAM [20] to segment camouflaged objects. Pioneering efforts [21], [22] showcase that it is non-trivial to generalize SAM to COD without any task-related modification. To this end, vision-language models (VLMs) are adopted to acquire the coordinates of objects to prompt SAM for segmentation [23], [24]. In addition, GenSAM [25] extracts points from class activation maps of CLIP [26], which serve as prompts for SAM. However, these methods either suffer from hallucinations [27] of VLMs not being able to localize objects accurately or get noisy activations due to the struggling performance of CLIP on dense prediction tasks. Despite being training-free, there is considerable scope for these methods to narrow the gap with fully supervised learning approaches.

In contrast to the above approaches, we propose a new training-free paradigm termed Retrieval-Augmented Camouflaged Object Detection (RA-COD), as depicted in Fig. 1. Our motivation stems from the aspiration to distinguish foreground from background by directly comparing test samples with those in the prototype repository. While conceptually straightforward, RA-COD confronts two key challenges: ❶ *how to tailor a high-quality prototype repository containing diverse and well-differentiated samples for COD*, and ❷ *how to integrate high-level retrieval matching into low-level dense prediction tasks.*

To address challenge ❶, we present GenPro, a pipeline for automatic generation of diverse and distinguishable sample prototypes. The entire pipeline comprises three stages: category acquisition, sample image generation, and prototype repository establishment. In the initial stage, to mitigate the

Received 10 June 2024; revised 1 October 2025; accepted 1 May 2026. Date of publication 14 May 2026; date of current version 19 May 2026. This work was supported in part by the Major Program of the National Natural Science Foundation of China under Grant 62233011; and in part by The Hong Kong Polytechnic University under Grant P0048387, Grant P0044520, Grant P0049586, and Grant P0050657. The associate editor coordinating the review of this article and approving it for publication was Dr. Fabrizio Guerrini. *(Corresponding authors: Jiesheng Wu; Jing Xu; Ping Li.)*

Ji Du is with the College of Artificial Intelligence, Nankai University, Tianjin 300071, China, and also with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: duji@mail.nankai.edu.cn).

Jiesheng Wu, Desheng Kong, Fangwei Hao, and Jing Xu are with the College of Artificial Intelligence, Nankai University, Tianjin 300071, China (e-mail: jasonwu@mail.nankai.edu.cn; kongds@mail.nankai.edu.cn; haofangwei@mail.nankai.edu.cn; xujing@nankai.edu.cn).

Ping Li is with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong (e-mail: p.li@polyu.edu.hk).

Data is available on-line at https://github.com/xiaohainku/RA-COD

Digital Object Identifier 10.1109/TIP.2026.3691679

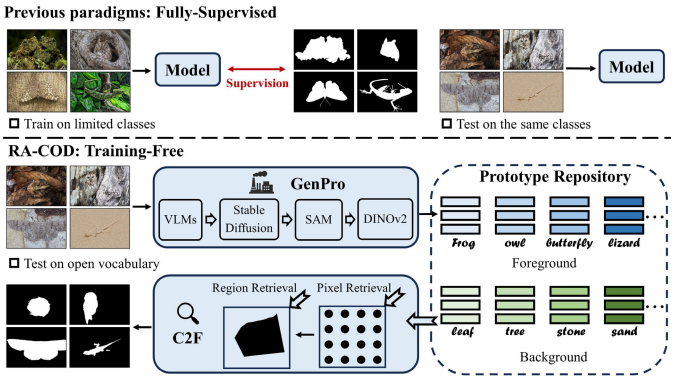


Fig. 1. Comparison of our RA-COD with the previous fully supervised learning paradigm. The prior paradigm demands training on meticulously labeled datasets, which encompass a limited number of categories. Based on generative prototypes (GenPro) as well as the innovative retrieval scheme (C2F), RA-COD segments camouflaged objects by retrieving the categories in the prototype repository that are most similar to the target. Being free of training, RA-COD could generalize to the segmentation of unknown categories easily.

scarcity of systematic category labels in COD datasets, we leverage Vision-Language Models (VLMs) such as LLaVA [28] to identify categories of camouflaged objects, environmental elements, and surrounding objects. Here, camouflaged objects are categorized as foreground, while environmental and surrounding objects serve as background categories. In the subsequent stage, harnessing the remarkable conditional generation capabilities of diffusion models [29], [30], [31], [32], [33], we generate a diverse array of sample images for each foreground and background category. To establish the prototype repository, a conventional approach involves mapping images corresponding to each category into embedding vectors using a feature extractor like DINOv2 [34]. However, images generated by diffusion models may include irrelevant environmental or object features alongside the target categories, potentially leading to insufficient differentiation between prototype samples. To mitigate this issue, we leverage cross-attention maps from diffusion models and SAM to generate masks for the target categories. These masks are then employed to filter out irrelevant features in the feature space, ensuring the creation of distinct and representative prototype samples.

To address challenge ❷, one of the most intuitive ways is to perform pixel-level retrieval in the feature space and assign foreground or background labels to each pixel. However, due to the low resolution of the feature space, the segmentation after up-sampling struggles to capture fine details. Another straightforward approach is to provide SAM with uniformly sampled points to generate a set of mask proposals, which are then applied to the feature space for region-level retrieval. Nevertheless, previous work has demonstrated that SAM encounters difficulty in distinguishing camouflaged objects from the environment without explicit prompts. In this context, we propose C2F, a Coarse-to-Fine retrieval scheme. Recognizing that pixel-level retrieval offers well-defined class discrimination and localization, we initially employ pixel-level retrieval to obtain coarse segmentation. Subsequently, a set of bounding boxes is extracted from the coarse segmentation to prompt SAM to generate mask proposals. These masks are utilized in region-level retrieval for generating fine segmentation.

By addressing challenges ❶ and ❷, RA-COD emerges as a possible paradigm for COD. Our proposed GenPro and C2F seamlessly integrate various foundation models such as VLMs, diffusion models, SAM, and DINOv2, synergistically providing diverse expertise from different angles for COD. Before inference, we utilize GenPro to construct the prototype repository, which comprises two category labels: foreground and background. At inference, C2F is employed to conduct pixel-level and region-level retrieval of images mapped to the feature space by a feature extractor. By computing cosine similarity, pixels or regions that are most similar to the prototype samples are assigned corresponding labels.

Extensive experiments on four benchmarks demonstrate that RA-COD performs comparably to the state-of-the-art fully supervised learning methods. In addition, RA-COD substantially outperforms all weakly supervised learning and training-free foundation-model-based methods.

In summary, our contributions are three-fold:

- We propose GenPro for generative prototypes, which incorporates a series of expert models across different domains to generate diverse and differentiated prototypes.
- We propose C2F, a retrieval strategy for transforming coarse-grained retrieval into fine-grained dense prediction tasks.
- With GenPro and C2F, we propose RA-COD, a new paradigm to segment camouflaged objects through retrieval in a training-free regime. Comprehensive experiments substantiate the effectiveness of our proposed RA-COD in COD.

II. Related Work

A. Camouflaged Object Detection

The central principle behind Camouflaged Object Detection (COD) is to discern and isolate objects that closely resemble their surroundings [1], [2], [35]. To address this challenging task, most existing work fixates on fully supervised training on well-annotated datasets. Among this, designing novel modules for mining details is the thrust of the research [1], [2], [3], [4], [5], [6], [7], [36], [37], [38], [39]. In addition, extra information, including edges [11], [12], texture [8], frequency [10], [40], depth [9], categories [14] and referring images [13], is exploited to unearth critical clues. Another line of work is dedicated to adopting new learning paradigms that are attuned to the characteristics of COD. ICEG+ [15] adopts adversarial training to augment data, thus alleviating overfitting of the model. To address the boundary uncertainty, diffCOD [17] adopts the diffusion denoising paradigm to directly generate segmentation masks. Inspired by biological visual mechanisms, HitNet [4], SegMaR [41], and ZoomNet [42] progressively localize and segment camouflaged objects by iterative decoding, or employing a multi-scale strategy. While these efforts have significantly spurred the development of COD, they all inevitably entail optimizations on large amounts of labeled data.

Unlike these methods, our method is training-free. In this paper, "training-free COD" denotes a paradigm in which SAM is guided by adaptive prompts to segment camouflaged objects without any training on COD datasets. This qualifies as a zero-shot setting because it avoids training on COD data; however, it is not unsupervised, as the underlying SAM is trained in a supervised manner on large-scale datasets.

B. Foundation Models

Trained on large-scale datasets, foundation models, such as vision, vision-language, and diffusion models, are being increasingly deployed in various downstream tasks [43], [44], [45]. However, the application of foundation models to COD is still in its nascent stage, primarily attributed to the inherent challenges associated with COD.

Existing work utilizing foundation models for COD revolves around leveraging VLMs to generate different prompts for SAM. MLKG [14] employs VLMs to generate object descriptions from various perspectives, which are then fed to SAM for fine-tuning. Similarly, MMCPF [23] and GenSAM [25] resort to VLMs to extract object coordinates or categories, complemented by additional models for prompt refinement. However, these approaches either necessitate training or suffer from inaccurate prompts credited to hallucinations of VLMs.

Another strand of work focuses on diffusion models. As a branch of generative models, diffusion models exhibit immensely powerful image generation [29], [30], [31], [32], [33] and editing [46], [47], [48] capabilities. In addition, diffusion models have been broadly explored for downstream tasks such as semantic segmentation [45], open-vocabulary segmentation [49], [50], object localization [51], object detection [52], and data synthesis [53], [54]. In the realm of COD, CamoDiffusion [55] and diffCOD [17] take the original image as the condition and directly generate masks by diffusion denoising. CamDiff [56] augments camouflage datasets by adding realistic salient objects into scenes. LAKE-RED [57] utilizes diffusion models to generate multi-scene camouflaged images from the image-inpainting perspective.

