CIRCOD: Co-saliency Inspired Referring Camouflaged Object Discovery

Avi Gupta[1], Koteswar Rao Jerripothula[2,1], Tammam Tillo[1]
[1]Indraprastha Institute of Information Technology Delhi (IIIT-Delhi)
[2]Indian Institute of Technology Kanpur (IIT Kanpur)
{avig,tammam}@iiitd.ac.in, kotesrj@iitk.ac.in

Abstract

Camouflaged object detection (COD), the task of identifying objects concealed within their surroundings, is often quite challenging due to the similarity that exists between the foreground and background. By incorporating an additional referring image where the target object is clearly visible, we can leverage the similarities between the two images to detect the camouflaged object. In this paper, we propose a novel problem setup: referring camouflaged object discovery (RCOD). In RCOD, segmentation occurs only when the object in the referring image is also present in the camouflaged image; otherwise, a blank mask is returned. This setup is particularly valuable when searching for specific camouflaged objects. Current COD methods are often generic, leading to numerous false positives in applications focused on specific objects. To address this, we introduce a new framework called Co-Saliency Inspired Referring Camouflaged Object Discovery (CIRCOD). Our approach consists of two main components: Co-Saliency-Aware Image Transformation (CAIT) and Co-Salient Object Discovery (CSOD). The CAIT module reduces the appearance and structural variations between the camouflaged and referring images, while the CSOD module utilizes the similarities between them to segment the camouflaged object, provided the images are semantically similar. Covering all semantic categories in current COD benchmark datasets, we collected over 1,000 referring images to validate our approach. Our extensive experiments demonstrate the effectiveness of our method and show that it achieves superior results compared to existing methods. Code is available at https://github.com/avigupta2798/CIRCOD/.

1. Introduction

Camouflaged object detection (COD) [13] is a crucial problem in computer vision, focusing on identifying objects concealed within their surroundings. This task has significant applications in wildlife monitoring, search and rescue operations, and military surveillance. Recent COD methods [1, 3, 22, 37, 44, 50, 65, 72] have improved detection accuracy, but they remain unreliable in out-of-distribution scenarios. Detecting camouflaged objects is challenging, even for humans. However, if the object is known, it becomes easier for them to detect it through careful examination of unclear or broken boundaries. Similarly, if such knowledge is fed through a referring image, it should be possible to handle out-of-distribution scenarios.

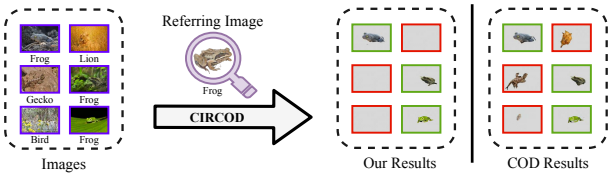

Figure 1. CIRCOD: A specialized approach for detecting camouflaged objects as specified by referring images, minimizing false positives in comparison to traditional COD [19] while searching for specific camouflaged objects.

To date, only [70] has used referring images for referring camouflaged object detection (Ref-COD), but its robustness is limited. It often detects objects even when no match exists between camouflaged and referring images, resulting in false positives. Even with referring-based large vision models [1, 22, 50], we observe a similar phenomenon, as revealed in our experiments (see Table 4). This highlights the unresolved problems of out-of-distribution COD and Ref-COD. We believe these challenges can be effectively addressed through co-saliency modeling [9, 14, 20, 21, 71], as one can then focus on highlighting common features between camouflaged and referring images, and ensure occurrence of segmentation only when there is a match, i.e., the two images are semantically similar.

Motivated by these challenges, we propose a framework called Co-Saliency Inspired Referring Camouflaged Object Discovery (CIRCOD). It segments a camouflaged object only when there is a match (see Fig. 1). Only those camouflaged images got segmented out that had frogs in them. However, existing COD methods are quite generic and are not tailored to detect specified camouflaged objects, lead-

ing to numerous false positives in applications focused on specific objects (see Fig. 1). The same issue persists in referring-based methods, as discussed above. To overcome this, we introduce a new problem setup: Referring Camouflaged Object Discovery (RCOD). In RCOD, segmentation occurs only when the object in the referring image matches the one in the camouflaged image; otherwise, a blank mask is returned.

We use the term "discovery" instead of "detection," as in Ref-COD [70], to emphasize that RCOD does not assume the presence of the specified object in the camouflaged image. RCOD occurs only when the objects in the two images match. This new setup necessitates robust image matching, and our CIRCOD framework ensures that. Our CIRCOD has applications in fields like medical imaging [8, 23], visual object search [43,53], military operations, aerial search and rescue [62], and wherever identifying specific objects is crucial (for e.g., searching for a particular animal concealed within its surrounding).

For co-saliency to be effective, objects must not only be similar but also salient. Therefore, we developed a new dataset named Ref-1K to serve as a pool of salient referring images. It contains more than 1,000 salient images, covering all categories of current COD benchmark datasets [13, 29, 39]. Additionally, we trained a Saliency Enhancement Network (SEN) to enhance the saliency of camouflaged images. SEN is part of our larger Co-Saliency-Aware Image Transformation (CAIT) module, which also helps in aligning the referring image with the saliency-enhanced camouflaged image to minimize any structural variations. Our Co-Salient Object Discovery (CSOD) module then uses a novel joint attention mechanism to make segmentation and similarity predictions. Together, the CAIT and CSOD modules form the core of our CIRCOD framework.

Our CIRCOD method achieves state-of-the-art (SOTA) performance in both Ref-COD (COD assisted by referring images) and RCOD tasks. While our primary goal is not to solve the general COD problem, our SEN can be further enhanced to achieve results comparable to SOTA methods for COD, serving as a valuable by-product of our work.

Our contributions are: (i) a new problem setup, RCOD, where segmentation occurs only when the referring and camouflaged images match; otherwise, a blank is returned; (ii) the novel CIRCOD framework to solve the RCOD problem; (iii) the Ref-1K dataset, covering all categories of current COD benchmark datasets; (iv) SOTA results in Ref-COD and RCOD settings, and competitive COD results.

2. Related Works

Camouflaged Object Detection: Detecting camouflaged objects is inherently challenging due to their close resemblance to the surrounding background, complicating segmentation and detection tasks [13, 30, 39]. Early works [6,51] focused on manually annotated, low-level features. The advent of deep learning and large datasets has led to significant advances, with early CNN-based methods [42, 47, 48, 63] extracting low-level features to identify camouflaged objects. However, their performance in complex, low-contrast scenes remained limited. Methods such as [15] introduced wavelet transformations to enhance edge reconstruction, while others [25, 44, 72] used zoom-in and zoom-out strategies to improve detection. Despite these advances, CNNs' limited receptive fields constrained their ability to model complex real-world environments. Transformer-based approaches [19, 45] addressed this limitation by capturing long-range dependencies, offering improved segmentation performance. Recent works like [1] demonstrated the adaptability of large vision foundational models [28] for COD tasks with minimal trainable parameters.

