Beyond Single Images: Retrieval Self-Augmented Unsupervised Camouflaged Object Detection

Ji Du[1,2], Xin Wang[2], Fangwei Hao[1,*], Mingyang Yu[1]
Chunyuan Chen[1], Jiesheng Wu[3], Bin Wang[1], Jing Xu[1,*], Ping Li[2,*]
[1]College of Artificial Intelligence, Nankai University, China
[2]Department of Computing, The Hong Kong Polytechnic University, Hong Kong
[3]School of Computer and Information, Anhui Normal University, China

Abstract

*At the core of Camouflaged Object Detection (COD) lies segmenting objects from their highly similar surroundings. Previous efforts navigate this challenge primarily through image-level modeling or annotation-based optimization. Despite advancing considerably, this commonplace practice hardly taps valuable dataset-level contextual information or relies on laborious annotations. In this paper, we propose RISE, a **R**etr**I**eval **SE**lf-augmented paradigm that exploits the entire training dataset to generate pseudo-labels for single images, which could be used to train COD models. RISE begins by constructing prototype libraries for environments and camouflaged objects using training images (without ground truth), followed by K-Nearest Neighbor (KNN) retrieval to generate pseudo-masks for each image based on these libraries. It is important to recognize that using only training images without annotations exerts a pronounced challenge in crafting high-quality prototype libraries. In this light, we introduce a Clustering-then-Retrieval (CR) strategy, where coarse masks are first generated through clustering, facilitating subsequent histogram-based image filtering and cross-category retrieval to produce high-confidence prototypes. In the KNN retrieval stage, to alleviate the effect of artifacts in feature maps, we propose Multi-View KNN Retrieval (MVKR), which integrates retrieval results from diverse views to produce more robust and precise pseudo-masks. Extensive experiments demonstrate that RISE outperforms state-of-the-art unsupervised and prompt-based methods. Code is available at* https://github.com/xiaohainku/RISE.

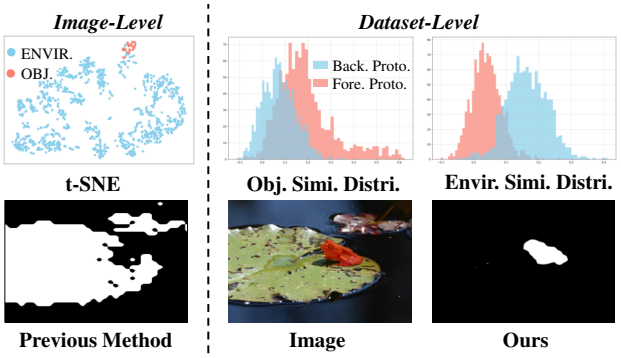

Figure 1. **LEFT**: t-SNE visualization of DINOv2 [54] features and segmentation results from previous unsupervised methods. The highly similar characteristics between the camouflaged object and its environment cause the intra-image similarity-based approach to underperform on the COD task. **RIGHT**: Similarity distributions of global features of camouflaged objects and environments on the respective foreground and background prototype libraries. These prototype libraries are constructed from the COD dataset. Both the camouflaged object and the environment show a higher similarity to their corresponding prototype libraries, suggesting that dataset-level information can be effectively utilized to distinguish between the highly similar foreground and background in a single image.

1. Introduction

Camouflaged Object Detection (COD) [14] is dedicated to segmenting objects meticulously concealed within their surroundings. Existing research efforts have predominantly concentrated on leveraging contextual information **within individual images** to accurately delineate and extract camouflaged objects from visually homogeneous backgrounds.

Mainstream supervised learning methods, including fully supervised [15, 32, 48, 55, 71, 75], weakly supervised [7, 8, 20, 25, 73], and semi-supervised [37, 79], exploit varying degrees of **single-image-level** supervised signals, such as dense annotations, scribbles, bounding boxes, and categories, to model the intrinsic relationship between

*Corresponding authors: Fangwei Hao, Jing Xu (xujing@nankai.edu.cn) and Ping Li (p.li@polyu.edu.hk).

camouflaged objects and their environments. Despite advancements, these methods remain constrained, to varying degrees, by the time-consuming annotation process.

To address this limitation, prompt-based segmentation [28, 29, 65] has emerged as a promising alternative. This approach leverages pre-trained foundation models and task-specific prompt words to localize camouflaged objects, subsequently generating prompts for the Segment Anything Model (SAM) [35] to perform segmentation. However, this paradigm still relies on some form of supervision and is fundamentally constrained by the limited camouflage-specific contextual understanding embedded **in a single image**.

Unsupervised COD, which further omits task-specific prompts compared to prompt-based segmentation, thus far represents a largely uncharted research territory. A straightforward practice is to migrate general unsupervised methods [53, 61, 68] to COD. However, as highlighted in Fig. 1, these methods, which rely solely on feature similarity **within a single image**, often struggle to differentiate between the foreground and background due to the high similarity between the camouflaged object and its surroundings.

Both prompt-based segmentation and unsupervised approaches demonstrate substantial performance gaps compared to supervised methods, primarily attributable to insufficient COD-specific contextual understanding. This limitation manifests in two critical aspects: firstly, foundation models trained on general datasets are not optimized for the unique challenges posed by COD; more significantly, both paradigms focus predominantly on **intra-image relationships**, ignoring potentially valuable **dataset-level contextual information** that could significantly enhance the differentiation between foreground and background objects.

Without fine-tuning foundation models to enrich camouflage-specific contextual understanding, we would like to ask: *how could dataset-level contextual information be effectively leveraged to segment targets in single images?*

In this light, we introduce RISE, a retrieval self-augmented paradigm designed to effectively utilize dataset-level contextual information. RISE segments camouflaged objects by retrieving prototypes from libraries derived directly from the COD dataset. Unlike conventional retrieval-augmented methods [3, 34, 69], which rely on external data sources for prototype libraries, RISE extracts high-quality prototypes directly from the camouflaged dataset itself, in an annotation-free manner.

In the absence of annotations, we introduce Clustering-then-Retrieval (CR), a method to extract high-quality, noise-resistant prototypes for both the environment and camouflaged objects from available images. CR begins by generating a coarse mask for each image using spectral clustering. Based on this mask, CR then constructs prototypes for the camouflaged object and environment by retrieving features that exhibit the least similarity to the respective background or foreground regions.

During the retrieval stage, we use K-Nearest Neighbor (KNN) to identify feature categories. However, artifacts in the foundation model's feature maps may introduce noise into the retrieval results. While fine-tuning the model with new tokens is a common solution [11], we propose a simpler yet effective alternative: Multi-View KNN Retrieval (MVKR). Recognizing that the location of these artifacts can vary with the image's viewpoint, MVKR mitigates their impact by combining retrieval results from images captured from multiple viewpoints, without any fine-tuning.

Comprehensive experiments across a range of benchmark datasets substantiate the effectiveness of our proposed method in segmenting camouflaged objects from their highly similar surroundings. Additionally, our approach significantly reduces the time required to generate high-quality pseudo-masks. In contrast to prompt-based segmentation methods, which may take **days** to generate masks for the entire dataset (4,040 images), our method completes the task in **hours**, with far less GPU memory usage. We succinctly summarize the key contributions of our work as follows:

- We present a new paradigm for unsupervised COD by leveraging the COD dataset to construct prototype libraries, which serve as the foundation for retrieving camouflaged objects. In contrast to existing methods that rely on the features of individual images, our approach harnesses the global information at the dataset level, enabling a more accurate capture of the subtle distinctions between camouflaged objects and their background.
- We propose CR to extract environmental and camouflaged object prototypes from COD datasets. CR utilizes available camouflaged images, rather than external ones, and incorporates a unique joint clustering and retrieval mechanism to distinguish and mine prototypes, eliminating the need for manual annotations. Based on the prototype libraries, we propose Multi-View KNN Retrieval to segment camouflaged objects from the environment.
- Our extensive experiments validate the effectiveness of the proposed method.