Different from these methods, our RA-COD provides a novel paradigm that assembles various foundation models in a synergistic and complementary manner to solve the challenging COD task.

C. Retrieval-Augmented Methods

In the Natural Language Processing community, Retrieval-Augmented Generation (RAG) leverages retrieval mechanisms to enable the introduction of external knowledge into the generation process, resulting in more accurate, varied, and relevant texts [58], [59], [60]. Similar to Transformer [61], the success of RAG in NLP quickly proliferated into computer vision. RAG has demonstrated its superiority in various vision domains, including zero-shot classification [62], semantic segmentation [63], open-vocabulary segmentation [50], [64], [65], and continual learning [66], [67]. Sharing similar spirits with these efforts, our method provides new perspectives in terms of generating distinguishable prototype samples and applying retrieval to fine-grained segmentation tasks.

III. Methodology

In this paper, we introduce a new paradigm named RA-COD to segment camouflaged objects via retrieval. Given an image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, RA-COD aims to assign each pixel a foreground or background label by retrieving the most similar samples in the prototype repository and generate a segmentation map $\mathbf{M} \in \mathbb{R}^{H\times W\times 1}$. As demonstrated in Fig. 2, RA-COD consists of two main stages. At the first stage, GenPro crafts a wealth of sample prototypes for each category of camouflaged objects, surrounding objects, and environment. These samples are further categorized into foreground (camouflaged objects) and background (surrounding objects, environment) to form the final prototype repository. At the retrieval stage, we first carry out point-wise retrieval in the embedding space, attributing corresponding labels to the pixels by matching the prototype samples that are most similar to the pixels. The coarse segmentation maps produced from this process are used to extract a set of boxes, which are employed in the SAM to generate mask proposals. Given the mask proposals, we further perform region-level retrieval to filter noisy proposals to get the final segmentation.

A. Generative Prototypes

We propose GenPro as a pipeline for generative prototypes, including fetching categories, generating images, and establishing the prototype repository.

1) Fetch Categories: COD involves localizing camouflaged objects from the complex environment and segmenting them from similar surrounding objects. If the prototype repository merely contains camouflaged object categories, retrieval could be accomplished by setting a similarity threshold. This not only introduces extra hyperparameters but may result in misclassification due to the limited variety of samples. In this light, we define three categories: Camouflaged Objects (CO), Surrounding Objects (SO), and the ENvironment (EN). Inspired by the promising performance of VLMs [28], [68], [69] in image captioning, VQA, and text-image retrieval, we adopt VLMs for automatically fetching categories.

Given an image $\mathbf{I}_i$ in the test set $T = \{\mathbf{I}_i\}_{i=1}^{N}$, ($N$ denotes the total number of images), the three categories could be derived by

$$C_i^{co} = \text{VLM}(\mathbf{I}_i, P_{co}), \tag{1}$$

$$C_i^{so} = \text{VLM}(\mathbf{I}_i, P_{so}), \tag{2}$$

$$C_i^{en} = \text{VLM}(\mathbf{I}_i, P_{en}), \tag{3}$$

where P_{co}, P_{so}, and P_{en} refer to the prompts required by the VLMs.

Furthermore, to adapt to the binary classification problem, we take the CO category as the foreground class F and the SO and EN categories as the background classes B. Aggregating the categories of each test image, we get the category set $C = \{(F : \{C_i^{co}\}), (B : \{C_i^{so}, C_i^{en}\})\}$.

2) Generate Images: Camouflaged objects of the same category may vary in shape, color, and size. This requires sufficiently diverse prototypes to achieve favorable retrieval. Intuitively, we can craft the prototype repository by sampling images from the real world. However, it is a tedious task to

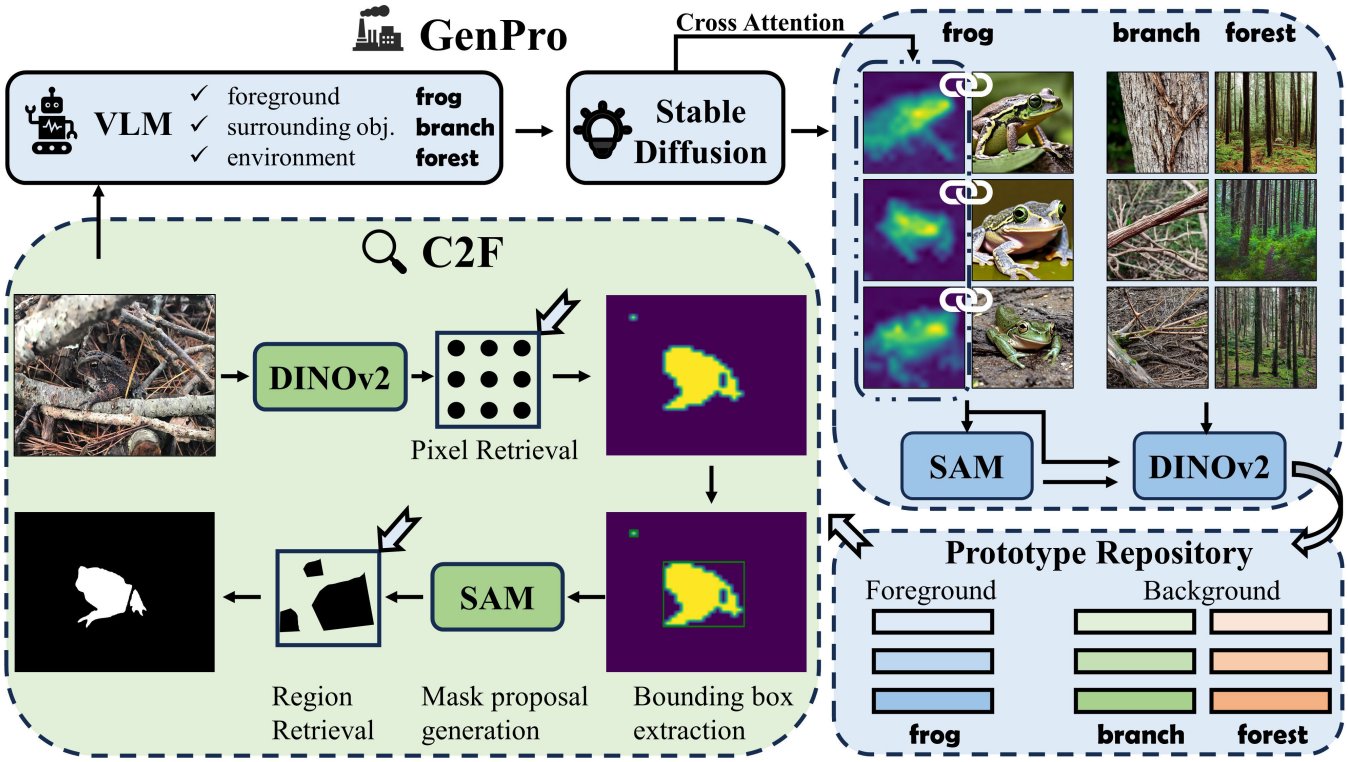


Fig. 2. The overall framework of RA-COD, consisting of the prototype generation pipeline GenPro and the retrieval segmentation scheme C2F. At the stage of GenPro, VLMs are first employed to obtain the categories of foreground, surrounding objects, and environment, which are fed into the diffusion model to generate multiple samples. DINOv2 is then utilized to map each sample into an embedding vector. In particular, for foreground samples, the corresponding attention maps are extracted for enhancing discrimination. During the C2F stage, the coarse segmentation generated by pixel-level retrieval is used to generate candidate bounding boxes, which are processed through SAM to generate mask proposals. Region-level retrieval is used to filter these masks and generate fine segmentation. This figure illustrates an example of a single test image. For multiple images, the foreground and background of the prototype repository entail different categories.

sample a large number of real images that need to be well aligned with the categories. Meanwhile, for segmentation in the wild, it is not convenient to collect new samples for new categories. To this end, we resort to the diffusion model to synthesize images. Benefiting from its powerful conditional generation capability, we could sample images with diverse scenarios and varying categories. For each category in the category set C, we provide the diffusion model with the prompt `a photo of [cls]` to generate a fixed number of images.

3) Prototype Repository: We adopt DINOv2 [34] to map each sample image to an embedding vector. Let $\mathbf{I_s} \in \mathbb{R}^{H\times W\times 3}$ denote the synthesized image from the diffusion model. We extract the feature map of the last layer and reshape it to $\mathbf{F_s} \in \mathbb{R}^{H'\times W'\times d}$ (d denotes the embedding dimension). Global average pooling (GAP) is applied to the feature map to generate the embedding vector $\mathbf{v_s} \in \mathbb{R}^d$. Let $f^k_{i,j}$ denote the k-th channel dimension at spatial point (i, j) of $\mathbf{F_s}$, the k-th element v^k_s of $\mathbf{v_s}$ could be derived by

$$v^k_s = \frac{1}{H'W'}\sum_{i=1}^{H'}\sum_{j=1}^{W'} f^k_{i,j}. \tag{4}$$

Integrating the embedding vectors of all the categories, we attain the prototype repository $P = \{(F : \{\mathbf{v}^{co}_i\}), (B : \{\mathbf{v}^{so}_i, \mathbf{v}^{en}_i\})\}$, where F and B indicate foreground and background, respectively.

In particular, for foreground categories, images sampled from the diffusion model often contain irrelevant environment, which may reduce the distinction between foreground and background samples, leading to misclassification of the background as camouflaged objects during retrieval.