Referring-Based Segmentation: Referring segmentation involves segmenting objects in a query image based on guidance from text, images, or both. [18] pioneered this task using recurrent and convolutional neural networks to generate visual masks from linguistic queries. Later methods incorporated multi-level visual features [31], enabling better handling of complex contexts. Numerous approaches have since been developed for image and video segmentation using linguistic features [5, 7, 33, 34, 46, 66], often assuming that the target object exists and matches the query. To address cases where no matching object is present, generalized approaches [32, 60, 61] have been proposed, allowing greater flexibility for real-world applications. Recent work, such as [70], has also explored using image-based references to localize objects in query images when both belong to similar categories.

Our work bridges these two areas, focusing on segmenting camouflaged objects only when they match the object in the referring image.

3. Proposed Method

This section introduces our CIRCOD framework for solving the Referring Camouflaged Object Discovery (RCOD) problem, focusing on segmenting camouflaged objects only when they match the object in the referring image. Our approach leverages co-saliency to discover the camouflaged object with the guidance of a referring image. Unlike traditional camouflaged object segmentation methods, our framework segments the object only when the camouflaged image (denoted as I_c) shares the same semantic category as the referring image (denoted as I_r). Consequently, our proposed framework generates two outputs: the predicted segmentation mask (denoted as $\hat{m}_c$) and the semantic similarity prediction (denoted as $\hat{d}$). The corresponding ground-truth labels are denoted as m_c and d, respectively.

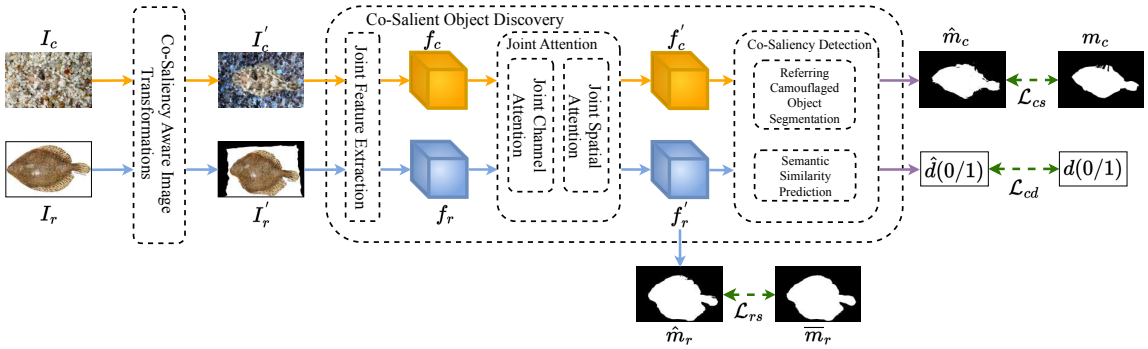

Figure 2. Overview of our CIRCOD framework: The camouflaged (I_c) and the referring (I_r) images are passed through the Co-saliency aware image transformations module, which enhances saliency of the camouflaged image and aligns the referring image to the camouflaged image. The transformed images $(I_c^{'}, I_r^{'})$ are sent to the co-salient object discovery module, where both branches are jointly processed by extracting the co-saliency features $(f_c^{'}, f_r^{'})$ to predict the required outputs, $\hat{m}_c$ and $\hat{d}$. m_c and d are the respective target labels. $\hat{m}_r$ and $\overline{m_r}$ are predicted and pseudo-labelled masks, respectively, for the aligned referring image. $\mathcal{L}_{cs}, \mathcal{L}_{rs}$, and $\mathcal{L}_{cd}$ are different losses used to optimize the network.

3.1. Overview

The overall architecture of our proposed framework is illustrated in Fig. 2, comprising two main components: (i) co-saliency aware image transformations; and (ii) co-salient object discovery.

In the first component, we pre-process the camouflaged and referring image to facilitate co-salient discovery by applying appropriate transformations. Since co-salient object discovery relies on saliency and minimal shape variation between paired images, we enhance the saliency of the camouflaged image and structurally align the referring image. This results in the transformed images, I_c' and I_r', which exhibit enhanced saliency and reduced shape variation, as shown in the figure. In the second component, the transformed images are processed jointly through feature extraction, attention mechanisms, and co-saliency detection (similarity computation), ultimately producing the desired outputs, $\hat{m}_c$ and $\hat{d}$.

3.2. Co-saliency Aware Image Transformations

We perform two key transformations: saliency enhancement and image alignment. Accurate boundary definition is crucial for co-saliency detection, particularly when addressing the challenge of distinguishing camouflaged objects. Therefore, enhancing the image is essential to improve the visibility of camouflaged objects against their background. Furthermore, co-saliency detection becomes increasingly challenging when significant shape variations are present. To address this, we align the referring image with the transformed (saliency-enhanced) camouflaged image. Detailed explanations of the saliency enhancement and image alignment transformations are provided below and visualized in Fig. 3.

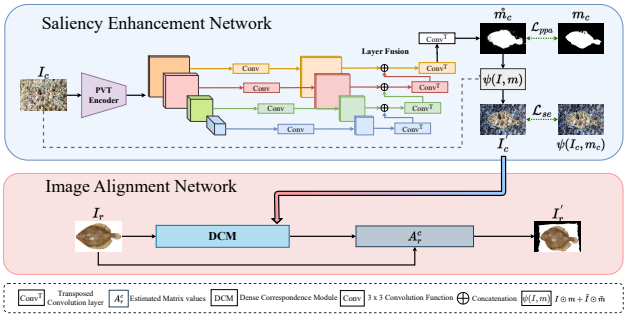

Figure 3. Co-saliency Aware Image Transformations: The first part is the proposed Saliency Enhancement Network (SEN). It consists of a Pyramid Vision Transformer (PVT) encoder and a Layer Fusion Module (LFM). The representations from the layers of PVT are passed through convolution blocks to get adjusted features. These features are aggregated using the LFM to get the required mask. In the Image Alignment Network (IAN), the referring image is aligned to the enhanced camouflaged image using a Dense Correspondence Module (DCM).

3.2.1 Saliency Enhancement Network

Since objects in camouflaged images often exhibit appearances similar to their background, changing the background's appearance significantly enhances the visibility of object boundaries, making them more distinguishable and detectable. This is accomplished using an object mask predicted using our saliency enhancement network (SEN). If m is the predicted mask of camouflaged image I, the transformation ψ (detailed below) is used to enhance the saliency of I:

$$\psi(I, m) = I \odot m + \tilde{I} \odot \tilde{m} \tag{1}$$

where $\tilde{I}$ and $\tilde{m}$ are negatives of I and m, respectively. A visualization of this transformation is shown in Fig. 4. This transformation enhances saliency in camouflaged images (except when the object's RGB pixel values are close to

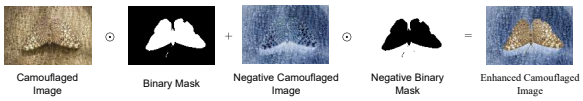

Figure 4. Enhancing saliency in camouflaged images: It involves the addition of two element-wise multiplication operations: one between the camouflaged image and its binary mask, and another between their negatives. This causes a significant change in the background, making it very different from the camouflaged object.

[128,128,128], which is obviously very rare). By applying this transformation on a camouflaged image (I_c) using the initially predicted object mask ($\mathring{m}_c$), we obtain $\psi(I_c, \mathring{m}_c)$. For simplicity, we refer to this transformed image as I'_c.