2. Related Work

2.1. Camouflaged Object Detection

Camouflaged Object Detection has garnered growing research interest credited to its unique challenges and vast utility [16, 72]. Existing methods can be categorized based on their degree of dependence on annotations into fully supervised learning, weakly supervised learning, prompt-based segmentation, and unsupervised learning.

As the most fully developed and mature branch, advances in fully supervised methods have consistently embraced the theme of maximizing the use of dense annota-

tions. By developing camouflage-specific modules [14, 15, 17, 19, 30, 31, 42, 51, 52, 70, 74, 76, 77, 80], incorporating additional supervised signals (boundary [6, 64, 86], frequency [43, 44, 85], depth [71], category [56, 81]), or adopting innovative learning paradigms (joint modeling [40, 49], uncertainty-guided learning [75], bio-inspired mechanism [32, 55], unfolding network [24], generation-based strategy [9, 21, 38, 63, 78, 82], general model [47, 48, 83, 84]), fully supervised learning seeks to effectively model the relationship between camouflaged objects and their environment based on manual labeling. However, this advancement comes with a non-negligible labeling cost, as labeling a single image can take up to an hour [14].

To alleviate the burden of pixel-level dense annotations, weakly supervised methods leverage easily accessible sparse annotations—such as scribbles [25], bounding boxes [79], points [7], or categories [26]—to segment camouflaged objects in a cost-effective manner. This is achieved through carefully designed loss functions [25], refined pseudo-masks from pre-trained segmentation models [20, 22], or bio-inspired detection processes [7]. To further reduce the need for labeling, another line of research, prompt-based segmentation [28, 29, 65], explores the "foundation models + SAM" paradigm for training-free COD using task-level prompt words. However, both types of methods exhibit significant performance gaps compared to fully supervised methods, due to limited information in sparse annotations or limited generalization abilities of foundation models when applied to camouflage contexts rather than general scenarios [45].

Compared to previous methods, unsupervised approaches are particularly challenging, as they do not rely on labels or task-specific prompt words, making them an underexplored area in COD. Conventional unsupervised methods [53, 60, 61, 66] typically differentiate between foreground and background based on feature similarity within a single image. A common paradigm [1, 10, 67, 68] is to construct a graph where features serve as nodes and the similarities between them form the edges, then apply Normalized Cut [59] to the graph for node segmentation. However, the highly similar characteristics of foreground and background in a single image cause these unsupervised methods, which perform well on general images, to underperform in COD.

Our approach pertains to unsupervised COD. Unlike traditional unsupervised methods that rely on image-level feature similarity, our proposed RISE focuses on dataset-level similarity and seeks to utilize cross-image information to separate camouflaged objects from the background.

2.2. Retrieval-Augmented Methods

In the context of natural language processing, Retrieval-Augmented Generation (RAG) [4] aims to enhance the accuracy and richness of generated content by combining information retrieval and generative models to retrieve relevant information from external knowledge bases. In the field of computer vision, the emergence of foundation models, especially self-supervised [5, 54] and generative models [23, 27, 58], has spawned a new paradigm for segmentation: retrieval-augmented semantic segmentation [3, 34, 69]. This line of research focuses on crafting prototype libraries of different categories through these foundation models and accomplishing segmentation of specific categories through retrieving from the prototype libraries.

Our proposed retrieval self-augmented method, RISE, draws parallels to retrieval-augmented semantic segmentation. However, instead of relying on external models to generate prototype libraries, our approach extracts high-quality prototypes from the COD dataset itself through our unique clustering-then-retrieval mechanism.

3. Method

In this paper, we tackle the problem of unsupervised camouflaged object detection with a simple retrieval self-augmented pipeline. Our work draws inspiration from conventional unsupervised segmentation methods [61, 68], which capitalize on the similarity between self-supervised representations to segment objects. While these methods primarily focus on feature similarity within a single image, our proposed retrieval-based approach leverages dataset-level information to effectively segment camouflaged objects. The overview of our retrieval self-augmented pipeline is illustrated in Fig. 2. First, we introduce Clustering-then-Retrieval (CR), a method that generates prototype libraries for camouflaged objects and environments using the COD dataset (Sec. 3.1). Second, leveraging the prototype libraries from CR, we apply KNN retrieval to each image in the training set to generate pseudo-masks, which could be used to train the model (Sec. 3.2).

3.1. Clustering-then-Retrieval

Unlike retrieval-augmented segmentation [34], which uses diffusion models to generate prototypes, we aim to mine domain-adapted prototypes directly from the COD dataset. However, in the absence of annotations, the key challenge lies in distinguishing between foreground and background to ensure that the resulting prototype library contains minimal noise. To this end, we propose CR to mine high-quality prototypes from the dataset. CR first employs spectral clustering to generate coarse masks, which roughly distinguish camouflaged objects from the environment. It then uses mask-averaged pooling to obtain global features that characterize either the foreground or the background. Based on these global features, CR employs retrieval to select appropriate local features as the prototypes.