To address this, we first extract cross-attention maps from the diffusion model in a similar vein to DiffuMask [53]. Specifically, we choose Stable Diffusion [31] as the diffusion model. Stable Diffusion consists of three key components: a text encoder $\mathcal{T}$, a variational autoencoder (including the encoder $\mathcal{E}$ and decoder $\mathcal{D}$), and a denoising UNet $\mathcal{U}$. Let $\mathbf{I_s} \in \mathbb{R}^{H\times W\times 3}$ and y denote the synthesized image and the corresponding text prompt, respectively. The encoder $\mathcal{E}$ encodes $\mathbf{I_s}$ into the latent space $\mathbf{z} = \mathcal{E}(\mathbf{I_s})$, and the diffusion process (adding noise) could be formulated as

$$\mathbf{z_t} \triangleq \sqrt{\bar{\alpha}_t}\mathbf{z} + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \epsilon_i \sim \mathcal{N}(0, 1), \tag{5}$$

where $\bar{\alpha}_t$ refers to the hyperparameter to control the degree of noise $\boldsymbol{\epsilon}$ addition at diffusion step t.

In the denoising UNet, the interaction between images and text prompts occurs in cross-attention layers of different resolutions. Let $r \in \{\times 8, \times 16, \times 32, \times 64\}$ represent the resolution of the cross-attention layer. At diffusion step t, the cross attention maps (CA^t_r) could be derived by

$$\mathrm{CA}^t_r = \mathrm{Softmax}\left(\frac{\mathbf{QK}^\top}{\sqrt{d'}}\right), \tag{6}$$

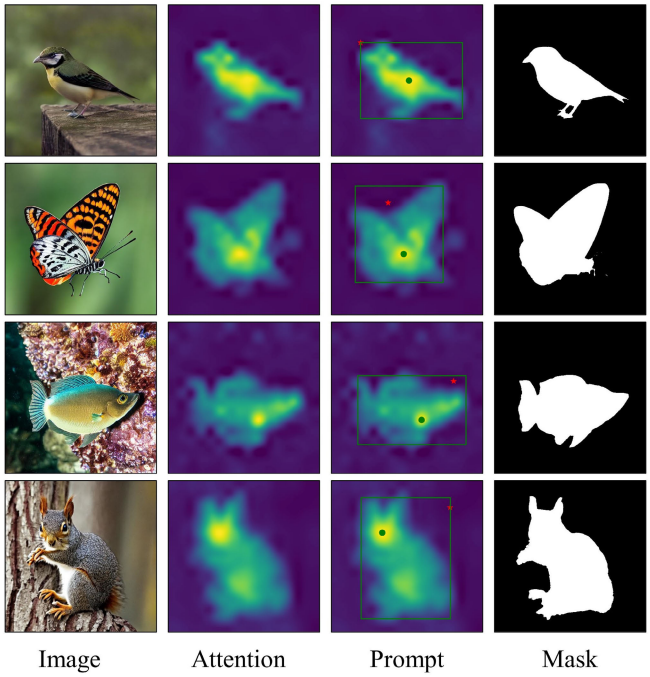


Fig. 3. Generating masks for foreground categories. We first extract cross-attention maps corresponding to the foreground categories from the diffusion model. A threshold approach [70] is adopted to extract bounding boxes from the attention maps. In addition, we also sample a pair of positive (•) and negative (⋆) points: the positive point is the location with the highest value within the bounding box, and the negative point is the location with the lowest value. These prompts are input to SAM for segmentation of the foreground.

$$\mathbf{Q} = \mathbf{W_Q}\mathbf{z_t}, \mathbf{K} = \mathbf{W_K}\mathcal{T}(y). \tag{7}$$

$\mathbf{W_Q}$ and $\mathbf{W_K}$ represent the learned projections. By integrating cross-attention maps at multiple diffusion steps and resolutions, the final attention maps could be derived by

$$\mathrm{CA} = \frac{1}{RT}\sum_{r,t}\mathrm{Upsample}(\mathrm{CA}_r^t), \tag{8}$$

where R and T refer to the number of resolutions and time steps, respectively.

After this, bounding boxes and high-confidence points are sampled from the attention maps CA to prompt SAM to output the foreground mask, as exemplified by Fig. 3. Based on the reshaped foreground mask $\mathbf{M_s} \in \mathbb{R}^{H'\times W'\times 1}$, we apply it to the features $\mathbf{F_s}$ and adopt mask average pooling to get the refined embedding vector. In addition, by reversing $\mathbf{M_s}$, we derive extra environment samples to complement the background prototypes. Let $m_{i,j}$ denote the spatial point of $\mathbf{M_s}$ at (i, j). Let v_f^k and v_b^k be the k-th elements of the foreground embedding $\mathbf{v_f} \in \mathbb{R}^d$ and background embedding $\mathbf{v_b} \in \mathbb{R}^d$, respectively. The above process could be formulated as

$$v_f^k = \frac{\sum_{i=1}^{H'}\sum_{j=1}^{W'} m_{i,j} f_{i,j}^k}{\sum_{i=1}^{H'}\sum_{j=1}^{W'} m_{i,j}}, v_b^k = \frac{\sum_{i=1}^{H'}\sum_{j=1}^{W'}(1-m_{i,j}) f_{i,j}^k}{\sum_{i=1}^{H'}\sum_{j=1}^{W'} 1-m_{i,j}}. \tag{9}$$

B. Coarse-to-Fine Retrieval

To apply retrieval to the segmentation task, one of the most straightforward ways is to adopt a point-wise approach to match the most similar categories for each pixel. Given the test image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$, the feature maps $\mathbf{F} \in \mathbb{R}^{H'\times W'\times d}$ could be derived by DINOv2. Each spatial point of $\mathbf{F}$ corresponds to an embedding vector $\mathbf{f}_i \in \mathbb{R}^d, i = 1, 2, \ldots, H'W'$. For a sample $\mathbf{p}_j \in \mathbb{R}^d$ in the prototype repository P, the cosine similarity between $\mathbf{f}_i$ and $\mathbf{p}_j$ could be derived by

$$s_{ij} = \frac{\mathbf{f}_i \mathbf{p}_j}{||\mathbf{f}_i||||\mathbf{p}_j||}. \tag{10}$$

The index of the sample that is most similar to $\mathbf{f}_i$ could be derived by

$$j^* = \arg\max_j \frac{\mathbf{f}_i \mathbf{p}_j}{||\mathbf{f}_i||||\mathbf{p}_j||}. \tag{11}$$

After that, the corresponding category of $\mathbf{p}_{j^*}$ is assigned to $\mathbf{f}_i$. By point-wise retrieval, the low-resolution segmentation map $\mathbf{M}_\mathrm{c}^{\mathrm{low}} \in \mathbb{R}^{H'\times W'\times 1}$ is generated and could be upsampled to yield the final mask $\mathbf{M}_\mathrm{c} \in \mathbb{R}^{H\times W\times 1}$. However, as the retrieval is performed on a low-resolution feature space, it is difficult for the upsampled mask $\mathbf{M}_\mathrm{c}$ to capture detailed information such as edges and textures, as shown in Fig. 2.

Another approach is to adopt the off-the-shelf segmentation models [20], [71] or superpixel algorithms [73] to produce a series of mask proposals for region-level retrieval in feature space. For instance, point prompts sampled uniformly from an image can be input into SAM to generate a set of fine-grained masks.[1] These proposals are applied to the feature map, and the corresponding embedding vectors are generated by mask average pooling. Through retrieval, each proposal is assigned a category label, foreground or background. The final segmentation map could be derived by summing up all the classified mask proposals. However, attributed to the high degree of similarity between the camouflaged object and the environment, adopting generic segmentation models or superpixel algorithms does not guarantee desirable mask proposals, as shown in Fig. 4.

Within this perspective, we propose C2F to retrieve camouflaged objects in a coarse-to-fine manner. Noting that the segmentation map $\mathbf{M}_\mathrm{c}$ produced through pixel-level retrieval accurately delineates object positioning, we modify the bounding box extraction algorithm [70] to derive multiple candidate bounding boxes from it. Specifically, for $\mathbf{M}_\mathrm{c}$, we extract all connected components and compute the minimal bounding box for each. Compared to uniformly sampled point prompts, these bounding boxes provide a more direct representation of potential camouflaged objects. Leveraging the explicit bounding boxes, SAM generates a collection of fine-grained mask proposals tailored for region-level retrieval.

IV. Experiments

A. Experimental Setup

1) Datasets and Evaluation Metrics: We evaluate our method on four commonly used COD datasets, including CHAMELEON [74], CAMO [75], COD10K [1], and NC4K [76], which contain 76, 250, 2,026, and 4,121 test images,

[1] https://github.com/facebookresearch/segment-anything/blob/main/notebooks/automatic_mask_generator_example.ipynb

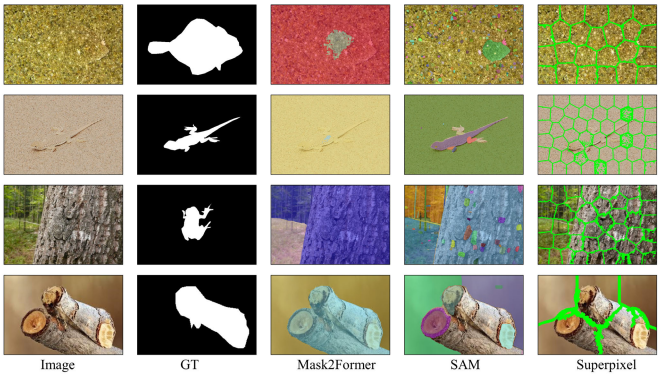


Fig. 4. Different methods for generating mask proposals. For Mask2Former [71], we adopt the checkpoint trained on ADE20K [72] for semantic segmentation. For SAM, we follow the official automatic mask generation pipeline. For superpixel algorithms, we adopt SLIC [73].

respectively. Being free of training, the training sets are discarded in our experiments. Following previous practice, we adopt structure measure (S_α) [77], mean E-measure (E_ϕ) [78], weighted F-measure (F_β^ω) [79] and mean absolute error (MAE) [80] as evaluation metrics.