To obtain the initial predicted object mask, the camouflaged image $I_c \in \mathbb{R}^{H\times W\times 3}$ is initially passed through a pyramid vision transformer (PVT) encoder [57]. As shown in Fig. 3, the encoder extracts feature representations at multiple levels, denoted as $\mathcal{X}_i \in \mathbb{R}^{\frac{H}{2^{i+1}}\times\frac{W}{2^{i+1}}\times C_i}$, where C_i is the i^{th} element in $C \in \{64, 128, 320, 512\}$.

For decoding the feature representations, they are first passed through convolution function $Conv(\cdot) = BN(ReLU(Convolution(\cdot)))$ which resizes each representation to 64 channel size (using 3×3 kernel size). The resized representations are then passed through the Layer Fusion Module (LFM), aggregating them using a series of transposed convolution (of 4×4 kernel size) and concatenation steps, as shown in Fig. 3. Essentially, a low-resolution representation is upsampled and concatenated with the immediately higher one [8, 59]. The resultant representation is further upsampled to yield a binary mask ($\mathring{m}_c \in \mathbb{R}^{H\times W\times 1}$).

Using $\mathring{m}_c$, we can now obtain our saliency-enhanced camouflaged image I'_c through the transformation given in Eq. (1). We train this network using pixel position-aware loss ($\mathcal{L}_{ppa}$) [26] while comparing the generated object mask and the ground-truth object mask. We also have saliency enhancement loss ($\mathcal{L}_{se}$), which computes MSE (mean squared error) between saliency-enhanced images obtained via the generated object mask and the ground truth object mask.

$$\mathcal{L}_{se}(I_c, \mathring{m}_c, m_c) = MSE\big(\psi(I_c, \mathring{m}_c), \psi(I_c, m_c)\big) \quad (2)$$

$$\mathcal{L}_{ct}(\mathring{m}_c, m_c) = \mathcal{L}_{ppa}(\mathring{m}_c, m_c) + \mathcal{L}_{se}(I_c, \mathring{m}_c, m_c) \quad (3)$$

where L_{ct} denotes the final loss that needs to be minimized for transforming the camouflaged image.

It's interesting to note that SEN performs a task similar to that of traditional camouflaged object detection networks.

3.2.2 Image Alignment Network

We align the referring image I_r using a pre-trained dense correspondence model (DCM) from GLUNet [54] with I'_c serving as the reference. This approach minimizes the structural variation between the transformed camouflaged and referring images. Specifically, the DCM provides an estimated flow $\mathcal{A}_r^c$ between the two images, represented as a matrix of displacement vectors. This flow is then used to warp (align) the referring image, as formulated below:

$$\mathcal{A}_r^c = DCM(I_r, I'_c) \quad (4)$$

$$I'_r = \mathcal{A}_r^c(I_r) \quad (5)$$

where I'_r denotes the transformed referring image.

3.3. Co-salient Object Discovery

This section describes how our co-salient object discovery module takes the two transformed images and performs co-salient object discovery in the camouflaged image, i.e., the camouflaged image gets segmented only when the objects present in the two images match. The module extracts features, computes joint attention, and predicts desired outputs, $\hat{m}_c$ and $\hat{d}$, all while attempting to leverage co-saliency.

3.3.1 Joint Feature Extraction

We jointly process the transformed camouflaged and referring images to extract feature representations through a Siamese encoder with shared weights, utilizing PVT and LFM modules similar to the ones discussed in Sec. 3.2. These feature representations are now more reliable because we have already accounted for appearance and structural variations between the camouflaged and referring images. Additionally, using shared weights ensures that the feature representations are comparable, which is essential for similarity extraction to estimate co-saliency.

3.3.2 Joint Attention

Joint Channel Attention: Channels in the extracted feature volumes do not contribute equally to the task. To address this, we introduce a shared weight vector that can provide the weights for each channel, amplifying the significant ones and suppressing the less relevant ones. This vector highlights the common characteristics of the two branches. It is computed by obtaining spatial summaries of the feature volumes from the two branches using separate global average pooling (gap) and global maximum pooling (gmp). These pooled vectors are then passed through a joint fully connected layer, followed by addition and softmax operations, as illustrated below for a feature volume x:

$$SS(x) = \phi\Big(\mathcal{FC}\big(gap(x)\big) + \mathcal{FC}\Big(gmp(x)\Big)\Big) \quad (6)$$

where $\mathcal{FC}$ represents a fully connected layer, and ϕ denotes the softmax function. SS denotes the spatial summary.

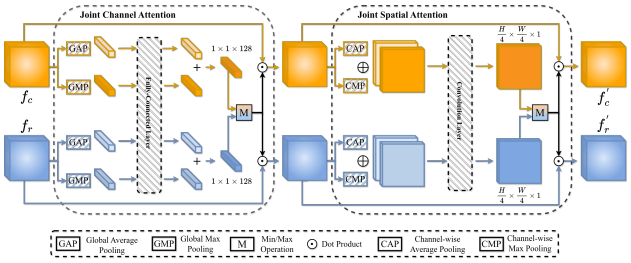

Figure 5. Joint attention module: The extracted feature volumes are first passed through joint channel attention to fetch spatial summary at each channel and then element-wise min-over-max operation is carried out across the two branches. The output vector is used weight vector to highlight relevant channels. Then, the updated volumes are further passed through the joint spatial attention to carry to similarly highlight relevant spatial locations now. (Best viewed in color).

Note that the weights involved in the generation of these summaries are shared across the two volumes. These summaries are then fused using an element-wise min-over-max ratio, resulting in a weight vector with higher weights for similar elements and lower weights for dissimilar ones. As a result, when this vector is multiplied by the two branches, similar channels get highlighted, and dissimilar ones get suppressed, resulting in channel attention. The entire process has been illustrated in Fig. 5.

Joint Spatial Attention: Similar to channels, some spatial locations are important, and some are not. Therefore, we need a spatial weight matrix that can provide the weights for each spatial, ensuring highlighting of important spatial locations get highlighted and suppression of less important ones. We do that by computing a channel summary for each spatial location in the two volumes and computing similarities between the two matrices formed, as we did in the case of channels earlier. We use channel-wise average pooling (cap) and channel-wise maximum pooling (cmp) to summarize the channels. The two summaries are then concatenated and passed through a convolutional layer followed by a sigmoid function, as formulated below for a feature volume x:

$$CS(x) = \sigma\Big(Conv\big(cap(x) \oplus cmp(x)\big)\Big), \tag{7}$$

where $Conv$ represents a single convolutional layer of 7×7 kernel size, and σ denotes the sigmoid function. CS denotes the channel-wise summary. Note that the weights are shared across the two branches while learning these summaries, which are then fused using element-with min-over-max operation to obtain the required weight matrix. Element-wise multiplication of that matrix with the two feature volume highlights highlights similar spatial locations and suppresses dissimilar ones, resulting in spatial attention. This process also has been illustrated in Fig. 5.

3.3.3 Co-Saliency Detection

Jointly attention-enhanced feature volumes, $(f_r^{'})$ and $(f_c^{'})$, are now exploited for co-saliency detection, which involves two tasks: referring camouflaged object segmentation and semantic similarity prediction.