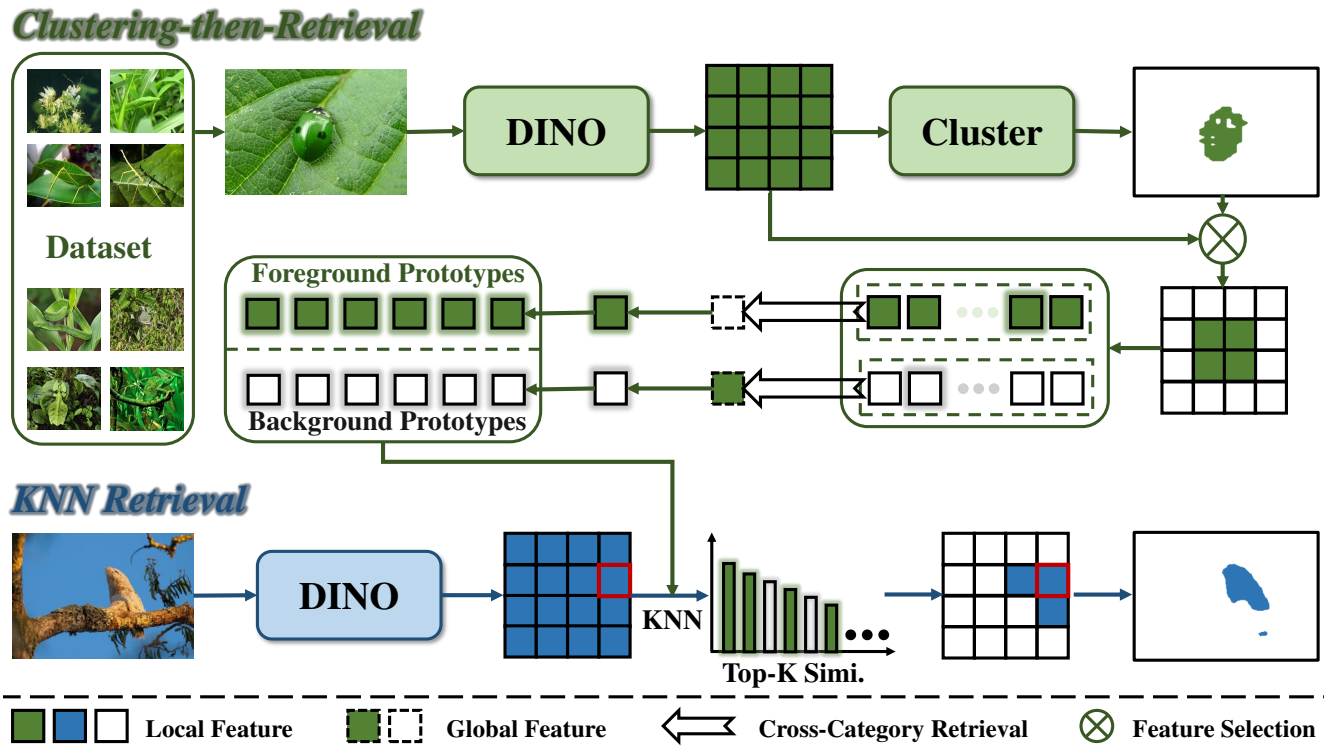

Figure 2. Overview of RISE. RISE is composed of two main stages: Clustering-then-Retrieval (CR) and KNN Retrieval. These stages work together to generate prototype libraries and retrieve camouflaged objects. In the CR phase, we start by clustering the feature maps from DINO for each image in the dataset, which generates a coarse mask. This mask allows us to extract both local and global features for the foreground and background. Using cross-category retrieval, we then retrieve high-quality prototypes from the local features. By aggregating the prototypes across all images, we create the final foreground and background prototype libraries. In the second stage, KNN retrieval, we apply KNN to identify the top-K most similar prototypes in prototype libraries for each feature in the feature map. A voting mechanism is then used to classify each feature as either foreground or background.

Spectral clustering The first step for CR is to employ spectral clustering to generate coarse masks. Compared to clustering high-dimensional features directly, spectral clustering maps the data to a low-dimensional space using feature decomposition by constructing the feature similarity matrix and Laplacian matrix, and then clustering using KMeans.

Given the image $\mathbf{I} \in \mathbb{R}^{H\times W\times 3}$ from COD training set, we utilize self-supervised model DINOv2 [54] to extract the feature map $\mathbf{F} \in \mathbb{R}^{h\times w\times d}$, where d denotes feature dimension. We then construct an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ based on the feature maps, with local features $\mathbf{F_{i,j}} \in \mathbb{R}^{\mathbf{d}}$ as nodes and cosine similarity between features as edges. After getting the flattened feature map $\mathbf{F'} \in \mathbb{R}^{hw\times d}$, the adjacency matrix $\mathbf{W}$ of the graph $\mathcal{G}$ is formulated as

$$\mathbf{W_{i,j}} = \frac{\mathbf{F'_i} \cdot \mathbf{F'_j}}{\|\mathbf{F'_i}\|\|\mathbf{F'_j}\|}. \tag{1}$$

In this paper, we consider negative similarity as 0,

$$\mathbf{W_{i,j}} = \max(\mathbf{W_{i,j}}, 0). \tag{2}$$

The Laplacian matrix for $\mathcal{G}$ is then given by $\mathbf{D} - \mathbf{W}$, where $\mathbf{D}$ with $\mathbf{D_{i,i}} = \sum_j \mathbf{W_{i,j}}$ indicates the diagonal matrix. The normalized Laplacian could be derived by

$$\mathbf{L} = \mathbf{D}^{-\frac{1}{2}}(\mathbf{D} - \mathbf{W})\mathbf{D}^{-\frac{1}{2}}. \tag{3}$$

We take the eigenvectors of $\mathbf{L}$ as new features and adopt KMeans to cluster the features into two clusters. A simple prior is then adopted to assign foreground or background categories to these two clusters: the foreground is usually at the center of the image, and thus occupies a lower proportion of border pixels compared to the background. In this way, we obtain the coarse mask $\mathbf{M} \in \mathbb{R}^{h\times w}$.

Cross-category retrieval The second step for CR is to employ a retrieval-based method to extract prototypes from images. Given the mask $\mathbf{M}$ from spectral clustering, We employ mask-averaged pooling to obtain the global features for the foreground and background, respectively.

$$\mathbf{F_f^g} = \frac{\sum_{i,j} \mathbf{F} \odot \mathbf{M}}{\sum_{i,j} \mathbf{M}}, \tag{4}$$

$$\mathbf{F_b^g} = \frac{\sum_{i,j} \mathbf{F} \odot (\mathbf{1} - \mathbf{M})}{\sum_{i,j} (\mathbf{1} - \mathbf{M})}. \tag{5}$$

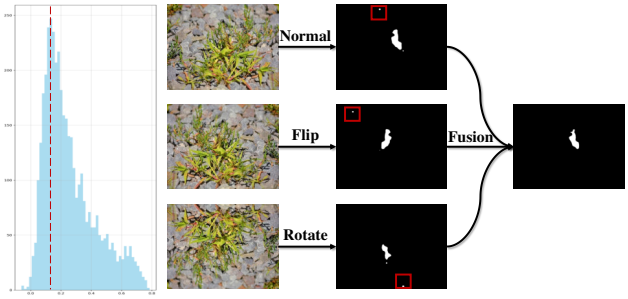

Figure 3. **LEFT:** Histogram of the global feature similarity distribution across all images in the dataset. In this study, we focus only on the similarity values to the left of the peak (indicated by the red dashed line). **RIGHT:** Multi-View KNN Retrieval. To mitigate noisy retrieval results (highlighted in the red box) caused by artifacts in the feature map, we combine the masks generated by transforming the image through different viewpoints, resulting in a final fused mask.

It is notable that despite some noise in M, average pooling enables each global feature to roughly characterize the corresponding category. We then derive the foreground and background feature sets by

$$\mathbf{S_f} = \{\mathbf{F_{i,j}} \mid \mathbf{M_{i,j}} = \mathbf{1}, 1 \le i \le h, 1 \le j \le w\}, \quad (6)$$

$$\mathbf{S_b} = \{\mathbf{F_{i,j}} \mid \mathbf{M_{i,j}} = \mathbf{0}, 1 \le i \le h, 1 \le j \le w\}. \quad (7)$$

To select the foreground prototype, we identify the feature from the foreground set $\mathbf{S_f}$ that is least similar to the global background feature $\mathbf{F_b^g}$. This can be expressed as:

$$\mathbf{P^f} = \arg\min_{\mathbf{s} \in \mathbf{S_f}} \left(\frac{\mathbf{s} \cdot \mathbf{F_b^g}}{\|\mathbf{s}\| \|\mathbf{F_b^g}\|} \right). \quad (8)$$

Notably, we do not select the foreground prototype based on the most similar feature to the global foreground feature $\mathbf{F_f^g}$. This design choice is intentional: by choosing the foreground feature least similar to the global background feature, we enhance the distinction between the foreground and background prototypes. This differentiation is crucial for distinguishing camouflaged objects from the environment during the KNN retrieval stage. Our ablation experiments in Tab. 2 show that this cross-category retrieval strategy significantly improves performance.

Similarly, the background prototype is chosen by selecting the feature that is least similar to the global foreground feature from the background feature set:

$$\mathbf{P^b} = \arg\min_{\mathbf{s} \in \mathbf{S_b}} \left(\frac{\mathbf{s} \cdot \mathbf{F_f^g}}{\|\mathbf{s}\| \|\mathbf{F_f^g}\|} \right). \quad (9)$$

Histogram-based image filtering The prototype libraries are constructed by aggregating prototypes from each image in the dataset. However, since spectral clustering may yield inaccurate segments for certain images, the prototypes from these images may not accurately represent the corresponding categories. Given that erroneous clustering reduces the distinction between foreground and background, thereby increasing the similarity between their global features, we propose a histogram-based adaptive thresholding method to filter out images with high similarity.