2) Implementation Details: We choose LLaVA-1.5-7B [81] as the VLM to acquire categories. Specifically, we design three different prompts, "`What is the animal hiding in the environment`", "`What is around the hiding animal`", and "`What environment does the animal hide in`" for fetching categories of camouflaged objects, surrounding objects, and environment, respectively.

For the diffusion model, Stable Diffusion V1-5 [31] is leveraged to generate prototype images. We sample 64 and 32 images for each foreground and background category, respectively. Cross attention maps from Stable Diffusion with resolutions $\{8 \times 8, 16 \times 16\}$ at diffusion step 0 are upsampled and averaged to get the final cross attention map. DINOv2-ViT-L/14 [34] is adopted as the vision backbone to map images of size 518×518 into embedding vectors. As for SAM, we utilize HQ-SAM ViT-H [82] for high-quality segmentation. For efficient similarity computation and category matching, the coarse-to-fine retrieval is performed based on the FAISS library [83]. All experiments are conducted on a single A40.

B. Comparison With the Sota Methods

1) Comparison Methods: To validate the efficacy of RA-COD, we compare it with fully-supervised learning, weakly-supervised learning and training-free methods. For fully-supervised learning, we select 21 methods ranging from 2020 to 2024, including SINET [1], C2FNet [84], TINet [85], JSCOD [86], LSR [76], S-MGL [87], R-MGL [87], PFNET [88], UGTR [89], BGNet [90], PreyNet [91], FAP-Net [3], BSANet [92], FDNet [10], ZoomNet [42], SegMaR [41], SINETV2 [2], FSNet [7], HitNet [4], PopNet [9] and ICEG+ [15]. Weakly-supervised methods (WSCOD [18], WS-SAM [19]), and training-free SAM-based methods (MMCPF [23], GenSAM [25], ProMaC [93]) are also employed for comparisons.

2) Experimental Results and Analysis: Initially, we compare RA-COD with fully supervised methods. As shown in Table I, RA-COD achieves competitive performance against methods published up to 2022. For example, on COD10K, it surpasses SINET [1], PFNET [88], and BSANet [92]. Moreover, RA-COD delivers performance that is on par with or better than methods introduced after 2023. On the more rigorous COD10K dataset, the differences between RA-COD and the state-of-the-art FSNet [7] across S_α, E_ϕ, and M are minimal. We further extend the comparison to weakly supervised approaches. RA-COD shows a clear advantage over WSCOD [18] and WS-SAM [19] across datasets and metrics. Finally, we evaluate against recent training-free methods built on foundation models such as SAM and VLMs. RA-COD maintains a distinct lead over the latest training-free approaches, indicating that its gains derive from the design of generative prototypes and hierarchical retrieval rather than solely from the capabilities of pretrained foundation models.

In addition, a qualitative comparison is conducted with conventional supervised learning approaches, as depicted in Fig. 5. This comparison spans various camouflage scenarios encompassing large objects, small objects, multiple objects, occlusion, complex shapes, uncertain boundaries, and complex environments. Relative to alternative methodologies, RA-COD demonstrates superior capabilities in mitigating False-Negative and False-Positive detections, precise localization of camouflaged objects, and achieving refined segmentation. Notably, RA-COD exhibits exceptional performance in segmenting small objects, as evidenced in the final three rows, attributed to the coarse-to-fine retrieval scheme of C2F.

C. Ablation Study

We conduct comprehensive ablation experiments on all four benchmarks to validate the effectiveness of each key component of RA-COD.

1) Effectiveness of GenPro: The core of GenPro lies in the generation of high-quality foreground prototypes without irrelevant backgrounds, as well as the generation of background prototypes to complement the retrieval. GenPro achieves these two goals by mask average pooling and generating background category images, respectively.

The baseline for our analysis is established using the prototype repository comprising solely foreground samples without mask average pooling. During retrieval, the categories are determined by setting a similarity threshold (We search for the optimal parameter in the interval 0 to 1 with a step size of 0.1), as shown in row **a** of Table II. The comparison between rows **a** and **b** unmistakably reveals a significant enhancement across all metrics on the datasets. This improvement can be attributed to the utilization of mask average pooling, which effectively isolates foreground objects from the background, thereby producing embedding vectors that exhibit superior alignment with the target categories. In contrast, relying solely on global average pooling would result in the dilution of foreground category semantics by irrelevant objects, thereby generating prototype samples that are indistinguishable.

Although mask average pooling enhances sample representativeness, class discrimination via similarity thresholds

TABLE I

QUANTITATIVE COMPARISON ON FOUR COD BENCHMARKS. F: FULLY-SUPERVISED. W: WEAKLY-SUPERVISED. TF: TRAINING-FREE. ↑/↓: THE HIGHER/LOWER THE BETTER. "-": NOT AVAILABLE. "*": RESULTS DIRECTLY FROM THE PAPER. THE BEST RESULTS ARE IN BOLD

Method	Pub./Year	Sup.	CHAMELEON				CAMO				COD10K				NC4K			
			$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$
SINET	CVPR20	F	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051	0.808	0.871	0.723	0.058
C2FNet	IJCAI21	F	0.888	0.935	0.828	0.032	0.796	0.854	0.719	0.080	0.813	0.890	0.686	0.036	0.838	0.897	0.762	0.049
TINet	AAAI21	F	0.874	0.916	0.783	0.038	0.781	0.847	0.678	0.087	0.793	0.848	0.635	0.043	0.829	0.879	0.734	0.055
JSCOD	CVPR21	F	0.891	0.945	0.833	0.030	0.800	0.859	0.728	0.073	0.809	0.884	0.684	0.035	0.842	0.898	0.771	0.047
LSR	CVPR21	F	0.890	0.935	0.822	0.030	0.787	0.838	0.696	0.080	0.804	0.880	0.673	0.037	0.840	0.895	0.766	0.048
S-MGL	CVPR21	F	0.892	0.912	0.802	0.032	0.772	0.806	0.664	0.089	0.811	0.844	0.654	0.037	0.829	0.862	0.731	0.055
R-MGL	CVPR21	F	0.893	0.917	0.812	0.031	0.776	0.812	0.673	0.088	0.814	0.851	0.666	0.035	0.833	0.867	0.739	0.053
PFNET	CVPR21	F	0.882	0.931	0.810	0.033	0.782	0.841	0.695	0.085	0.800	0.877	0.660	0.040	0.829	0.887	0.745	0.053
UGTR	ICCV21	F	0.888	0.911	0.796	0.031	0.785	0.823	0.686	0.086	0.818	0.853	0.667	0.035	0.839	0.874	0.747	0.052
BGNet	IJCAI22	F	0.901	0.943	0.850	0.027	0.812	0.870	0.749	0.073	0.831	0.901	0.722	0.033	0.851	0.907	0.788	0.044
PreyNet	MM22	F	0.895	0.952	0.844	0.028	0.790	0.842	0.708	0.077	0.813	0.881	0.697	0.034	0.834	0.887	0.763	0.050
FAPNet	TIP22	F	0.893	0.940	0.825	0.028	0.815	0.865	0.734	0.076	0.822	0.888	0.694	0.036	0.851	0.899	0.775	0.047
BSANet	AAAI22	F	0.895	0.946	0.841	0.027	0.794	0.851	0.717	0.079	0.818	0.891	0.699	0.034	0.841	0.897	0.771	0.048
FDNet	CVPR22	F	0.895	0.951	0.849	0.027	0.840	0.896	0.782	0.063	0.838	0.921	0.747	0.030	0.832	0.895	0.759	0.052
ZoomNet	CVPR22	F	0.902	0.943	0.845	0.023	0.820	0.877	0.752	0.066	0.838	0.888	0.729	0.029	0.853	0.896	0.784	0.043
SegMaR	CVPR22	F	0.906	0.951	0.860	0.025	0.815	0.874	0.753	0.071	0.833	0.899	0.724	0.034	0.841	0.896	0.781	0.046
SINETV2	TPAMI22	F	0.888	0.942	0.816	0.030	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.770	0.048
FSNet	TIP23	F	0.905	0.961	0.854	0.023	**0.879**	**0.931**	**0.838**	**0.042**	0.870	**0.936**	0.787	**0.023**	**0.891**	**0.938**	**0.845**	**0.031**
HitNet	AAAI23	F	**0.921**	**0.967**	**0.897**	**0.019**	0.849	0.906	0.809	0.055	**0.871**	0.935	**0.806**	**0.023**	0.875	0.926	0.834	0.037
PopNet	ICCV23	F	0.917	0.965	0.875	0.020	0.808	0.859	0.744	0.077	0.851	0.910	0.757	0.028	0.861	0.909	0.802	0.042
ICEG+*	ICLR24	F	0.908	0.961	-	0.022	0.871	**0.931**	**-**	**0.042**	0.862	0.934	-	**0.023**	0.883	0.937	-	0.033
WS-SAM	NIPS23	W	0.820	0.887	0.723	0.048	0.759	0.814	0.667	0.092	0.803	0.877	0.680	0.038	0.829	0.886	0.757	0.052
WSCOD	AAAI23	W	0.794	0.870	0.708	0.048	0.731	0.814	0.637	0.095	0.719	0.814	0.552	0.052	0.765	0.847	0.673	0.066
MMCPF	MM24	TF	-	-	-	-	0.749	0.820	0.680	0.101	0.733	0.803	0.592	0.066	0.767	0.826	0.681	0.083
GenSAM	AAAI24	TF	0.767	0.827	0.673	0.075	0.738	0.803	0.674	0.106	0.773	0.832	0.667	0.065	0.810	0.866	0.751	0.065
ProMaC*	NeurIPS24	TF	0.833	0.899	-	0.044	0.767	0.846	-	0.090	0.805	0.876	-	0.042	-	-	-	-
Ours	-	TF	**0.868**	**0.914**	**0.826**	**0.032**	**0.823**	**0.884**	**0.781**	**0.063**	**0.863**	**0.926**	**0.798**	**0.024**	**0.872**	**0.924**	**0.833**	**0.034**