Referring Camouflaged Object Segmentation: First, we select $(f_r^{'})$ and highlight its critical areas using self-attention. We basically compute the matrix multiplication of the feature volume with itself, scale it (value-wise), and then applying a softmax function to it. The resultant map is again multiplied with the original feature volume to emphasize essential features of the referring object, as detailed below:

$$f_r^{\gamma}(f_r^{'}) = \phi\left(\frac{f_r^{'} \otimes (f_r^{'})^T}{\xi}\right) \otimes f_r^{'}, \tag{8}$$

where $f_r^{\gamma}(f_r^{'})$ denotes spatial attention enhanced feature volume of the referring image and ξ denotes the scaling factor.

Taking inspiration from [8, 55], we now apply cross attention using (f_r^{γ}) as query(Q) and $(f_c^{'})$ as both key(K) and value (V). Specifically, we compute the matrix multiplication of the query with the key, scale it (value-wise), and apply the softmax function to obtain cross-attention. The resultant attention map is multiplied with value representations to obtain cross-attention enhanced feature volume $(\mathcal{CA}(f_c^{'}|f_r^{\gamma}))$, as detailed below:

$$\mathcal{CA}(f_c^{'}|f_r^{\gamma}) = \phi\left(\frac{f_r^{\gamma} \otimes (f_c^{'})^T}{\xi}\right) \otimes f_c^{'}, \tag{9}$$

which is then added to the original feature volume and passed through two transposed convolutional layers (represented as $\mathcal{C}onv_2^T$) to: (i) reduce the channel size to 1; and (ii) increase the spatial size to that of the original camouflaged image. After softmax, we obtain a co-saliency mask $\hat{m}_c$ for the camouflaged image, as computed below:

$$\hat{m}_c = \phi\left(\mathcal{C}onv_2^T\left(\mathcal{C}onv_1^T\left(f_c^{'} + \mathcal{CA}(f_c^{'}|f_r^{\gamma})\right)\right)\right) \tag{10}$$

where $\mathcal{C}onv_1^T$ and $\mathcal{C}onv_2^T$ are two 4×4 transposed convolution operations to first reduce the channel size from 128 to 8 and then from 8 to 1, respectively, while gradually increasing the spatial size to match the size of the original image for mask generation.

Semantic Similarity Prediction: Feature volumes $(f_c^{'}, f_r^{'}$ are also passed through a similarity prediction network to predict semantic similarity between the camouflaged image and the referring image. Specifically, both representations are globally average pooled to generate two

vectors of size 128. We then compute their absolute difference and pass through two linear layers: one with ReLU activation and another with sigmoid activation. This yields a probability value, which is used as the decision value $\hat{d}$ to indicate how similar are the two images, as computed below:

$$\hat{d} = \sigma\Big(\mathcal{FC}_2(\mathcal{FC}_1^{relu}\big(|gap(f_c^{'}) - gap(f_r^{'})|\big))\Big) \tag{11}$$

where $\mathcal{FC}_1^{relu}$ and $\mathcal{FC}_2$ denote the two fully connected layers just discussed.

3.4. Training Objectives

While training our CIRCOD model, we have two main objectives: (i) learning to detect semantic similarity and (ii) learning to generate an output mask to fulfill the requirements of RCOD setup, *i.e.*, a segmentation mask is generated for positive samples (referring and camouflaged images match) and a blank mask for negative samples (referring and camouflaged images don't match). We can generate such a mask by multiplying $\hat{d}$ and $\hat{m}_c$. These objectives are achieved through our similarity detection and camouflaged image segmentation losses, respectively. Additionally, we incorporate an auxiliary loss, the referring image segmentation loss, to further support the second objective and improve the model's overall performance.

Similarity Detection Loss: We use binary cross-entropy loss ($\mathcal{L}_{cd}$) to facilitate semantic similarity prediction, as detailed below:

$$\mathcal{L}_{cd}(\hat{d}, d) = d\log(\hat{d}) + (1-d)\log(1-\hat{d}) \tag{12}$$

where $\hat{d}$ and d are the predicted and ground truth values for semantic similarity.

Camouflaged Image Segmentation Loss. For optimizing the predicted mask, we use the sum of weighted cross-entropy loss ($\mathcal{L}_{ce}$) and weighted intersection-over-union loss ($\mathcal{L}_{iu}$) [26], as detailed below:

$$\begin{aligned}\mathcal{L}_{cs}(\hat{m}_c, m_c|\hat{d}, d) = \mathcal{L}_{ce}(\hat{d} * \hat{m}_c, d * m_c) + \\ d * \mathcal{L}_{iu}(\hat{d} * \hat{m}_c, d * m_c)\end{aligned} \tag{13}$$

Referring Image Segmentation Loss: Solving two related problems simultaneously can enhance their performance. Thus, we additionally try to predict the object mask of the aligned referring image. For the target, we generate a pseudo-label generated via SEN and IAN. More specifically, we use SEN to generate a saliency map for the original referring image and then align it using $\mathcal{A}_r^c$ of IAN. This pseudo label is denoted as $\overline{m}_r$. The referring segmentation loss ($\mathcal{L}_{rs}$) is defined as follows:

$$\begin{aligned}\mathcal{L}_{rs}(\hat{m}_r, \overline{m}_r|\hat{d}, d) = \mathcal{L}_{ce}(\hat{d} * \hat{m}_r, d * \overline{m}_r) + \\ d * \mathcal{L}_{iu}(\hat{d} * \hat{m}_r, d * \overline{m}_r)\end{aligned} \tag{14}$$

Note how we multiply the decision values (for predicted and target) with corresponding masks before computing the loss values $\mathcal{L}_{cs}$ and $\mathcal{L}_{rs}$. Our final loss is sum of all three losses: $\mathcal{L}_{cd}$, $\mathcal{L}_{cs}$ and $\mathcal{L}_{rs}$.

4. Experiments

4.1. Experimental Setup

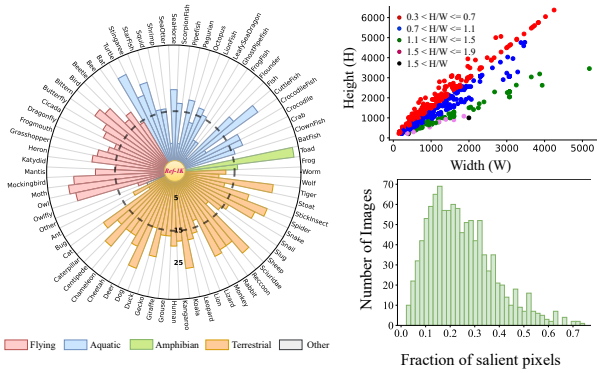

Figure 6. Detailed statistics of Ref-1K dataset: [Left] Taxonomic representation of proposed Ref-1K dataset. Each bar represents the log approximation of the number of images within each category. [Top-Right] Distribution of image resolutions. [Bottom-Right] Histogram of salient pixel fractions per image.

Datasets: We consider popular COD datasets: CAMO [29], COD10K [13], and NC4K [39]. CAMO comprises 1250 camouflaged images, COD10K contains 5066 camouflaged images, and NC4K includes 4121 camouflaged images.

While the above datasets provide camouflaged images, we also need referring images for Ref-COD and RCOD settings. Recently, adataset named R2C7K [70] was released, which comprises of 5015 camouflaged and 1600 referring images. However, we observed that these 1600 images are not comprehensive.