We first employ spectral clustering to generate coarse masks for each image in the dataset and calculate the global feature similarity $\frac{\mathbf{F_f^g} \cdot \mathbf{F_b^g}}{\|\mathbf{F_f^g}\| \|\mathbf{F_b^g}\|}$ for each image based on these masks. The histogram of these similarities is shown on the left of Fig. 3. The distribution roughly forms a single peak. In this paper, we select the similarity corresponding to the peak's vertex as the adaptive threshold, considering only images with a similarity below this threshold. We denote the final foreground and background prototype libraries by $\{\mathbf{P_i^f}\}_{i=1}^n$ and $\{\mathbf{P_i^b}\}_{i=1}^n$, respectively, where n indicates the number of prototypes.

3.2. Multi-View KNN Retrieval

In this part, we introduce using KNN retrieval to generate pseudo-masks for each image in the dataset based on prototype libraries from CR. Similarly, given the image $\mathbf{I} \in \mathbb{R}^{H \times W \times 3}$, we first extract its feature map $\mathbf{F} \in \mathbb{R}^{h \times w \times d}$ using DINOv2. For each local feature $\mathbf{F_{i,j}} \in \mathbb{R}^d$, we apply KNN to find the top-K most similar prototypes from $\{\mathbf{P_i^f}\}_{i=1}^n$ and $\{\mathbf{P_i^b}\}_{i=1}^n$, using cosine similarity to measure the similarity between $\mathbf{F_{i,j}}$ and the prototypes. To classify each feature $\mathbf{F_{i,j}}$ as foreground or background, we use a voting mechanism over the top-K retrieved prototypes. This process is repeated for all features in $\mathbf{F}$, yielding the pseudo-mask $\mathbf{m} \in \mathbb{R}^{h \times w}$. After upsampling, the final mask matching the resolution of $\mathbf{I}$ could be obtained.

However, artifacts (features that do not reflect the true semantics of the image) in the feature maps may introduce noise into the retrieval results. Since the location of these artifacts varies with the image's viewpoint, we propose a Multi-View KNN Retrieval (MVKR) method that does not require fine-tuning. MVKR reduces the impact of artifacts by combining retrieval results from images derived from different viewpoints.

As illustrated on the right of Fig. 3, we first generate multiple views of the same image by applying transformations such as flipping or rotating. For each transformed image, we perform KNN retrieval to obtain the corresponding mask for the given viewpoint. To recover the mask for the normal view, we apply an inverse transformation. Finally, we use a voting mechanism to combine these masks into a single, unified result.

Method	Feat. Extra. / SAM	Venue	CHAMELEON				CAMO				COD10K				NC4K			
			$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
Unsupervised Segmentaion																		
LOST	DINO-ViT-S16	BMVC21	0.631	0.713	0.418	0.145	0.644	0.688	0.450	0.166	0.673	0.722	0.433	0.095	0.705	0.748	0.532	0.118
DeepSpectral	DINO-ViT-S8	CVPR22	0.662	0.677	0.465	0.187	0.640	0.659	0.457	0.215	0.621	0.625	0.363	0.162	0.719	0.739	0.542	0.133
DeepSpectral	DINOv2-ViT-L14	CVPR22	0.605	0.613	0.402	0.225	0.560	0.563	0.366	0.292	0.493	0.472	0.227	0.314	0.588	0.588	0.378	0.255
TokenCut	DINO-ViT-S16	CVPR22	0.687	0.757	0.516	0.115	0.661	0.717	0.498	0.157	0.682	0.729	0.468	0.095	0.752	0.804	0.614	0.093
TokenCut	DINOv2-ViT-L14	CVPR22	0.708	0.747	0.510	0.078	0.608	0.612	0.385	0.148	0.637	0.672	0.370	0.077	0.697	0.731	0.511	0.095
MaskCut	DINO-ViT-B8	CVPR23	0.636	0.662	0.448	0.194	0.646	0.653	0.454	0.206	0.626	0.628	0.363	0.170	0.708	0.717	0.518	0.152
MaskCut	DINOv2-ViT-L14	CVPR23	0.653	0.665	0.462	0.169	0.628	0.656	0.439	0.200	0.607	0.624	0.346	0.165	0.683	0.710	0.498	0.152
FOUND	DINO-ViT-S8	CVPR23	0.609	0.616	0.374	0.215	0.572	0.571	0.379	0.280	0.522	0.503	0.252	0.247	0.610	0.613	0.399	0.216
FOUND	DINOv2-ViT-L14	CVPR23	0.656	0.731	0.459	0.115	0.590	0.635	0.378	0.166	0.636	0.710	0.394	0.088	0.669	0.734	0.483	0.113
ProMerge	DINO-ViT-B8	ECCV24	0.712	0.757	0.534	0.087	0.692	0.732	0.542	0.134	0.714	0.756	0.504	0.078	0.773	0.817	0.639	0.081
ProMerge	DINOv2-ViT-L14	ECCV24	0.741	0.787	0.567	0.085	0.679	0.706	0.502	0.142	0.674	0.714	0.435	0.081	0.726	0.771	0.558	0.096
VoteCut	DINO(v2) Ensemble	CVPR24	0.699	0.763	0.556	0.138	0.663	0.708	0.506	0.158	0.702	0.763	0.504	0.092	0.756	0.812	0.628	0.094
VoteCut	DINOv2-ViT-L14	CVPR24	0.679	0.695	0.484	0.110	0.580	0.572	0.356	0.164	0.645	0.673	0.390	0.082	0.674	0.694	0.476	0.104
DiffCut	SSD-1B	NeurIPS24	0.574	0.613	0.390	0.220	0.627	0.662	0.454	0.185	0.628	0.667	0.372	0.120	0.693	0.739	0.514	0.122
RISE	DINO-ViT-S16	-	0.670	0.738	0.489	0.095	0.655	0.712	0.494	0.135	0.716	0.790	0.518	0.063	0.754	0.814	0.627	0.083
RISE	DINO-ViT-B16	-	0.679	0.741	0.508	0.090	0.664	0.715	0.507	0.132	0.724	0.795	0.526	0.062	0.759	0.818	0.633	0.081
RISE	DINOv2-ViT-S14	-	0.762	0.810	0.631	0.068	0.684	0.726	0.532	0.127	0.741	0.814	0.564	0.059	0.788	0.847	0.674	0.069
RISE	DINOv2-ViT-B14	-	0.805	0.858	0.675	0.052	0.722	0.775	0.587	0.113	0.753	0.827	0.578	0.053	0.797	0.860	0.687	0.064
RISE	DINOv2-ViT-L14	-	**0.822**	**0.884**	**0.720**	**0.050**	**0.734**	**0.787**	**0.610**	**0.109**	**0.763**	**0.840**	**0.600**	**0.049**	**0.805**	**0.868**	**0.705**	**0.061**
Prompt-based Segmentation																		
WS-SAM*	SAM-ViT-H	NeurIPS23	0.795	0.824	0.676	0.099	**0.781**	0.807	**0.658**	0.108	0.787	0.838	0.622	0.057	**0.829**	0.867	0.727	0.063
GenSAM	SAM-ViT-H	AAAI24	0.659	0.716	0.495	0.153	0.633	0.673	0.458	0.188	0.641	0.675	0.390	0.136	0.702	0.744	0.524	0.128
ProMac	SAM-ViT-H	NeurIPS24	0.786	0.842	0.665	0.066	0.754	**0.812**	0.645	**0.101**	0.774	0.835	0.609	0.052	0.812	0.862	0.711	0.065
RISE	SAM-ViT-H	-	**0.823**	**0.882**	**0.733**	**0.055**	0.760	0.807	0.651	0.102	**0.790**	**0.854**	**0.643**	**0.044**	0.825	**0.874**	**0.736**	**0.056**

Table 1. Comparisons with unsupervised and prompt-based segmentation methods across four commonly used benchmark datasets. "↑/↓": The higher/lower the better. "*": The prompts for SAM are derived by manual labeling. The best and second-best results are **bolded** and underlined to highlight, respectively.