TABLE II

ABLATION EXPERIMENTS ON EFFECTIVENESS OF GENPRO AND C2F

index	GenPro		C2F		CHAMELEON				CAMO				COD10K				NC4K			
	w/ mask average pooling	w/ background category	w/ pixel-level retrieval	w/ region-level retrieval	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi\uparrow$	$F_\beta^\omega\uparrow$	$M\downarrow$
a	✘	✘	✔	✔	0.647	0.691	0.518	0.185	0.609	0.637	0.508	0.240	0.586	0.606	0.378	0.192	0.635	0.662	0.493	0.203
b	✔	✘	✔	✔	0.854	0.896	0.812	0.041	0.716	0.734	0.614	0.100	0.857	0.921	0.790	0.026	0.862	0.916	0.822	0.038
c	✘	✔	✔	✔	0.501	0.578	0.324	0.257	0.492	0.558	0.324	0.273	0.535	0.589	0.292	0.189	0.549	0.605	0.371	0.224
d	✔	✔	✔	✘	0.826	0.868	0.689	0.051	0.783	0.830	0.662	0.095	0.766	0.817	0.579	0.050	0.813	0.863	0.690	0.061
e	✔	✔	✘	✔	0.789	0.845	0.694	0.048	0.738	0.807	0.657	0.096	0.802	0.883	0.694	0.033	0.823	0.883	0.759	0.048
f	✔	✔	✔	✔	0.868	0.914	0.826	0.032	0.823	0.884	0.781	0.063	0.863	0.926	0.798	0.024	0.872	0.924	0.833	0.034

lacks intuitiveness and necessitates hyperparameter tuning. The introduction of background category samples transforms the previous "hard" discrimination approach, reliant on thresholding, into a "soft" method centered on maximum similarity, thereby eliminating the need for hyperparameters and rendering more interpretable predictions, as evidenced by rows **b** and **f**. Moreover, the mere introduction of background categories without incorporating mask average pooling yields unsatisfactory results, as demonstrated in row **c**. This is due to the inclusion of irrelevant environment categories in the foreground samples.

2) Effectiveness of C2F: Taking into account the intrinsic correlation between camouflaged objects and their surroundings, C2F endeavors to progressively mitigate camouflage effects by synergistically leveraging pixel-level and region-level retrieval strategies.

Pixel-level retrieval facilitates the explicit categorization of each point within the feature space, thus proving valuable in discerning highly similar foregrounds from backgrounds, as demonstrated in row **d** of Table II. Moreover, fine-grained retrieval contributes to the effective localization of camouflaged objects, particularly those of small size. However, implemented within a low-resolution feature space, pixel-level retrieval generates segmentation maps characterized by prominently jagged edges and substantial areas of noise.

A pragmatic strategy involves assigning a class to a particular region rather than to individual pixels. Utilizing the pre-trained segmentation model SAM, fine masks are produced, which are subsequently refined through region-level retrieval to generate the ultimate segmentation outcome, as illustrated in row **e**.

However, SAM encounters challenges in generating mask proposals that maintain semantic consistency without explicit guidance. Considering this, C2F derives explicit prompts from the outcomes of pixel-level retrieval to guide SAM in generating fine-grained mask proposals. As depicted in rows **d**, **e**, and **f**, C2F significantly enhances the performance of RA-COD in detecting and segmenting camouflaged objects. In comparison to the SAM-based proposal generation approach, C2F achieves an average 10.9% improvement in F_β^ω metric across all datasets.

3) Versions of Stable Diffusion: We employ a range of diffusion models, including Stable Diffusion V1-1, V1-2, V1-3, V1-4, and V1-5, to investigate how the quality of generated images influences model performance. For simplicity, we adopt Score $= S_\alpha + E_\phi + F_\beta^\omega + 1 - M$ to denote the overall performance. As illustrated in Fig. 6, improvements in the generative quality of these models (with higher version numbers indicating better image quality) are accompanied by corresponding enhancements in overall model performance. These results demonstrate the scalability of our proposed framework:

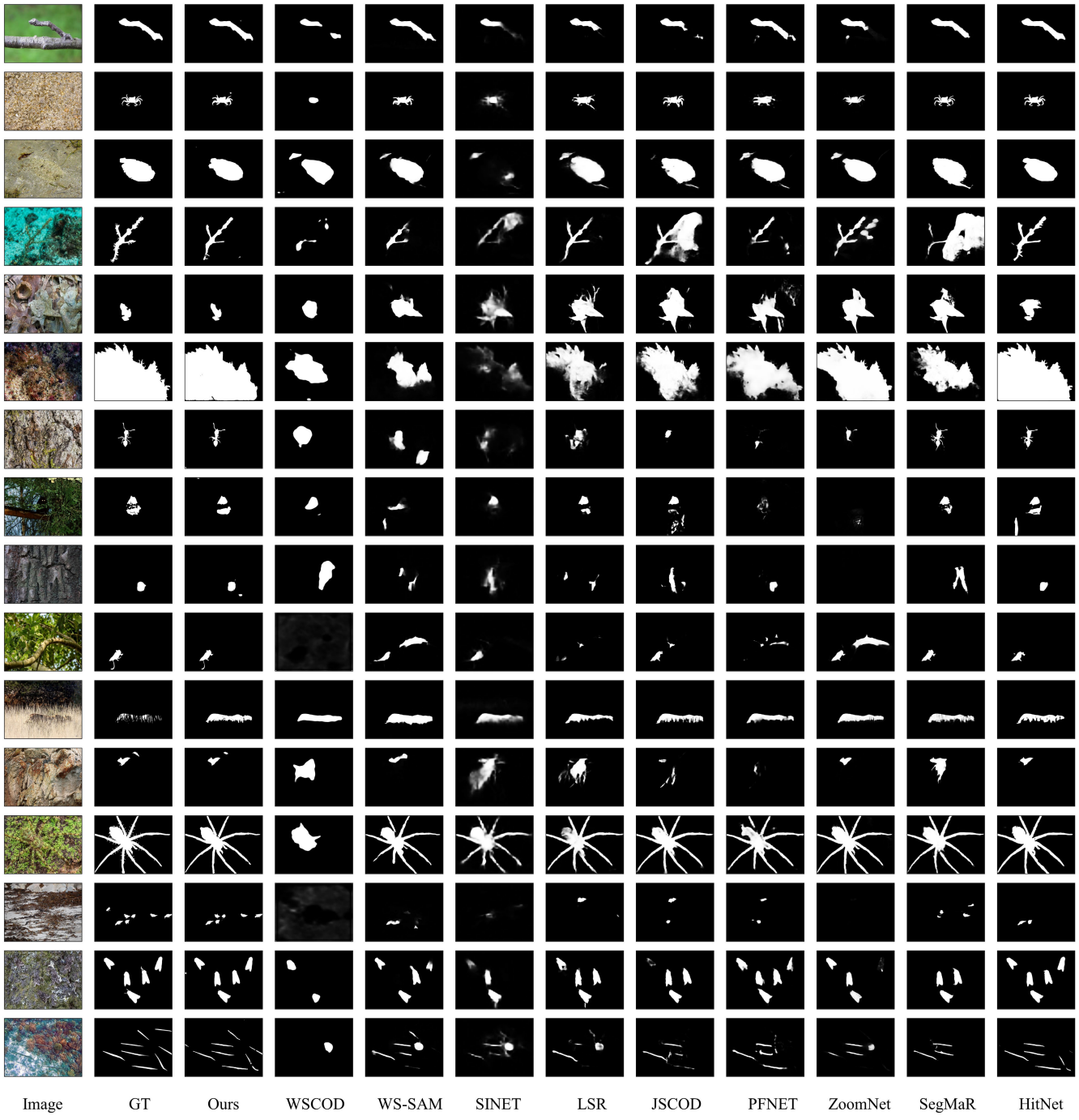


Fig. 5. Visual comparisons with weakly-supervised (WSCOD [18], WS-SAM [19]) and fully-supervised methods (SINET [1], LSR [76], JSCOD [86], PFNET [88], ZoomNet [42], SegMaR [41], HitNet [4]).

its retrieval-based paradigm enables effective integration of diverse independent models, allowing the framework's performance to advance in tandem with improvements in its individual components.

4) Versions of DINOv2 and SAM: In our RA-COD framework, DINOv2 is tasked with projecting RGB images into the feature space and generating embedding vectors, while SAM is entrusted with generating fine mask proposals.

We first explore various backbones of DINOv2, encompassing ViT-S/B/L/G, corresponding to feature vectors of 384/768/1024/1536 dimensions, respectively. Higher-dimensional embedding vectors can encapsulate more comprehensive information, particularly beneficial for detail-dependent COD tasks. As demonstrated in Table III, ViT-L outperforms ViT-S and ViT-B on the more representative COD10K dataset. However, further integration of ViT-G does not yield a noteworthy enhancement in performance, potentially attributable to feature redundancy.