To evaluate comprehensively Ref-COD and RCOD settings, we introduce a new referring image dataset, which we call ***Ref-1K***. It consists of 1,078 referring images (with 80:20 split for training and testing) divided among 5 classes and 76 categories. These images were collected from Flickr, Unsplash, and other public-domain repositories. Fig. 6 provides a statistical analysis of our proposed dataset, including a taxonomic representation of categories, image resolution distribution, and histogram of salient pixel fractions per image. It can be observed that our dataset covers every category, a good range of image resolutions and saliency pixel fractions. We ensured that the categories represent union of all categories present in existing camouflaged benchmark datasets. Furthermore, we structured the dataset to proportionally align the number of images per category with the representation in the COD datasets discussed.

Training Setting: Our experiments have been conducted under three training settings: (i) COD, where, following [13,39,44], SEN is trained using the combined training sets of COD10K and CAMO; (ii) Ref-COD, where CIRCOD is trained on training sets of R2C7K (involving both reference and camouflaged images); and (iii) RCOD, in addition to the training set used for camouflaged images, the training set of our Ref-1K dataset in used for referring images. In the RCOD setup, there are two types of training samples: (i) positive (referring and camouflaged images match); and (ii) negative (referring and camouflaged images don't match). For each sample in the training set of camouflaged images, two positive and two negative samples are generated by drawing appropriate referring images from the training set of Ref-1K. Similarly, during testing/inference, we create one positive and one negative sample for each image in the test set of a COD dataset. For any camouflaged image, the average accuracy of its positive and negative samples is calculated.

Evaluation Metrics: We adopt five standard metrics for evaluation: mean absolute error ($\mathcal{M}$), weighted F-measure ($\mathcal{F}_\beta^w$) [41], S-measure ($\mathcal{S}_m$) [10], mean E-measure ($\mathcal{E}_\phi^m$) [11], and adaptive E-measure ($\alpha\mathcal{E}$) [11], which can be directly used to evaluate positive samples. For negative samples, the binary masks (of both predicted and ground truth) are inversed before applying these metrics. That's required because they are not applicable on blank masks. Additionally, we calculate the decision accuracy ($\mathcal{A}$) by comparing the predicted decision value against the groundtruth decision value.

Implementation Details: We implemented our model using PyTorch framework on a single workstation of NVIDIA A100 GPU. The pre-trained PVT-V2 [57] model was adopted as the backbone. All the images were resized to 512×512 while maintaining their aspect ratio. We used a batch size of 8 and applied random augmentations, including flipping, mirroring, and rotations. The network parameters were optimized using the AdamW [36] optimizer with an initial learning rate of 5e-5 and a weight decay of 0.0001.

4.2. COD Results

We compare our proposed Saliency Enhanced Network (SEN) with current state-of-the-art COD methods, as summarized in Table 1. Note that "P4" denotes usage of PVT-B4, and "f" denotes fine-tuned version. The alrger SEN network and finetuning was required to get comparable results with SOTA, as our original SEN is just an intermediate step required to enhance saliency of the camouflaged image, and is trained for limited epochs. We can see superior performance compared to existing approaches [17, 19, 38, 40, 52, 73] and achieves comparable results with others [2, 4, 16, 56]. It's is essential to emphasize that SEN's primary objective is to produce saliency enhanced images as a by-product rather than being explicitly optimized for COD tasks. Qualitative results have been shown in Fig. 7.

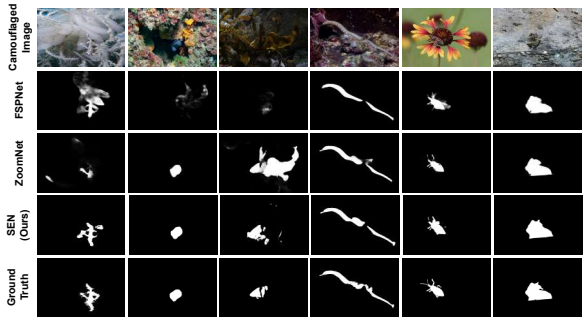

Figure 7. Visual comparison of our proposed SEN with SOTA COD methods.

Table 1. Quantitative comparison of proposed SEN with fifteen methods on three benchmark datasets in a COD setting. ↑ / ↓ denotes the larger/smaller is better. "-" denotes results are not available. "P4" denotes PVT-B4, "f" denotes fine-tuned version. The best three results are marked in **red**, blue, and violet respectively.

Models	CAMO				COD10K				NC4K			
	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
ZoomNet [44]	0.820	0.878	0.752	0.066	0.838	0.888	0.729	0.029	0.853	0.896	0.784	0.043
FSPNet [19]	0.856	0.899	0.799	0.050	0.851	0.895	0.735	0.026	0.879	0.915	0.816	0.035
GenSAM [17]	0.719	0.775	-	0.113	0.775	0.838	-	0.067	-	-	-	-
DINet [73]	0.821	0.874	-	0.068	0.832	0.903	-	0.031	0.856	0.909	-	0.043
UEDG [38]	0.868	0.922	0.819	0.048	0.858	0.924	0.766	0.025	0.881	0.928	0.829	0.035
CINet [40]	0.847	0.899	0.794	0.055	0.841	0.914	0.744	0.028	0.868	0.924	0.815	0.037
GIFE [52]	0.817	0.873	-	0.070	0.832	0.899	-	0.032	0.856	0.908	-	0.044
DCNet [67]	0.870	0.922	0.831	0.050	0.873	0.934	0.810	0.022	-	-	-	-
SAM-Adapter [1]	0.847	0.873	0.765	0.070	0.883	0.918	0.801	0.025	-	-	-	-
PGT-P4 [56]	**0.882**	0.935	**0.884**	**0.042**	0.879	0.938	0.801	0.021	0.896	0.942	0.854	**0.029**
MLKG [4]	0.828	-	0.744	0.075	**0.910**	-	**0.829**	**0.019**	**0.900**	-	0.833	0.036
CamoFocus-P4 [27]	0.873	-	0.842	0.043	0.873	-	0.802	0.021	0.889	-	0.853	0.030
CamoDiffusion [2]	0.880	**0.939**	0.855	**0.042**	0.883	0.942	0.819	**0.019**	0.894	0.941	**0.859**	**0.029**
ICEG [16]	0.871	0.931	-	**0.042**	0.862	0.934	-	0.023	0.883	0.937	-	0.033
CamoFormer-P4 [64]	0.878	-	0.839	0.044	0.872	-	0.793	0.022	0.893	-	0.850	0.030
SEN (Ours)	0.824	0.894	0.772	0.063	0.835	0.916	0.741	0.030	0.858	0.920	0.808	0.040
SEN$_f$(Ours)	0.857	0.924	0.824	0.050	0.863	0.941	0.789	0.023	0.878	0.939	0.840	0.033
SEN$_f^{P4}$(Ours)	0.867	0.933	0.842	0.045	0.873	**0.943**	0.804	0.021	0.887	**0.943**	0.854	0.030

Table 2. Quantitative comparison of proposed SEN [Left] under COD setting and CIRCOD [Right] under Ref-COD setting on R2C7K. The results of other methods were taken from [70].