4. Experiment

4.1. Experimental setup

Datasets and evaluation metrics Following previous works, our training dataset consists of 3,040 images from COD10K [14] and 1,000 images from CAMO [39]. We test our method on four widely used benchmarks: CHAMELEON (76 test images) [62], CAMO (250 test images), COD10K (2,026 test images), and NC4K (4,121 test images) [49]. We employ four metrics for evaluation, including structure measure (S_α) [12], mean E-measure (E_ϕ) [13], weighted F-measure (F_β^ω) [50] and mean absolute error (M) [57].

Implementation details In this paper, we focus on generating pseudo-masks for training COD models, rather than designing the models themselves. We choose SINet-V2 [15] as the training model, and the whole training process is identical to SINet-V2. We employ the self-supervised model DINOv2-ViT-L14 [54] as the feature extractor and extract features from the last layer. During Multi-View KNN Retrieval, we retrieve the top-K most similar prototypes from the prototype libraries, with K set to 512. To generate multiple views of each image, we apply horizontal and vertical flipping, as well as rotations (90°, 180°, and 270°). The FAISS library [33] is used for efficient retrieval. All images are resized to 476×476.

4.2. Comparisons with state-of-the-arts

Comparison methods We compare our method with both unsupervised and prompt-based segmentation approaches. For the unsupervised setting, we re-implement eight state-of-the-art methods on COD using their official codes. These include seven DINO(v2)-based methods (DeepSpectral [53], LOST [60], FOUND [61], TokenCut [68], MaskCut [67], VoteCut [1], ProMerge [41]), and one diffusion model-based method, DiffCut [10]. For the DINO(v2)-based methods, following CuVLER [1], we resize images to 480×480 for DINO and 476×476 for DINOv2. To ensure a fair comparison, we remove all post-processing techniques, such as bilateral solvers [2] and conditional random fields [36]. For methods like MaskCut that produce multiple masks, we combine them into a single binary mask. We also observe that some DINO-based methods, such as ProMerge, perform poorly with DINOv2 as the feature extractor. In this situation, we tune the main hyperparameters and experiment with features at different locations (Query, Key, Value, QKV, and last layer) to optimize performance. For prompt-based setting, we integrate our method with SAM [35] by extracting bounding boxes from pseudo-masks, which serve as prompts for SAM. We select three methods for comparisons: GenSAM [28], ProMac [29], and WS-SAM [20]. For GenSAM and ProMac, we use their official codes to generate pseudo-labels on the training set. It is notable that for ProMac we adopt LLaVA-1.5-7B [46] as LMM instead of LLaVA-1.5-13B due to limited GPU memory. For WS-SAM, we use the pseudo-masks provided by the authors.

Quantitative comparison In comparison to unsupervised methods, as shown in Tab. 1, RISE outperforms all state-of-the-art techniques. On the COD10K dataset, our method achieves a minimum improvement of 8% and 9% in metrics E_ϕ and F_β^ω, respectively. Notably, RISE delivers top-tier performance on the more challenging COD10K and NC4K datasets, regardless of the DINO configuration. These results highlight the advantage of our approach, which lever-

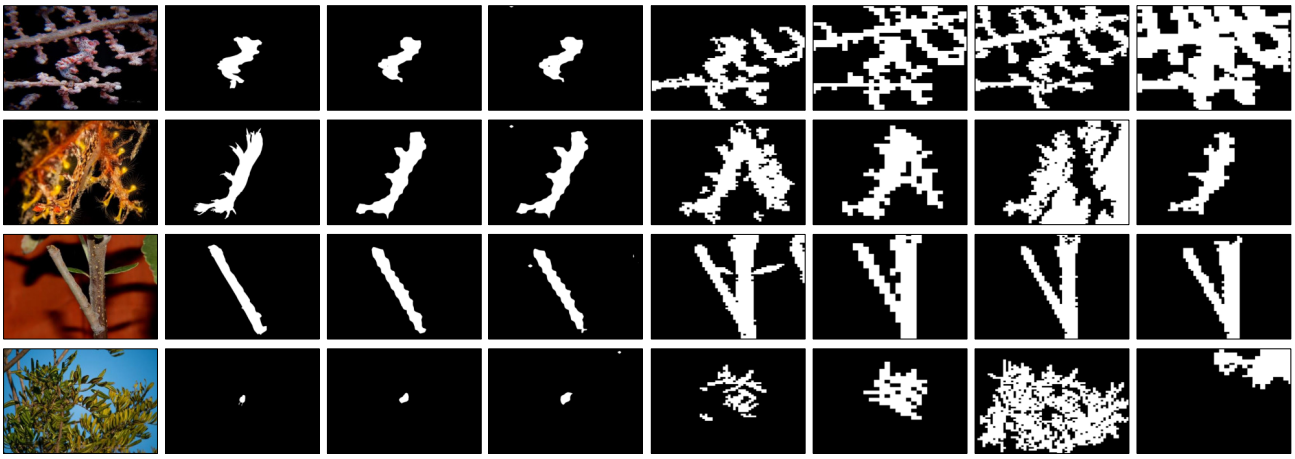

Figure 4. Qualitative comparisons with four SOTA unsupervised methods. RISE* denotes RISE without Multi-View Retrieval.

Row index	Component	Ablation	CHAMELEON				CAMO				COD10K				NC4K			
			$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
a	Paradigm	w image-level modeling	0.743	0.782	0.607	0.133	0.665	0.690	0.506	0.197	0.641	0.662	0.414	0.169	0.727	0.760	0.569	0.134
b	CR	w/o Cross-Category Retrieval	0.787	0.851	0.660	0.063	0.743	0.808	0.615	0.107	0.710	0.781	0.518	0.065	0.764	0.833	0.639	0.076
c		w/o Histogram-based Filtering	0.820	0.878	0.718	0.052	0.733	0.791	0.609	0.112	0.744	0.822	0.575	0.055	0.793	0.859	0.688	0.065
d	KNN Retrieval	w/o Multi-View Retrieval	0.829	0.885	0.719	0.052	0.735	0.789	0.604	0.115	0.759	0.832	0.584	0.052	0.803	0.863	0.694	0.064
e	Ours	-	0.822	0.884	0.720	0.050	0.734	0.787	0.610	0.109	0.763	0.840	0.600	0.049	0.805	0.868	0.705	0.061

Table 2. Ablation experiments on the effect of each component.

ages dataset-level information to retrieve camouflaged objects, in contrast to traditional unsupervised methods that focus on intra-image similarity for foreground segmentation. When compared to prompt-based segmentation methods, our approach also demonstrates superior performance. Despite WS-SAM [20] providing prompts for SAM based on weakly supervised signals from manually labeled data, RISE outperforms it in most metrics, showcasing its enhanced ability to localize camouflaged objects. Additionally, GenSAM [28] and ProMac [29] leverage multimodal LLMs or diffusion models to generate prompts. However, due to the high parameter intensity of these models, generating pseudo-masks for the entire training dataset can take days. In contrast, RISE uses only self-supervised models, significantly reducing inference time.

Qualitative comparison To illustrate RISE's ability to accurately localize and segment camouflaged objects in complex environments, we present a qualitative comparison with four leading unsupervised methods for generating pseudo-masks, as shown in Fig. 4. These methods, which rely on intra-image similarity, often fail to differentiate between targets and backgrounds with similar appearances, leading to suboptimal segmentation. In contrast, RISE overcomes this challenge by incorporating both foreground and background semantics at the dataset level, resulting in more effective separation. Notably, our method performs well in localizing small objects, as shown in the last row of Fig. 4.