Furthermore, we evaluate various SAM configurations, including vanilla SAM-B/L/H and the SAM variant

TABLE III
ABLATION EXPERIMENTS ON DIFFERENT VERSIONS OF DINOV2 AND SAM

Model	Backbone	CHAMELEON				CAMO				COD10K				NC4K			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
DINOv2	ViT-S/14	0.830	0.867	0.759	0.041	0.784	0.832	0.715	0.078	0.844	0.890	0.762	0.026	0.858	0.902	0.811	0.037
	ViT-B/14	0.850	0.887	0.793	0.038	0.813	0.871	0.767	0.069	0.858	0.917	0.789	0.024	0.866	0.914	0.824	0.036
	ViT-L/14	**0.868**	**0.914**	**0.826**	**0.032**	**0.823**	**0.884**	**0.781**	**0.063**	**0.863**	**0.926**	**0.798**	0.024	**0.872**	**0.924**	0.833	**0.034**
	ViT-G/14	0.862	0.912	0.816	0.035	0.801	0.854	0.743	0.073	**0.863**	0.924	**0.798**	**0.023**	**0.872**	**0.924**	**0.834**	**0.034**
SAM	ViT-B	0.823	0.874	0.754	0.044	0.728	0.786	0.643	0.099	0.817	0.893	0.730	0.033	0.815	0.875	0.754	0.054
	ViT-L	0.846	0.902	0.799	0.039	0.773	0.838	0.715	0.081	0.841	0.910	0.765	0.028	0.844	0.901	0.796	0.043
	ViT-H	0.867	**0.920**	0.822	**0.031**	0.791	0.859	0.739	0.073	0.848	0.917	0.775	0.026	0.852	0.908	0.805	0.041
	ViT-B-HQ	0.794	0.821	0.712	0.048	0.770	0.826	0.706	0.083	0.830	0.895	0.751	0.031	0.833	0.885	0.780	0.048
	ViT-L-HQ	0.843	0.895	0.790	0.040	0.803	0.866	0.754	0.071	0.852	0.919	0.780	0.026	0.859	0.917	0.817	0.038
	ViT-H-HQ	**0.868**	0.914	**0.826**	0.032	**0.823**	**0.884**	**0.781**	**0.063**	**0.863**	**0.926**	**0.798**	**0.024**	**0.872**	**0.924**	**0.833**	**0.034**

TABLE IV
ABLATION EXPERIMENTS ON DIFFERENT MASK PROPOSAL GENERATORS

Mask Proposal Generator	CHAMELEON				CAMO				COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
Point-wise	0.826	0.868	0.689	0.051	0.783	0.830	0.662	0.095	0.766	0.817	0.579	0.050	0.813	0.863	0.690	0.061
Mask2Former	0.607	0.545	0.368	0.096	0.645	0.626	0.469	0.112	0.643	0.591	0.399	0.055	0.646	0.610	0.456	0.087
SuperPixel	0.724	0.826	0.589	0.074	0.631	0.711	0.468	0.134	0.692	0.800	0.508	0.058	0.700	0.816	0.567	0.085
SAM	0.789	0.845	0.694	0.048	0.738	0.807	0.657	0.096	0.802	0.883	0.694	0.033	0.823	0.883	0.759	0.048
C2F	**0.868**	**0.914**	**0.826**	**0.032**	**0.823**	**0.884**	**0.781**	**0.063**	**0.863**	**0.926**	**0.798**	**0.024**	**0.872**	**0.924**	**0.833**	**0.034**

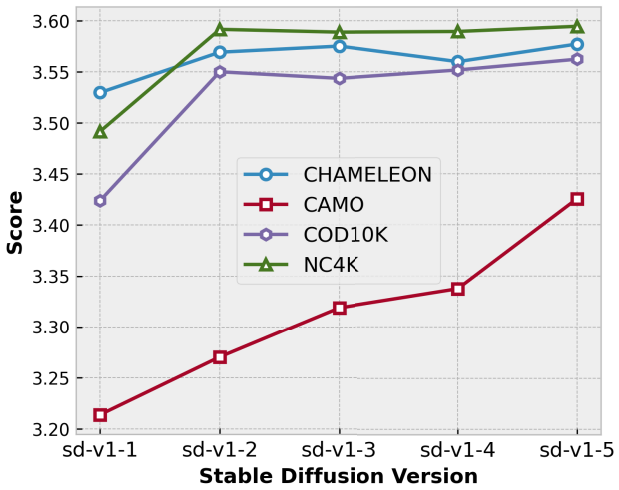


Fig. 6. Ablation experiments on different versions of diffusion models.

HQ-SAM-B/L/H. The results presented in Table III reveal that (1) larger ViT backbones yield improved segmentation performance, and (2) HQ-SAM consistently outperforms the original SAM in terms of segmentation quality. Additionally, our experiments indicate that RA-COD does not exhibit a trend of performance saturation, suggesting its adaptability and ongoing competitiveness with advancements in foundation models.

5) Mask Proposal Generators: We conduct a comparative analysis of C2F against various mask proposal generators, including point-wise (treating pixels as proposals), superpixel [73], Mask2Former [71], and SAM, as illustrated in Table IV. In contrast to conventional proposal generation approaches, C2F initially isolates camouflaged objects from the environment through pixel-level retrieval. From this, it derives high-confidence bounding boxes to guide SAM in generating semantically consistent proposals. By adopting the retrieval-based mask generation scheme, C2F effectively mitigates issues such as misalignment between the capabilities of pre-trained segmentation models (e.g., Mask2Former) and the requirements of the COD task. Furthermore, by leveraging a proficient SAM, C2F addresses the limitations of superpixel segmentation in accurately delineating the boundaries of camouflaged objects. Our experimental findings demonstrate that C2F significantly outperforms the aforementioned methods and offers a novel approach to mask proposal generation.

6) Number of Prototype Samples: We investigate the influence of the quantity of prototype samples per foreground category on RA-COD. As depicted in Fig. 7, RA-COD demonstrates resilience to sample quantity across CHAMELEON, COD10K, and NC4K datasets, where promising outcomes are achieved with merely two samples. Conversely, performance on CAMO displays a direct correlation with sample quantity. Upon further examination, we observe that the CAMO test set comprises a substantial proportion of human camouflaged images (117 out of 250), presenting more intricate scenarios compared to animal-centric camouflage images, as shown in Fig. 8. To effectively discern humans concealed within intricate environments, the prototype repository necessitates a diverse and extensive sample pool. To validate this assertion, we eliminate test images featuring humans from CAMO, resulting in the CAMO_WO_HUMAN dataset. As illustrated in Fig. 7, CAMO_WO_HUMAN similarly demonstrates robustness to sample quantity.

7) Number of Categories: To further demonstrate the robustness of RA-COD, we assess how performance varies with the number of object and environment categories. In our default setting, there are 79 object categories and 49 environment categories. For each setting, we randomly sample

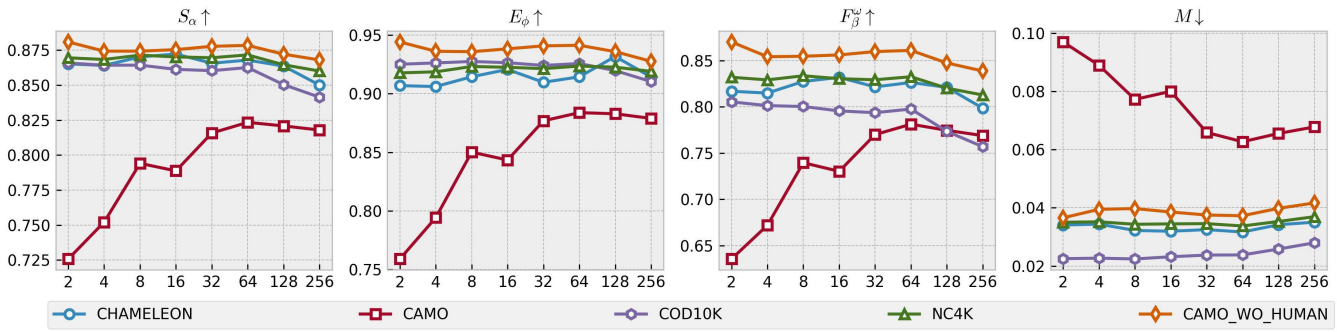


Fig. 7. Effect of prototype sample quantity. CAMO comprises a large number of human camouflage images spanning diverse scenes, necessitating a larger sample size to attain satisfactory segmentation outcomes.

TABLE V
EXPERIMENTS ON THE EFFECT OF THE NUMBER OF ENVIRONMENT/OBJECT CATEGORIES

Num	CHAMELEON				CAMO				COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
Environment Category																
9	0.868	0.927	0.827	0.032	0.821	0.882	0.775	0.066	0.858	0.920	0.788	0.025	0.870	0.923	0.830	0.035
19	0.868	0.917	0.827	0.032	0.824	0.884	0.780	0.065	0.859	0.920	0.791	0.025	0.870	0.922	0.830	0.035
29	0.869	0.918	0.829	0.031	0.821	0.880	0.776	0.066	0.861	0.921	0.793	0.025	0.870	0.922	0.831	0.035
39	0.869	0.916	0.829	0.031	0.820	0.878	0.776	0.066	0.862	0.922	0.795	0.024	0.871	0.922	0.831	0.035
49	0.868	0.914	0.826	0.032	0.823	0.884	0.781	0.063	0.863	0.926	0.798	0.024	0.872	0.924	0.833	0.034
Object Category																
19	0.863	0.905	0.820	0.034	0.758	0.796	0.680	0.086	0.865	0.920	0.804	0.023	0.868	0.914	0.829	0.035
29	0.864	0.906	0.822	0.033	0.764	0.806	0.691	0.085	0.864	0.921	0.802	0.023	0.869	0.917	0.830	0.035
39	0.866	0.909	0.823	0.032	0.781	0.827	0.715	0.080	0.865	0.924	0.803	0.023	0.871	0.920	0.833	0.034
49	0.868	0.913	0.827	0.032	0.802	0.854	0.748	0.072	0.864	0.924	0.801	0.023	0.871	0.921	0.833	0.034
59	0.868	0.914	0.828	0.032	0.818	0.875	0.773	0.067	0.863	0.923	0.799	0.024	0.872	0.922	0.833	0.034
69	0.870	0.916	0.830	0.031	0.819	0.876	0.775	0.067	0.863	0.923	0.798	0.024	0.871	0.922	0.833	0.034
79	0.868	0.914	0.826	0.032	0.823	0.884	0.781	0.063	0.863	0.926	0.798	0.024	0.872	0.924	0.833	0.034

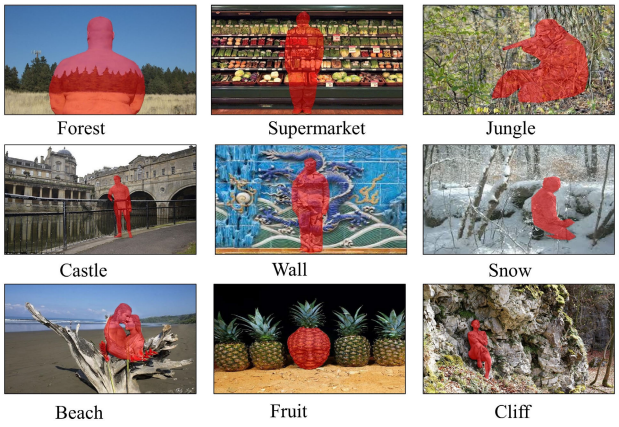


Fig. 8. Images of the "human" category in CAMO contain a range of complex scenarios.

the specified number of categories from the full pools, run the experiment five times with different random seeds, and report the mean performance.