Models	$\mathcal{S}_m\uparrow$	$\alpha\mathcal{E}\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
PFNet [42]	0.791	0.876	0.651	0.040
PreyNet [68]	0.806	0.890	0.690	0.034
SINetV2 [12]	0.813	0.874	0.678	0.036
BSANet [74]	0.818	0.893	0.702	0.034
BGNet [49]	0.818	0.901	0.679	0.036
ZoomNet [44]	0.813	0.884	0.688	0.032
DGNet [24]	0.816	0.883	0.684	0.034
R2CNet [70]	0.772	0.847	0.604	0.044
SEN (Ours)	**0.828**	**0.910**	**0.729**	**0.030**

Models	$\mathcal{S}_m\uparrow$	$\alpha\mathcal{E}\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
PFNet-Ref	0.811	0.885	0.687	0.036
PreyNet-Ref	0.817	0.900	0.704	0.032
SINetV2-Ref	0.823	0.888	0.700	0.033
BSANet-Ref	0.830	0.912	0.727	0.030
BGNet-Ref	0.840	0.909	0.738	0.029
ZoomNet-Ref	0.834	0.886	0.720	0.029
DGNet-Ref	0.821	0.891	0.696	0.032
R2CNet-Ref	0.805	0.879	0.669	0.036
CIRCOD (Ours)	**0.848**	**0.918**	**0.756**	**0.026**

Table 2 [left] also reports additional COD results obtained on R2C7K dataset, and, here, our SEN could outperform all other methods.

4.3. Ref-COD Results

Under Ref-COD setting, we compare the performance of CIRCOD with existing methods in Table 2 [right]. Notably, our method achieved significant improvements over R2CNet [70] with the gains of 5.34%, 4.44%, 13.01%, 27.78% with referring supervision in terms of $\mathcal{S}_m$, $\alpha\mathcal{E}$, $\mathcal{F}_\beta^w$, and $\mathcal{M}$ metrics, respectively. This demonstrates both our approaches outperform previous methods. Furthermore, by incorporating referring supervision, we can see CIRCOD's shows better performance over SEN by 2.42%, 0.88%, 3.70%, and 13.33% margins, emphasizing the role of the referring component. Similarly, all prior referring-based methods show notable improvements compared to their respective COD baselines, as summarized in Table 2. We also show visual comparisons in Fig. 8.

We conducted additional experiments to draw comparisons in out-of-distribution scenarios using R2C7K (only camouflaged images part), NC4K, and CAMO as camouflaged datasets and R2C7K-Ref as the referring dataset. Here, we trained CIRCOD and R2CNet on three classes and tested on the remaining one to assess robustness of the two approaches. Test results have been reported in Table 3. CIRCOD achieved significant improvements over R2CNet.

Table 3. Quantitative comparison under Ref-COD setting in terms of results obtained through cross-domain analysis. We train on 3/4 classes and test on the remaining one. The best results are highlighted in **bold**.

Test Domain	Model	R2C7K				NC4K				CAMO			
		$\mathcal{S}_m\uparrow$	$\alpha\mathcal{E}\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\alpha\mathcal{E}\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\alpha\mathcal{E}\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
Terrestrial	R2CNet	0.712	0.803	0.507	0.052	0.769	0.850	0.647	0.069	0.597	0.702	0.411	0.156
	Ours	**0.795**	**0.900**	**0.686**	**0.030**	**0.834**	**0.909**	**0.784**	**0.044**	**0.672**	**0.718**	**0.560**	**0.122**
Flying	R2CNet	0.779	0.851	0.606	0.040	0.801	0.877	0.690	0.057	0.772	0.861	0.643	0.066
	Ours	**0.849**	**0.905**	**0.735**	**0.025**	**0.872**	**0.925**	**0.809**	**0.033**	**0.870**	**0.928**	**0.804**	**0.037**
Aquatic	R2CNet	0.717	0.809	0.560	0.081	0.727	0.808	0.614	0.102	0.752	0.820	0.649	0.098
	Ours	**0.821**	**0.897**	**0.729**	**0.045**	**0.835**	**0.888**	**0.770**	**0.061**	**0.838**	**0.895**	**0.780**	**0.064**
Amphibian	R2CNet	0.810	0.860	0.665	0.045	0.840	0.893	0.727	0.045	0.743	0.862	0.582	0.078
	Ours	**0.855**	**0.911**	**0.748**	**0.032**	**0.875**	**0.919**	**0.795**	**0.032**	**0.812**	**0.899**	**0.721**	**0.055**

Table 4. Quantitative comparison of CIRCOD with large generic referring-based segmentation methods under RCOD setting. The best results are highlighted in **bold**.

Models	CAMO				COD10K				NC4K			
	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$
SegGPT [58]	0.734	0.775	0.764	0.188	0.758	0.787	0.736	0.150	0.765	0.796	0.652	0.121
Per-SAM [69]	0.709	0.732	0.703	0.205	0.698	0.732	0.653	0.199	0.724	0.750	0.715	0.199
Per-SAM (fine-tuned) [69]	0.705	0.742	0.699	0.184	0.699	0.730	0.654	0.202	0.718	0.745	0.710	0.206
Matcher [35]	0.628	0.670	0.569	0.248	0.605	0.637	0.511	0.276	0.680	0.717	0.619	0.217
CIRCOD (Ours)	**0.867**	**0.887**	**0.833**	**0.056**	**0.875**	**0.890**	**0.806**	**0.027**	**0.869**	**0.875**	**0.817**	**0.045**

4.4. RCOD Results

Under RCOD setting, we compared CIRCOD's performance against some of the large generic referring-based image segmentation methods [35, 58, 69], where an additional image is used to perform segmentation. Interestingly, CIRCOD surpassed them by a significant margin (see Table 4), suggesting their biasness towards generic objects.

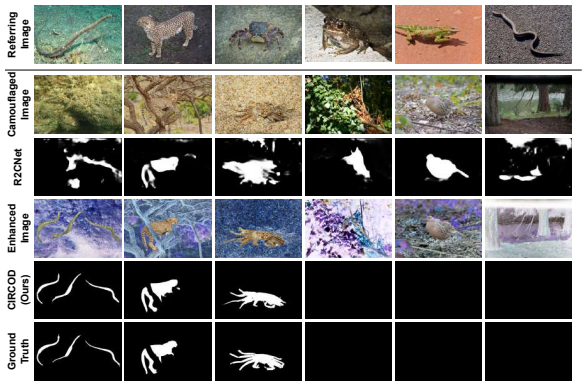

Figure 8. Visual comparison of our proposed CIRCOD with R2CNet under RCOD setting.

Table 5. Ablation Study under RCOD setting on COD10K and Ref-1K datasets. "SEN" denotes Saliency Enhancement Network, "IAN" denotes Image Alignment Network, "JCA" denotes Joint Channel Attention, "JSA" denotes Joint Spatial Attention, "RCOS" denotes Referring Camouflaged Object Segmentation, "$\mathcal{L}_{rs}$" denotes Referring Segmentation Loss, and "$\hat{d}$" denotes semantic similarity detection value.