4.3. Ablation Study

Paradigm Our retrieval self-augmented pipeline combines a prototype generator (CR) with a KNN retrieval scheme. We first assess the effectiveness of our proposed paradigm for retrieving camouflaged objects using the COD dataset, comparing it to previous approaches that rely on single-image modeling. Spectral clustering, used in the CR stage, serves as the baseline. As shown in rows **a** and **e** of Tab. 2, our method outperforms the baseline by more than 10% on average across all datasets and metrics. This underscores the importance of fully utilizing dataset-level semantics to distinguish between highly similar foregrounds and backgrounds in a single image.

Clustering-then-Retrieval The CR module incorporates two key components: cross-category retrieval and histogram-based image filtering, both designed to generate high-quality, distinguishable prototype libraries. As shown in rows **b** and **e**, incorporating cross-category retrieval improves RISE performance on the challenging COD10K dataset by 5.3%, 5.9%, and 8.2% for the S_α, E_ϕ, and F_β^ω metrics, respectively. Furthermore, histogram-based image filtering not only boosts performance across all datasets (as shown in rows **c** and **e**) but also reduces the size of the prototype libraries, enhancing inference efficiency.

Multi-View KNN Retrieval In the KNN retrieval stage, we address artifacts in the self-supervised model by aggregating retrieval results from multiple viewpoints. This not only improves performance (as indicated by rows **d** and **e**)

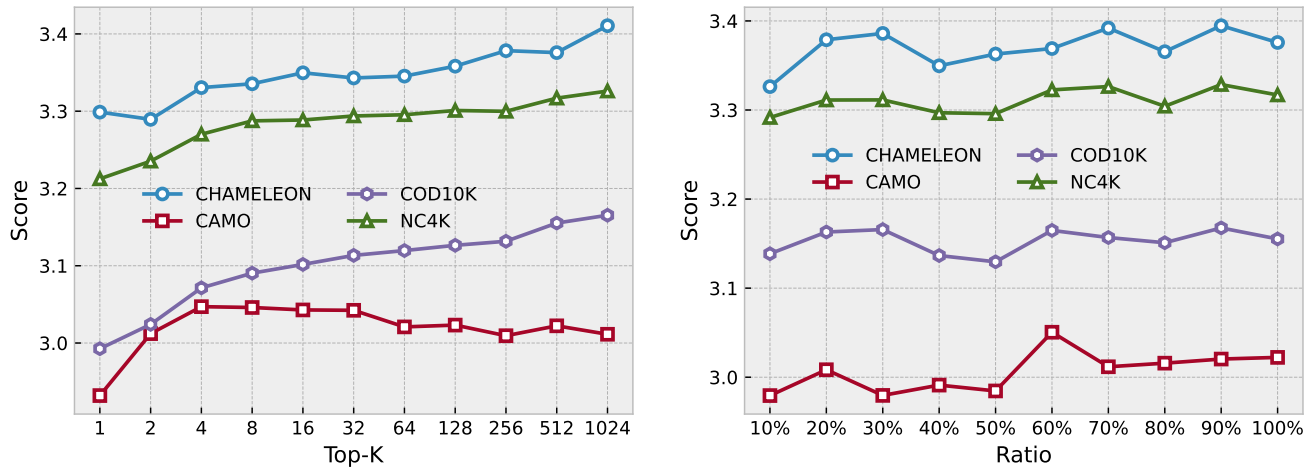

Figure 5. **Left:** Sensitivity analysis of the top-K hyperparameter; **Right:** Impact of the dataset size.

Method	COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
Previous SOTA	0.714	0.756	0.504	0.078	0.773	0.817	0.639	0.081
KMeans	0.400	0.404	0.155	0.410	0.494	0.514	0.303	0.343
KMeans+	**0.723**	**0.788**	**0.536**	**0.073**	**0.777**	**0.836**	**0.656**	**0.078**
GMM	0.450	0.440	0.192	0.371	0.572	0.578	0.371	0.280
GMM+	**0.719**	**0.789**	**0.534**	**0.069**	**0.771**	**0.835**	**0.652**	**0.078**
HCA	0.474	0.484	0.206	0.317	0.579	0.600	0.376	0.257
HCA+	**0.733**	**0.805**	**0.556**	**0.064**	**0.786**	**0.848**	**0.672**	**0.071**
Spectral	0.641	0.662	0.414	0.169	0.727	0.760	0.569	0.134
Spectral+	**0.763**	**0.840**	**0.600**	**0.049**	**0.805**	**0.868**	**0.705**	**0.061**

Table 3. Experiments on different clustering methods. "+": Integrating clustering methods with RISE.

but also ensures that noise-resistant results can be obtained without fine-tuning models, as demonstrated in the third and fourth columns of Fig. 4 (best zoomed-in).

4.4. Further Analysis

Hyperparameter sensitivity Our method has only one hyperparameter: top-K. We adopt $\text{Score} = S_\alpha + E_\phi + F_\beta^\omega + 1 - M$ [18] to indicate the overall performance. As shown in the left of Fig. 5, RISE performs well even with a small K. This robustness stems from the unique CR scheme, which reduces noise in the prototype libraries and enhances the distinction between different libraries.

Dataset size In the right of Fig. 5, we investigate the impact of the dataset size. Leveraging CR's distinctive prototype mining mechanism, RISE demonstrates consistent and robust performance even when only 10% of the training images (randomly selected) are used to construct the prototype libraries. This highlights the efficiency and effectiveness of RISE, particularly in scenarios with limited data.

Clustering methods We explore several clustering methods, including KMeans, Gaussian Mixture Model (GMM), Hierarchical Clustering Analysis (HCA), and Spectral Clustering (Spectral). As shown in Tab. 3, while the performance of each method varies, integrating our approach significantly boosts the performance of all methods.

Method	COD10K				NC4K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$M \downarrow$
LOST	0.673	0.722	0.433	0.095	0.705	0.748	0.532	0.118
LOST+	**0.766**	**0.854**	**0.615**	**0.046**	**0.799**	**0.869**	**0.708**	**0.061**
DeepSpectral	0.621	0.625	0.363	0.162	0.719	0.739	0.542	0.133
DeepSpectral+	**0.765**	**0.842**	**0.605**	**0.050**	**0.804**	**0.860**	**0.702**	**0.063**
TokenCut	0.682	0.729	0.468	0.095	0.752	0.804	0.614	0.093
TokenCut+	**0.769**	**0.848**	**0.607**	**0.048**	**0.813**	**0.876**	**0.715**	**0.059**
MaskCut	0.626	0.628	0.363	0.170	0.708	0.717	0.518	0.152
MaskCut+	**0.763**	**0.848**	**0.608**	**0.047**	**0.793**	**0.856**	**0.698**	**0.063**
ProMerge	0.714	0.756	0.504	0.078	0.773	0.817	0.639	0.081
ProMerge+	**0.770**	**0.850**	**0.622**	**0.045**	**0.805**	**0.868**	**0.715**	**0.059**
VoteCut	0.702	0.763	0.504	0.092	0.756	0.812	0.628	0.094
VoteCut+	**0.766**	**0.845**	**0.608**	**0.048**	**0.807**	**0.869**	**0.710**	**0.061**

Table 4. Experiments on the plug-and-play attribute of RISE. "+": Integrating unsupervised methods with RISE.

Plug-and-play By replacing spectral clustering in CR with unsupervised methods, our approach can be integrated with unsupervised methods. This enhances the performance of these methods, as shown in Tab. 4.

5. Conclusion

In this paper, we offer a new perspective on camouflaged object detection: while a camouflaged object may be indistinguishable from its surroundings in a single image, it may become distinguishable when considered within the context of a dataset. Building on this insight, we introduce RISE, a retrieval self-augmented unsupervised COD paradigm. RISE is designed to separate hard-to-recognize targets within a single image by efficiently integrating and leveraging dataset-level information. To access this dataset-level information, we propose a Clustering-then-Retrieval approach. For segmenting camouflaged objects in a single image, we introduce Multi-View KNN Retrieval. Extensive experiments demonstrate that our method outperforms both unsupervised and prompt-based segmentation approaches.