As shown in Table V, across all three benchmarks (CHAMELEON, COD10K, and NC4K), RA-COD's performance is largely unaffected even when the number of environment or object categories is substantially reduced. We attribute this to two factors: (1) even if the prototype repository lacks the exact test categories, object and environment features in test images still align more closely with those in the repository; and (2) the Coarse-to-Fine (C2F) retrieval strategy enhances robustness—while missing exact categories may slightly affect the coarse stage, it only needs to localize objects (via bounding boxes) rather than produce precise masks. Consequently, reliable localization is attainable despite moderate noise in coarse retrieval.

However, on the CAMO dataset, RA-COD's performance declines as the number of object categories decreases. We attribute this to the reduced likelihood of including human-related prototypes in the repository when fewer categories are available. Unlike natural camouflage, human camouflage generally requires richer, more diverse prototypes (as discussed earlier). Given that CAMO contains many human-related camouflage images, this scarcity of relevant prototypes leads to a noticeable performance drop.

D. Limitations

As depicted in Fig. 9, RA-COD encounters challenges akin to those faced by state-of-the-art fully supervised learning methods, manifesting as False-Positive (FP) and False-Negative (FN) detections, and failure in some

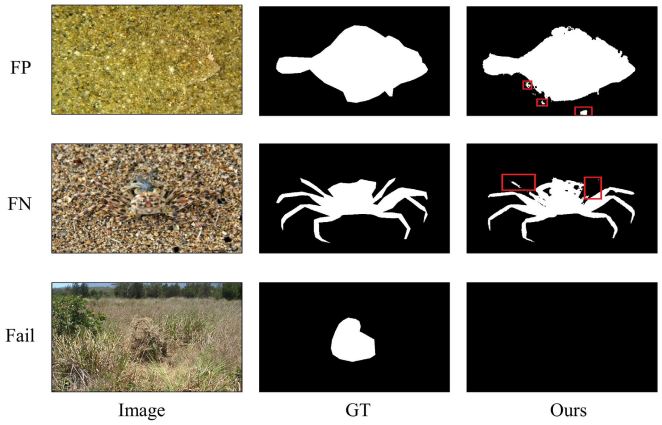


Fig. 9. Limitations of RA-COD.

challenging situations. We posit that these hurdles can be alleviated through generating more granular sample prototypes and improving the mask proposal generation scheme. In addition, RA-COD leverages a suite of pre-trained models, encompassing VLMs, Stable Diffusion, SAM, DINOv2, among others. While obviating the need for training, this engenders an additional parameter load. We propose that knowledge distillation [94] could expedite inference and alleviate the parameter burden. We leave the aforementioned issues for future exploration.

V. Conclusion

In this paper, we introduce RA-COD, a paradigm rooted in retrieval augmentation for COD. RA-COD endeavors to harness the robust capabilities of off-the-shelf foundation models to address COD challenges without task-specific training. It comprises two pivotal components, GenPro and C2F, where GenPro crafts a comprehensive and distinctive prototype repository, while C2F facilitates the seamless transition from retrieval to segmentation. Extensive experiments demonstrate the advantages of RA-COD over weakly-supervised and training-free methods.

References

[1] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 2774–2784.

[2] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 10, pp. 6024–6042, Oct. 2022.

[3] T. Zhou, Y. Zhou, C. Gong, J. Yang, and Y. Zhang, "Feature aggregation and propagation network for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 7036–7047, 2022.

[4] X. Hu et al., "High-resolution iterative feedback network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell.*, 2022, pp. 881–889.

[5] B. Yin et al., "CamoFormer: Masked separable attention for camouflaged object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 46, no. 12, pp. 10362–10374, Dec. 2024.

[6] Z. Huang et al., "Feature shrinkage pyramid for camouflaged object detection with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 5557–5566.

[7] Z. Song, X. Kang, X. Wei, H. Liu, R. Dian, and S. Li, "FSNet: Focus scanning network for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 2267–2278, 2023.

[8] P. Li, X. Yan, H. Zhu, M. Wei, X.-P. Zhang, and J. Qin, "FindNet: Can you find me? Boundary-and-texture enhancement network for camouflaged object detection," *IEEE Trans. Image Process.*, vol. 31, pp. 6396–6411, 2022.

[9] Z. Wu et al., "Source-free depth for object pop-out," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1032–1042.

[10] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4494–4503.

[11] C. He et al., "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 22046–22055.

[12] J. Wu, W. Liang, F. Hao, and J. Xu, "Mask-and-edge co-guided separable network for camouflaged object detection," *IEEE Signal Process. Lett.*, vol. 30, pp. 748–752, 2023.

[13] X. Zhang, B. Yin, Z. Lin, Q. Hou, D.-P. Fan, and M.-M. Cheng, "Referring camouflaged object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 5, pp. 3597–3610, May 2025.

[14] S. Cheng, G.-P. Ji, P. Qin, D.-P. Fan, B. Zhou, and P. Xu, "Large model based referring camouflaged object detection," 2023, *arXiv:2311.17122*.

[15] C. He et al., "Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects," in *Proc. Int. Conf. Learn. Represent.*, 2023, pp. 1–10.

[16] H. Lamdouar, W. Xie, and A. Zisserman, "The making and breaking of camouflage," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 832–842.

[17] Z. Chen, R. Gao, T.-Z. Xiang, and F. Lin, "Diffusion model for camouflaged object detection," in *Proc. Eur. Conf. Artif. Intell.*, 2023, pp. 445–452.

[18] R. He, Q. Dong, J. Lin, and R. W. H. Lau, "Weakly-supervised camouflaged object detection with scribble annotations," in *Proc. AAAI Conf. Artif. Intell.*, vol. 37, 2023, pp. 781–789.

[19] C. He et al., "Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 30726–30737.

[20] A. Kirillov et al., "Segment anything," in *Proc. Int. Conf. Comput. Vision (ICCV)*, 2023, pp. 3992–4003.

[21] G.-P. Ji, D.-P. Fan, P. Xu, B. Zhou, M.-M. Cheng, and L. Van Gool, "SAM struggles in concealed scenes—Empirical study on 'segment anything,'" *Sci. China Inf. Sci.*, vol. 66, no. 12, pp. 226101:1–226101:4, Dec. 2023.

[22] L. Tang, H. Xiao, and B. Li, "Can SAM segment anything? When SAM meets camouflaged object detection," 2023, *arXiv:2304.04709*.

[23] L. Tang, P.-T. Jiang, Z.-H. Shen, H. Zhang, J.-W. Chen, and B. Li, "Chain of visual perception: Harnessing multimodal large language models for zero-shot camouflaged object detection," in *Proc. 32nd ACM Int. Conf. Multimedia*, Oct. 2024, pp. 8805–8814.

[24] Y. Jiang et al., "Effectiveness assessment of recent large vision-language models," *Vis. Intell.*, vol. 2, no. 1, pp. 1–17, 2024.

[25] J. Hu, J. Lin, S. Gong, and W. Cai, "Relax image-specific prompt requirement in SAM: A single generic prompt for segmenting camouflaged objects," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 12511–12518.

[26] A. Radford et al., "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Mach. Learn.*, vol. 139, 2021, pp. 8748–8763.

[27] F. Liu, K. Lin, L. Li, J. Wang, Y. Yacoob, and L. Wang, "Aligning large multi-modal model with robust instruction tuning," 2023, *arXiv:2306.14565*.

[28] Y. J. Lee, C. Li, H. Liu, and Q. Wu, "Visual instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 34892–34916.

[29] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Proc. 34th Int. Conf. Neural Inf. Process. Syst.*, H. Larochelle, M. A. Ranzato, R. Hadsell, M.-F. Balcan, and H.-T. Lin, Eds., Red Hook, NY, USA, Dec. 2020, pp. 6840–6851.

[30] J. Song, C. Meng, and S. Ermon, "Denoising diffusion implicit models," in *Proc. Int. Conf. Learn. Represent.*, 2021, pp. 1–12.

[31] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2022, pp. 10684–10695.

[32] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with CLIP latents," 2022, *arXiv:2204.06125*.

[33] W. Chan et al., "Photorealistic text-to-image diffusion models with deep language understanding," in *Proc. Adv. Neural Inf. Process. Syst. 35*, 2022, pp. 36479–36494.

[34] M. Oquab et al., "DINOv2: Learning robust visual features without supervision," *Trans. Mach. Learn. Res.*, vol. 2024, pp. 1–32, Apr. 2023.

[35] F. Xiao et al., "A survey of camouflaged object detection and beyond," *CAAI Artif. Intell. Res.*, vol. 3, pp. 1–26, Dec. 2024.

[36] H. Li, C.-M. Feng, Y. Xu, T. Zhou, L. Yao, and X. Chang, "Zero-shot camouflaged object detection," *IEEE Trans. Image Process.*, vol. 32, pp. 5126–5137, 2023.