SEN	IAN	JCA	JSA	RCOS	$\mathcal{L}_{rs}$	$\hat{d}$	$\mathcal{S}_m\uparrow$	$\mathcal{E}_\phi^m\uparrow$	$\mathcal{F}_\beta^w\uparrow$	$\mathcal{M}\downarrow$	$\mathcal{A}\uparrow$
	-	✓	✓	✓	-	✓	0.848	0.864	0.765	0.036	75.0
✓		✓	✓	✓	✓	✓	0.862	0.880	0.789	0.031	78.5
✓	✓		✓	✓	✓	✓	0.859	0.861	0.770	0.031	77.5
✓	✓	✓		✓	✓	✓	0.856	0.856	0.764	0.030	71.5
✓	✓	✓	✓		✓	✓	0.760	0.724	0.623	0.051	60.0
✓	✓	✓	✓	✓		✓	0.850	0.846	0.757	0.029	77.2
✓	✓	✓	✓	✓	✓		0.835	0.865	0.775	0.053	-
✓	✓	✓	✓	✓	✓	✓	**0.875**	**0.890**	**0.806**	**0.027**	**80.7**

4.5. Ablation Studies

Here, under RCOD setting, we conduct ablation studies, as summarized in Table 5. It involved removing specific sub-modules to study their importance in our entire framework. These studies were conducted on COD10K using our proposed Ref-1K dataset. It can be observed that removing any of our sub-modules results in a deterioration of performance, highlighting their importance.

Conclusion

We presented a novel task called referring camouflaged object discovery (RCOD), where the camouflaged object is segmented out only when it matches with that in the referring image; otherwise, a blank mask is expected. We accomplish it using our co-saliency-inspired referring camouflaged object discovery (CIRCOD) framework, which leveraged co-saliency-based ideas at every stage. Our proposed method not only performed well under RCOD setting but also under COD and Ref-COD settings.

References

[1] Tianrun Chen, Lanyun Zhu, Chaotao Ding, Runlong Cao, Yan Wang, Shangzhan Zhang, Zejian Li, Lingyun Sun, Ying Zang, and Papa Mao. Sam-adapter: Adapting segment anything in underperformed scenes. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023 - Workshops, Paris, France, October 2-6, 2023*, pages 3359–3367. IEEE, 2023. 1, 2, 7

[2] Zhongxi Chen, Ke Sun, and Xianming Lin. Camodiffusion: Camouflaged object detection via conditional diffusion models. In Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 1272–1280. AAAI Press, 2024. 7

[3] Zhongxi Chen, Ke Sun, Xianming Lin, and Rongrong Ji. Camodiffusion: Camouflaged object detection via conditional diffusion models. *arXiv preprint arXiv:2305.17932*, 2023. 1

[4] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection. *CoRR*, abs/2311.17122, 2023. 7

[5] Shupeng Cheng, Ge-Peng Ji, Pengda Qin, Deng-Ping Fan, Bowen Zhou, and Peng Xu. Large model based referring camouflaged object detection, 2023. 2

[6] H.B. Cott. *Adaptive Coloration in Animals*. Methuen & Company, Limited, 1940. 2

[7] Henghui Ding, Chang Liu, Shuting He, Xudong Jiang, and Chen Change Loy. MeViS: A large-scale benchmark for video segmentation with motion expressions. In *ICCV*, 2023. 2

[8] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramid vision transformers. *CoRR*, abs/2108.06932, 2021. 2, 4, 5

[9] Deng-Ping Fan, Tengpeng Li, Zheng Lin, Ge-Peng Ji, Dingwen Zhang, Ming-Ming Cheng, Huazhu Fu, and Jianbing Shen. Re-thinking co-salient object detection. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(8):4339–4354, 2022. 1

[10] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Oct 2017. 7

[11] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, pages 698–704, 7 2018. 7

[12] 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. 7

[13] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1, 2, 6, 7

[14] Qi Fan, Deng-Ping Fan, Huazhu Fu, Chi-Keung Tang, Ling Shao, and Yu-Wing Tai. Group collaborative learning for co-salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 12288–12298. Computer Vision Foundation / IEEE, 2021. 1

[15] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22046–22055, June 2023. 2

[16] Chunming He, Kai Li, Yachao Zhang, Yulun Zhang, Chenyu You, Zhenhua Guo, Xiu Li, Martin Danelljan, and Fisher Yu. Strategic preys make acute predators: Enhancing camouflaged object detectors by generating camouflaged objects. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. 7

[17] 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 Michael J. Wooldridge, Jennifer G. Dy, and Sriraam Natarajan, editors, *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024, Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence, IAAI 2024, Fourteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2014, February 20-27, 2024, Vancouver, Canada*, pages 12511–12518. AAAI Press, 2024. 7

[18] Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2016. 2

[19] 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/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 1, 2, 7

[20] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Quality-guided fusion-based co-saliency estimation for image co-segmentation and colocalization. *IEEE Trans. Multim.*, 20(9):2466–2477, 2018. 1

[21] Koteswar Rao Jerripothula, Jianfei Cai, and Junsong Yuan. Efficient video object co-localization with co-saliency activated tracklets. *IEEE Trans. Circuits Syst. Video Technol.*, 29(3):744–755, 2019. 1

[22] Ge-Peng Ji, Deng-Ping Fan, Peng Xu, Bowen Zhou, Ming-Ming Cheng, and Luc Van Gool. SAM struggles in concealed scenes - empirical study on "segment anything". *Sci. China Inf. Sci.*, 66(12), 2023. 1

[23] Ge-Peng Ji, Guobao Xiao, Yu-Cheng Chou, Deng-Ping Fan, Kai Zhao, Geng Chen, and Luc Van Gool. Video polyp segmentation: A deep learning perspective. *Int. J. Autom. Comput.*, 19(6):531–549, 2022. 2

[24] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learn-

ing for efficient camouflaged object detection. *Machine Intelligence Research*, 20:92–108, 2023. 7

[25] 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/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 4703–4712. IEEE, 2022. 2

[26] Qingming Huang Jun Wei, Shuhui Wang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2020. 4, 6

[27] Abbas Khan, Mustaqeem Khan, Wail Gueaieb, Abdulmotaleb El-Saddik, Giulia De Masi, and Fakhri Karray. Camofocus: Enhancing camouflage object detection with split-feature focal modulation and context refinement. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2024, Waikoloa, HI, USA, January 3-8, 2024*, pages 1423–1432. IEEE, 2024. 7

[28] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 3992–4003. IEEE, 2023. 2

[29] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Journal of Computer Vision and Image Understanding*, 184:45–56, 2019. 2, 6

[30] Aixuan Li, Jing Zhang, Yunqiu Lyu, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2

[31] Ruiyu Li, Kaican Li, Yi-Chun Kuo, Michelle Shu, Xiaojuan Qi, Xiaoyong Shen, and Jiaya Jia. Referring image segmentation via recurrent refinement networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5745–5753, 2018. 2

[32] Chang Liu, Henghui Ding, and Xudong Jiang. GRES: Generalized referring expression segmentation. In *CVPR*, 2023. 2

[33] Chang Liu, Henghui Ding, Yulun Zhang, and Xudong Jiang. Multi-modal mutual attention and iterative interaction for referring image segmentation. *IEEE Transactions on Image Processing*, 32:3054–3065, 2023. 2

[34] Jiang Liu, Hui Ding, Zhaowei Cai, Yuting Zhang, Ravi Kumar Satzoda, Vijay Mahadevan, and R Manmatha. Polyformer: Referring image segmentation as sequential polygon generation. In *CVPR*, 2023. 2

[35] Yang Liu, Muzhi Zhu, Hengtao Li, Hao Chen, Xinlong Wang, and Chunhua Shen. Matcher: Segment anything with one shot using all-purpose feature matching. *CoRR*, abs/2305.13310, 2023. 8

[36] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2017. 7

[37] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Nick Barnes, and Deng-Ping Fan. Toward deeper understanding of camouflaged object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 33(7):3462–3476, 2023. 1

[38] Yixuan Lyu, Hong Zhang, Yan Li, Hanyang Liu, Yifan Yang, and Ding Yuan. Uedg:uncertainty-edge dual guided camouflage object detection. *IEEE Trans. Multim.*, 26:4050–4060, 2024. 7

[39] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 2, 6, 7

[40] Ming Ma and Bangyong Sun. A cross-level interaction network based on scale-aware augmentation for camouflaged object detection. *IEEE Trans. Emerg. Top. Comput. Intell.*, 8(1):69–81, 2024. 7

[41] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014. 7

[42] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8772–8781, June 2021. 2, 7

[43] Thao Nguyen, Vladislav Hrosinkov, Eric Rosen, and Stefanie Tellex. Language-conditioned observation models for visual object search. In *IROS*, pages 10894–10901, 2023. 2

[44] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2160–2170, June 2022. 1, 2, 7

[45] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *European conference on computer vision*. Springer, 2022. 2

[46] Mengxue Qu, Yu Wu, Yunchao Wei, Wu Liu, Xiaodan Liang, and Yao Zhao. Learning to segment every referring object point by point. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3021–3030, 2023. 2

[47] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *IJCAI*, pages 1025–1031, 2021. 2

[48] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1025–1031. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 2

[49] Yujia Sun, Shuo Wang, Chenglizhao Chen, and Tian-Zhu Xiang. Boundary-guided camouflaged object detection. In *IJCAI*, pages 1335–1341, 2022. 7

[50] Lv Tang, Haoke Xiao, and Bo Li. Can SAM segment anything? when SAM meets camouflaged object detection. *CoRR*, abs/2304.04709, 2023. 1

[51] G.H. Thayer and A.H. Thayer. *Concealing-coloration in the Animal Kingdom: An Exposition of the Laws of Disguise Through Color and Pattern: Being a Summary of Abbott H. Thayer's Discoveries*. Macmillan Company, 1909. 2

[52] Jinghui Tong, Yaqiu Bi, Cong Zhang, Hongbo Bi, and Ye Yuan. Local to global purification strategy to realize collaborative camouflaged object detection. *Comput. Vis. Image Underst.*, 241:103932, 2024. 7

[53] Xuan-The Tran, Thomas Tien-Thong Do, and Chin-Teng Lin. Early detection of human decision-making in concealed object visual searching tasks: An eeg-bilstm study. In *45th Annual International Conference of the IEEE Engineering in Medicine & Biology Society, EMBC 2023, Sydney, Australia, July 24-27, 2023*, pages 1–4. IEEE, 2023. 2

[54] Prune Truong, Martin Danelljan, and Radu Timofte. GLU-Net: Global-local universal network for dense flow and correspondences. In *2020 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2020*, 2020. 4

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. 5

[56] Rui Wang, Caijuan Shi, Changyu Duan, Weixiang Gao, Hongli Zhu, Yunchao Wei, and Meiqin Liu. Camouflaged object segmentation with prior via two-stage training. *Comput. Vis. Image Underst.*, 246:104061, 2024. 7

[57] Wenhai Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pvt v2: Improved baselines with pyramid vision transformer. *Computational Visual Media*, 8(3):415–424, 2022. 4, 7

[58] Xinlong Wang, Xiaosong Zhang, Yue Cao, Wen Wang, Chunhua Shen, and Tiejun Huang. Seggpt: Towards segmenting everything in context. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 1130–1140, 2023. 8

[59] Yukang Wang, Yongchao Xu, Stavros Tsogkas, Xiang Bai, Sven Dickinson, and Kaleem Siddiqi. Deepflux for skeletons in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5287–5296, 2019. 4

[60] Jianzong Wu, Xiangtai Li, Xia Li, Henghui Ding, Yunhai Tong, and Dacheng Tao. Towards Robust Referring Image Segmentation, Sept. 2022. arXiv:2209.09554 [cs]. 2

[61] Yixuan Wu, Zhao Zhang, Xie Chi, Feng Zhu, and Rui Zhao. Advancing Referring Expression Segmentation Beyond Single Image, May 2023. arXiv:2305.12452 [cs]. 2

[62] Jianhao Xu, Xiangtao Fan, Hongdeng Jian, Chen Xu, Weijia Bei, Qifeng Ge, and Teng Zhao. Yoloow: A spatial scale adaptive real-time object detection neural network for open water search and rescue from UAV aerial imagery. *IEEE Trans. Geosci. Remote. Sens.*, 62:1–15, 2024. 2

[63] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 2

[64] 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*, pages 1–14, 2024. 7

[65] Bowen Yin, Xuying Zhang, Qibin Hou, Bo-Yuan Sun, Deng-Ping Fan, and Luc Van Gool. Camoformer: Masked separable attention for camouflaged object detection. *arXiv preprint arXiv:2212.06570*, 2022. 1

[66] S. Yu, P. Seo, and J. Son. Zero-shot referring image segmentation with global-local context features. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19456–19465, Los Alamitos, CA, USA, jun 2023. IEEE Computer Society. 2

[67] Guanghui Yue, Houlu Xiao, Hai Xie, Tianwei Zhou, Wei Zhou, Weiqing Yan, Baoquan Zhao, Tianfu Wang, and Qiuping Jiang. Dual-constraint coarse-to-fine network for camouflaged object detection. *IEEE Trans. Circuits Syst. Video Technol.*, 34(5):3286–3298, 2024. 7

[68] Miao Zhang, Shuang Xu, Yongri Piao, Dongxiang Shi, Shusen Lin, and Huchuan Lu. Preynet: Preying on camouflaged objects. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 5323–5332, New York, NY, USA, 2022. Association for Computing Machinery. 7

[69] Renrui Zhang, Zhengkai Jiang, Ziyu Guo, Shilin Yan, Junting Pan, Hao Dong, Peng Gao, and Hongsheng Li. Personalize segment anything model with one shot. *arXiv preprint arXiv:2305.03048*, 2023. 8

[70] Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring Camouflaged Object Detection, July 2023. arXiv:2306.07532 [cs]. 1, 2, 6, 7, 8

[71] Zhao Zhang, Wenda Jin, Jun Xu, and Ming-Ming Cheng. Gradient-induced co-saliency detection. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XII*, volume 12357 of *Lecture Notes in Computer Science*, pages 455–472. Springer, 2020. 1

[72] Dehua Zheng, Xiaochen Zheng, Laurence T. Yang, Yuan Gao, Chenlu Zhu, and Yiheng Ruan. Mffn: Multi-view feature fusion network for camouflaged object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 6232–6242, January 2023. 1, 2

[73] Xiaofei Zhou, Zhicong Wu, and Runmin Cong. Decoupling and integration network for camouflaged object detection. *IEEE Trans. Multim.*, 26:7114–7129, 2024. 7

[74] 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. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36:3608–3616, 06 2022. 7