Acknowledgments. This work was supported in part by The Hong Kong Polytechnic University under Grants P0048387, P0044520, P0050657, and P0049586, and in part by the Tianjin Science and Technology Major Project under Grant 24ZXZSSS00420.

References

[1] Shahaf Arica, Or Rubin, Sapir Gershov, and Shlomi Laufer. CuVLER: Enhanced unsupervised object discoveries through exhaustive self-supervised transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 23105–23114, 2024. 3, 6

[2] Jonathan T. Barron and Ben Poole. The fast bilateral solver. In *European Conference on Computer Vision*, pages 617–632, 2016. 6

[3] Luca Barsellotti, Roberto Amoroso, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Training-free open-vocabulary segmentation with offline diffusion-augmented prototype generation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3689–3698, 2024. 2, 3

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240, 2022. 3

[5] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *IEEE International Conference on Computer Vision*, pages 9650–9660, 2021. 3

[6] Chunyuan Chen, Weiyun Liang, Donglin Wang, Bin Wang, and Jing Xu. Vision-inspired boundary perception network for lightweight camouflaged object detection. *IEEE Signal Processing Letters*, 32:1176–1180, 2025. 3

[7] Huafeng Chen, Dian Shao, Guangqian Guo, and Shan Gao. Just a hint: Point-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 332–348, 2025. 1, 3

[8] Huafeng Chen, Pengxu Wei, Guangqian Guo, and Shan Gao. SAM-COD: Sam-guided unified framework for weakly-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 315–331, 2025. 1

[9] Zhongxi Chen, Ke Sun, and Xianming Lin. CamoDiffusion: Camouflaged object detection via conditional diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 1272–1280, 2024. 3

[10] Paul Couairon, Mustafa Shukor, Jean-Emmanuel Haugeard, Matthieu Cord, and Nicolas Thome. Diffcut: Catalyzing zero-shot semantic segmentation with diffusion features and recursive normalized cut. In *Advances in Neural Information Processing Systems*, pages 1–24, 2024. 3, 6

[11] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, pages 1–21, 2024. 2

[12] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-Measure: A new way to evaluate foreground maps. In *IEEE International Conference on Computer Vision*, pages 4558–4567, 2017. 6

[13] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *International Joint Conference on Artificial Intelligence*, pages 698–704, 2018. 6

[14] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2774–2784, 2020. 1, 3, 6

[15] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6024–6042, 2022. 1, 3, 6

[16] Deng-Ping Fan, Ge-Peng Ji, Peng Xu, Ming-Ming Cheng, Christos Sakaridis, and Luc Van Gool. Advances in deep concealed scene understanding. *Visual Intelligence*, 1(1):1–16, 2023. 2

[17] Chao Hao, Zitong Yu, Xin Liu, Yuhao Wang, Weicheng Xie, Jingang Shi, Huanjing Yue, and Jingyu Yang. Distribution-specific learning for joint salient and camouflaged object detection. *Available at SSRN 5089840*, pages 1–33, 2024. 3

[18] Chao Hao, Zitong Yu, Xin Liu, Jun Xu, Huanjing Yue, and Jingyu Yang. A simple yet effective network based on vision transformer for camouflaged object and salient object detection. *IEEE Transactions on Image Processing*, 34:608–622, 2025. 8

[19] Chunming He, Kai Li, Yachao Zhang, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Camouflaged object detection with feature decomposition and edge reconstruction. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 22046–22055, 2023. 3

[20] Chunming He, Kai Li, Yachao Zhang, Guoxia Xu, Longxiang Tang, Yulun Zhang, Zhenhua Guo, and Xiu Li. Weakly-supervised concealed object segmentation with SAM-based pseudo labeling and multi-scale feature grouping. In *Advances in Neural Information Processing Systems*, pages 30726–30737, 2023. 1, 3, 6, 7

[21] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *International Conference on Learning Representations*, pages 1–10, 2024. 3

[22] Chunming He, Kai Li, Yachao Zhang, Ziyun Yang, Youwei Pang, Longxiang Tang, Chengyu Fang, Yulun Zhang, Linghe Kong, Xiu Li, and Sina Farsiu. Segment concealed objects with incomplete supervision. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–18, 2025. 3

[23] Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4630–4651, 2025. 3

[24] Chunming He, Rihan Zhang, Fengyang Xiao, Chenyu Fang, Longxiang Tang, Yulun Zhang, Linghe Kong, Deng-Ping Fan, Kai Li, and Sina Farsiu. RUN: Reversible unfolding network for concealed object segmentation. *International Conference on Machine Learning*, pages 1–14, 2025. 3

[25] Ruozhen He, Qihua Dong, Jiaying Lin, and Rynson W.H. Lau. Weakly-supervised camouflaged object detection with scribble annotations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 781–789, 2023. 1, 3

[26] Zhentao He, Changqun Xia, Shengye Qiao, and Jia Li. Text-prompt camouflaged instance segmentation with graduated camouflage learning. In *ACM International Conference on Multimedia*, page 5584–5593, 2024. 3

[27] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, pages 6840–6851, 2020. 3

[28] Jian Hu, Jiayi Lin, Shaogang Gong, and Weitong Cai. Relax image-specific prompt requirement in SAM: A single generic prompt for segmenting camouflaged objects. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12511–12518, 2024. 2, 3, 6, 7

[29] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. In *Advances in Neural Information Processing Systems*, pages 1–13, 2024. 2, 3, 6, 7

[30] Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Donghao Luo, Ying Tai, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 881–889, 2023. 3

[31] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. Feature shrinkage pyramid for camouflaged object detection with transformers. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5557–5566, 2023. 3

[32] Qi Jia, Shuilian Yao, Yu Liu, Xin Fan, Risheng Liu, and Zhongxuan Luo. Segment, magnify and reiterate: Detecting camouflaged objects the hard way. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4703–4712, 2022. 1, 3

[33] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[34] Laurynas Karazija, Iro Laina, Andrea Vedaldi, and Christian Rupprecht. Diffusion Models for Open-Vocabulary Segmentation. In *European Conference on Computer Vision*, pages 1–20, 2024. 2, 3

[35] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *IEEE International Conference on Computer Vision*, pages 3992–4003, 2023. 2, 6

[36] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *Advances in Neural Information Processing Systems*, pages 1–9, 2011. 6

[37] Xunfa Lai, Zhiyu Yang, Jie Hu, Shengchuan Zhang, Liujuan Cao, Guannan Jiang, Zhiyu Wang, Songan Zhang, and Rongrong Ji. Camoteacher: Dual-rotation consistency learning for semi-supervised camouflaged object detection. In *European Conference on Computer Vision*, pages 438–455, 2025. 1

[38] Hala Lamdouar, Weidi Xie, and Andrew Zisserman. The making and breaking of camouflage. In *IEEE International Conference on Computer Vision*, pages 832–842, 2023. 3

[39] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019. 6

[40] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10066–10076, 2021. 3

[41] Dylan Li and Gyungin Shin. ProMerge: Prompt and merge for unsupervised instance segmentation. In *European Conference on Computer Vision*, pages 1–17, 2024. 6

[42] Weiyun Liang, Jiesheng Wu, Xinyue Mu, Fangwei Hao, Ji Du, Jing Xu, and Ping Li. Weighted dense semantic aggregation and explicit boundary modeling for camouflaged object detection. *IEEE Sensors Journal*, 24(13):21108–21122, 2024. 3

[43] Weiyun Liang, Jiesheng Wu, Yanfeng Wu, Xinyue Mu, and Jing Xu. Finet: Frequency injection network for lightweight camouflaged object detection. *IEEE Signal Processing Letters*, 31:526–530, 2024. 3

[44] Jiaying Lin, Xin Tan, Ke Xu, Lizhuang Ma, and Rynson W. H. Lau. Frequency-aware camouflaged object detection. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 19(2):1–16, 2023. 3

[45] Fang Liu, Yuhao Liu, Ke Xu, Shuquan Ye, Gerhard Petrus Hancke, and Rynson W. H. Lau. Language-guided salient object ranking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 29803–29813, 2025. 3

[46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 6

[47] Weihuang Liu, Xi Shen, Chi-Man Pun, and Xiaodong Cun. Explicit visual prompting for low-level structure segmentations. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 19434–19445, 2023. 3

[48] Ziyang Luo, Nian Liu, Wangbo Zhao, Xuguang Yang, Dingwen Zhang, Deng-Ping Fan, Fahad Khan, and Junwei Han. VSCode: General visual salient and camouflaged object detection with 2d prompt learning. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17169–17180, 2024. 1, 3

[49] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11586–11596, 2021. 3, 6

[50] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2014. 6

[51] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8768–8777, 2021. 3

[52] Haiyang Mei, Ke Xu, Yunduo Zhou, Yang Wang, Haiyin Piao, Xiaopeng Wei, and Xin Yang. Camouflaged object segmentation with omni perception. *International Journal of Computer Vision*, 131(11):3019–3034, 2023. 3

[53] Luke Melas-Kyriazi, Christian Rupprecht, Iro Laina, and Andrea Vedaldi. Deep spectral methods: A surprisingly strong baseline for unsupervised semantic segmentation and localization. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8364–8375, 2022. 2, 3, 6

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024:1–32, 2024. 1, 3, 4, 6

[55] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2150–2160, 2022. 1, 3

[56] Youwei Pang, Xiaoqi Zhao, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Open-vocabulary camouflaged object segmentation. In *European Conference on Computer Vision*, pages 476–495, 2025. 3

[57] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 733–740, 2012. 6

[58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3

[59] Jianbo Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000. 3

[60] Oriane Siméoni, Gilles Puy, Huy V. Vo, Simon Roburin, Spyros Gidaris, Andrei Bursuc, Patrick Pérez, Renaud Marlet, and Jean Ponce. Localizing objects with self-supervised transformers and no labels. In *Proceedings of the British Machine Vision Conference*, pages 1–25, 2021. 3, 6

[61] Oriane Siméoni, Chloé Sekkat, Gilles Puy, Antonín Vobecký, Éloi Zablocki, and Patrick Pérez. Unsupervised object localization: Observing the background to discover objects. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3176–3186, 2023. 2, 3, 6

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

[63] Ke Sun, Zhongxi Chen, Xianming Lin, Xiaoshuai Sun, Hong Liu, and Rongrong Ji. Conditional diffusion models for camouflaged and salient object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–16, 2025. 3

[64] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In *International Joint Conference on Artificial Intelligence*, pages 1335–1341, 2022. 3

[65] Lv Tang, Peng-Tao Jiang, Zhi-Hao Shen, Hao Zhang, Jin-Wei Chen, and Bo Li. Chain of Visual Perception: Harnessing multimodal large language models for zero-shot camouflaged object detection. In *ACM International Conference on Multimedia*, page 8805–8814, 2024. 2, 3

[66] Xin Tian, Ke Xu, and Rynson Lau. Unsupervised salient instance detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2702–2712, 2024. 3

[67] Xudong Wang, Rohit Girdhar, Stella X. Yu, and Ishan Misra. Cut and learn for unsupervised object detection and instance segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3124–3134, 2023. 3, 6

[68] Yangtao Wang, Xi Shen, Shell Xu Hu, Yuan Yuan, James L. Crowley, and Dominique Vaufreydaz. Self-supervised transformers for unsupervised object discovery using normalized cut. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 14543–14553, 2022. 2, 3, 6

[69] Yuan Wang, Rui Sun, Naisong Luo, Yuwen Pan, and Tianzhu Zhang. Image-to-image matching via foundation models: A new perspective for open-vocabulary semantic segmentation. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3952–3963, 2024. 2, 3

[70] Jiesheng Wu, Weiyun Liang, Fangwei Hao, and Jing Xu. Mask-and-edge co-guided separable network for camouflaged object detection. *IEEE Signal Processing Letters*, 30: 748–752, 2023. 3

[71] Zongwei Wu, Danda Pani Paudel, Deng-Ping Fan, Jingjing Wang, Shuo Wang, Cédric Demonceaux, Radu Timofte, and Luc Van Gool. Source-free depth for object pop-out. In *IEEE International Conference on Computer Vision*, pages 1032–1042, 2023. 1, 3

[72] Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Chunming He, Longxiang Tang, Ziyun Yang, and Xiu Li. A survey of camouflaged object detection and beyond. *CAAI Artificial Intelligence Research*, pages 1–26, 2024. 2

[73] Ke Xu, Tsun Wai Siu, and Rynson W.H. Lau. Zoom: Learning video mirror detection with extremely-weak supervision. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 6315–6323, 2024. 1

[74] Tao Yan, Jiahui Gao, Ke Xu, Xiangjie Zhu, Hao Huang, Helong Li, Benjamin Wah, and Rynson W. H. Lau. Ghostingnet: A novel approach for glass surface detection with ghosting cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(1):323–337, 2025. 3

[75] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *IEEE International Conference on Computer Vision*, pages 4126–4135, 2021. 1, 3

[76] Bowen Yin, Xuying Zhang, Deng-Ping Fan, Shaohui Jiao, Ming-Ming Cheng, Luc Van Gool, and Qibin Hou. CamoFormer: Masked separable attention for camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):10362–10374, 2024. 3

[77] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 12992–13002, 2021. 3

[78] Haichao Zhang, Can Qin, Yu Yin, and Yun Fu. Camouflaged image synthesis is all you need to boost camouflaged detection. *arXiv preprint arXiv:2308.06701*, pages 1–11, 2023. 3

[79] Jin Zhang, Ruiheng Zhang, Yanjiao Shi, Zhe Cao, Nian Liu, and Fahad Shahbaz Khan. Learning camouflaged object detection from noisy pseudo label. In *European Conference on Computer Vision*, pages 158–174, 2025. 1, 3

[80] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. PreyNet: Preying on camouflaged objects. In *ACM International Conference on Multimedia*, page 5323–5332, 2022. 3

[81] Kai Zhao, Wubang Yuan, Zheng Wang, Guanyi Li, Xiaoqiang Zhu, Deng-ping Fan, and Dan Zeng. Open-vocabulary camouflaged object segmentation with cascaded vision language models. *arXiv preprint arXiv:2506.19300*, pages 1–17, 2025. 3

[82] Pancheng Zhao, Peng Xu, Pengda Qin, Deng-Ping Fan, Zhicheng Zhang, Guoli Jia, Bowen Zhou, and Jufeng Yang. LAKE-RED: Camouflaged images generation by latent background knowledge retrieval-augmented diffusion. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–10, 2024. 3

[83] Xiaoqi Zhao, Youwei Pang, Wei Ji, Baicheng Sheng, Jiaming Zuo, Lihe Zhang, and Huchuan Lu. Spider: A unified framework for context-dependent concept segmentation. In *International Conference on Machine Learning*, pages 1–21, 2024. 3

[84] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. Towards diverse binary segmentation via a simple yet general gated network. *International Journal of Computer Vision*, 132:4157–4234, 2024. 3

[85] Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4504–4513, 2022. 3

[86] Hongwei Zhu, Peng Li, Haoran Xie, Xuefeng Yan, Dong Liang, Dapeng Chen, Mingqiang Wei, and Jing Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3608–3616, 2022. 3