[37] C. He et al., "RUN: Reversible unfolding network for concealed object segmentation," in *Proc. Int. Conf. Mach. Learn.*, 2025, pp. 1–14.

[38] C. He et al., "Segment concealed objects with incomplete supervision," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 9, pp. 7832–7851, Sep. 2025.

[39] C. Hao, Z. Yu, X. Liu, J. Xu, H. Yue, and Y. Jing-yu, "A simple yet effective network based on vision transformer for camouflaged object and salient object detection," *IEEE Trans. Image Process.*, vol. 34, pp. 608–622, 2024.

[40] W. Liang, J. Wu, Y. Wu, X. Mu, and J. Xu, "FINet: Frequency injection network for lightweight camouflaged object detection," *IEEE Signal Process. Lett.*, vol. 31, pp. 526–530, 2024.

[41] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 4703–4712.

[42] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 2150–2160.

[43] J. Ma, Y. He, F. Li, L. Han, C. You, and B. Wang, "Segment anything in medical images," *Nature Commun.*, vol. 15, no. 1, p. 654, 2024.

[44] Y. Shen, K. Song, X. Tan, D. Li, W. Lü, and Y. Zhuang, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in hugging face," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 38154–38180.

[45] D. Baranchuk, I. Rubachev, A. Voynov, V. Khrulkov, and A. Babenko, "Label-efficient semantic segmentation with diffusion models," in *Proc. Int. Conf. Learn. Represent.*, 2022, pp. 1–15.

[46] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-prompt image editing with cross-attention control," in *Proc. 11th Int. Conf. Learn. Represent. (ICLR)*, Kigali, Rwanda, May 2023, pp. 1–36.

[47] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, Oct. 2023, pp. 3836–3847.

[48] C. He et al., "Diffusion models in low-level vision: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 47, no. 6, pp. 4630–4651, Jun. 2025.

[49] J. Xu, S. Liu, A. Vahdat, W. Byeon, X. Wang, and S. De Mello, "Open-vocabulary panoptic segmentation with text-to-image diffusion models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 2955–2966.

[50] L. Barsellotti, R. Amoroso, M. Cornia, L. Baraldi, and R. Cucchiara, "Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3689–3698.

[51] Y. Zhao, Q. Ye, W. Wu, C. Shen, and F. Wan, "Generative prompt model for weakly supervised object localization," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 6328–6338.

[52] S. Chen, P. Sun, Y. Song, and P. Luo, "DiffusionDet: Diffusion model for object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 19773–19786.

[53] W. Wu, Y. Zhao, M. Z. Shou, H. Zhou, and C. Shen, "DiffuMask: Synthesizing images with pixel-level annotations for semantic segmentation using diffusion models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 1206–1217.

[54] W. Wu et al., "DatasetDM: Synthesizing data with perception annotations using diffusion models," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 54683–54695.

[55] Z. Chen, K. Sun, and X. Lin, "CamoDiffusion: Camouflaged object detection via conditional diffusion models," in *Proc. AAAI Conf. Artif. Intell.*, vol. 38, 2024, pp. 1272–1280.

[56] X.-J. Luo et al., "CamDiff: Camouflage image augmentation via diffusion model," in *Proc. CAAI Artif. Intell. Res.*, 2023, pp. 1–10.

[57] P. Zhao et al., "LAKE-RED: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 4092–4101.

[58] K. Guu, K. Lee, K. Z. Tung, P. Pasupat, and M. Chang, "Retrieval augmented language model pre-training," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 3929–3938.

[59] P. Lewis et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 33, 2020, pp. 9459–9474.

[60] S. Borgeaud et al., "Improving language models by retrieving from trillions of tokens," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 2206–2240.

[61] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, 2021, pp. 1–12.

[62] V. Udandarao, A. Gupta, and S. Albanie, "SuS-X: Training-free name-only transfer of vision-language models," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2023, pp. 2725–2736.

[63] G. Shin, W. Xie, and S. Albanie, "ReCo: Retrieve and co-segment for zero-shot transfer," in *Proc. Adv. Neural Inf. Process. Syst.*, 2022, pp. 33754–33767.

[64] L. Karazija, I. Laina, A. Vedaldi, and C. Rupprecht, "Diffusion models for open-vocabulary segmentation," in *Proc. Eur. Conf. Comput. Vis.*, 2023, pp. 299–317.

[65] Y. Wang, R. Sun, N. Luo, Y. Pan, and T. Zhang, "Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 3952–3963.

[66] B. Peng, Z. Tian, S. Liu, M.-C. Yang, and J. Jia, "Scalable language model with generalized continual learning," in *Proc. Int. Conf. Learn. Represent.*, 2024, pp. 1–23.

[67] Z. Gui et al., "KNN-CLIP: Retrieval enables training-free segmentation on continually expanding large vocabularies," *Trans. Mach. Learn. Res.*, vol. 2024, pp. 1–17, Aug. 2024.

[68] J. Li, D. Li, S. Savarese, and S. C. H. Hoi, "BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 19730–19742.

[69] W. Dai et al., "InstructBLIP: Towards general-purpose vision-language models with instruction tuning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 49250–49267.

[70] X. Zhang, Y. Wei, J. Feng, Y. Yang, and T. Huang, "Adversarial complementary learning for weakly supervised object localization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1325–1334.

[71] B. Cheng, I. Misra, A. G. Schwing, A. Kirillov, and R. Girdhar, "Masked-attention mask transformer for universal image segmentation," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 1280–1289.

[72] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ADE20K dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5122–5130.

[73] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Süsstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.

[74] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," *Unpublished manuscript*, vol. 2, no. 6, p. 7, 2018.

[75] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Comput. Vis. Image Understand.*, vol. 184, pp. 45–56, Jul. 2019.

[76] Y. Lv et al., "Simultaneously localize, segment and rank the camouflaged objects," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 11586–11596.

[77] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4558–4567.

[78] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 698–704.

[79] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 248–255.

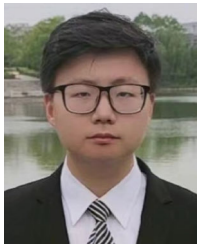
[80] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[81] H. Liu, C. Li, Y. Li, and Y. J. Lee, "Improved baselines with visual instruction tuning," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2024, pp. 26296–26306.

[82] K. Lei et al., "Segment anything in high quality," in *Proc. Adv. Neural Inf. Process. Syst.*, 2023, pp. 29914–29934.
[83] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," *IEEE Trans. Big Data*, vol. 7, no. 3, pp. 535–547, Jul. 2021.
[84] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," in *Proc. 30th Int. Joint Conf. Artif. Intell.*, Aug. 2021, pp. 1025–1031.
[85] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 4, 2021, pp. 3599–3607.
[86] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 10066–10076.
[87] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Nashville, TN, USA, Jun. 2021, pp. 12992–13002.
[88] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 8768–8777.
[89] F. Yang et al., "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Montreal, QC, Canada, Oct. 2021, pp. 4126–4135.
[90] Y. Sun, S. Wang, C. Chen, and T. Xiang, "Boundary-guided camouflaged object detection," in *Proc. 21st Int. Joint Conf. Artif. Intell.*, 2022, pp. 1335–1341.
[91] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "PreyNet: Preying on camouflaged objects," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 5323–5332.
[92] H. Zhu et al., "I can find you! Boundary-guided separated attention network for camouflaged object detection," in *Proc. AAAI Conf. Artif. Intell. (AAAI)*, 2022, pp. 3608–3616.
[93] S. Gong, J. Hu, J. Lin, and J. Yan, "Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2024, pp. 107171–107197.
[94] G. E. Hinton, O. Vinyals, and J. B. Dean, "Distilling the knowledge in a neural network," in *Proc. Adv. Neural Inf. Process. Syst. Workshops*, 2015, pp. 1–9.

Ji Du received the bachelor's degree from Nankai University, Tianjin, China, in 2023, where he is currently pursuing the Ph.D. degree with the College of Artificial Intelligence. His research interests include machine learning and computer vision.

Jiesheng Wu is currently pursuing the Ph.D. degree with the College of Artificial Intelligence, Nankai University, Tianjin, China. His main research focuses on image processing based on deep learning.



Desheng Kong received the master's degree from Nankai University, where he is currently pursuing the Ph.D. degree with the College of Artificial Intelligence, supervised by Prof. Jing Xu. His research interests include graph neural networks, representation learning, and quantum artificial intelligence.

Fangwei Hao is currently pursuing the Ph.D. degree with the College of Artificial Intelligence, Nankai University, Tianjin, China. His main research focuses on image processing based on deep learning.

Jing Xu (Member, IEEE) received the Ph.D. degree from Nankai University, Tianjin, China, in 2003. She is currently a Professor with the College of Artificial Intelligence, Nankai University. She has authored or co-authored more than 100 articles in software engineering, software security, and big data analytics. She was a recipient of the Second Prize of the Tianjin Science and Technology Progress Award twice in 2017 and 2018, respectively.

Ping Li (Member, IEEE) received the Ph.D. degree in computer science and engineering from The Chinese University of Hong Kong, Hong Kong, in 2013. He is currently an Assistant Professor with the Department of Computing, The Hong Kong Polytechnic University, Hong Kong. He has published over 300 top-tier scholarly research articles, pioneered several new research directions, and made a series of landmark contributions in his areas. He has an excellent research project reported by the *ACM TechNews*, which only reports the top breakthrough news in computer science worldwide. More importantly, however, many of his research outcomes have strong impacts on research fields, addressing societal needs and contributing tremendously to the people concerned. His current research interests include AIGC, image/video generation, stylization, colorization, artistic rendering and synthesis, realism in non-photorealistic rendering, computational art, and creative media. He is a fellow of the Computer Graphics Society. He is an Associate Editor of IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE.