CAReFuseNet: Cross Attention Fusion Network for Referring Camouflaged Object Detection

Anonymous authors
Paper under double-blind review

Abstract

The Referring Camouflaged Object Detection (Ref-COD) task aims to generate a binary segmentation mask to detect camouflaged objects of a specified category in an image, guided by reference image(s) containing salient example(s) of the target object of the same category. With only a few methods (e.g., R2CNet and UAT) proposed to date, Ref-COD remains challenging due to the similarity of camouflaged objects to their backgrounds and substantial feature gaps with salient references. At the same time, recent state-of-the-art approaches often rely on heavy transformer-based encoder–decoder stacks or large frozen vision backbones, resulting in substantial parameter footprints that hinder efficient deployment. This work proposes 'CAReFuseNet'[1], a novel framework featuring a cross-attention based reference feature fusion module that effectively extracts reference-conditioned feature representations from camouflaged images while targeting parameter efficiency. The proposed CAReFuse module leverages global interactions between reference and camouflaged image features via cross-attention, but constrains all fusion and decoding operations to a lower dimensional feature space and employs a lightweight convolutional decoder. Combined with a frozen Ref-Image Encoder, this design yields a compact Ref-COD model without sacrificing accuracy. Extensive experiments on the R2C7K dataset show that our method surpasses state-of-the-art, while using significantly fewer parameters. Further evaluations across multiple backbone architectures, including Swin Transformer, ConvNeXt, EfficientNet, and ResNet, demonstrate that the proposed reference feature fusion module provides a general and parameter-efficient building block for the referring camouflaged object detection task.

1 Introduction

Camouflaged Object Detection (COD) (Fan et al., 2020a), which aims to identify the objects that are seamlessly blended into their background, is a challenging yet useful computer vision task. The intrinsic similarities between the target object and the background, such as indistinguishable texture and ambiguous object boundaries, make COD a difficult task for computer vision systems. Owing to its real-world applications ranging from medical image segmentation (e.g., polyp segmentation (Fan et al., 2020b), lung infection segmentation (Fan et al., 2020c; Wu et al., 2021)) to video surveillance (e.g., camouflaged person or object detection), the COD task is gaining increased attention in the computer vision community (Lei et al., 2025; Le et al., 2025; Ren et al., 2025; Chen et al., 2025). One novel task setting of the COD problem is the Referring Camouflaged Object Detection (Ref-COD) (Zhang et al., 2025), where a reference image (or a set of reference images) containing a salient example of the target object is used to guide the detection of the camouflaged object in another image. Similar to Referring Image Segmentation (Hu et al., 2016; Lee et al., 2025), this task setting proves invaluable, especially in scenarios where one knows in advance what target object category one is looking for in the camouflaged images, as well as when the object is uncommon. However, there can be variation in pose, appearance, shape and size between the salient object instance in the reference images and the camouflaged objects in the input camouflaged image (see Figure 1). Zhang et al. (2025), who introduced the Ref-COD task, have proposed a dual-branch framework, dubbed R2CNet,

[1]Code is available in supplementary material.

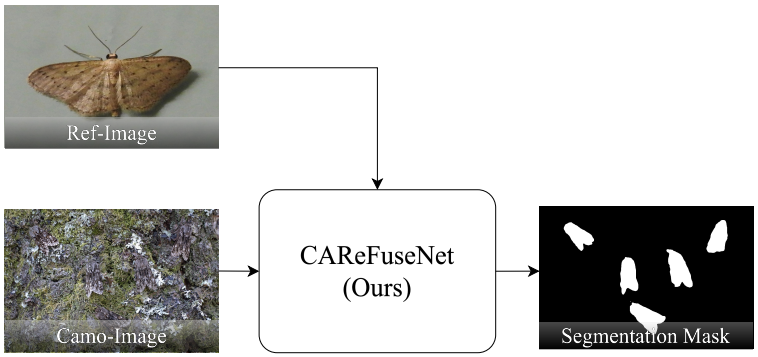

Figure 1: Despite variation in pose, appearance, shape and size between reference and target object instances, our CAReFuseNet segments camouflaged objects in the input image under the guidance of the salient reference image.

to perform the Ref-COD task by fusing the features extracted from a set of reference images with the feature representations of the input camouflaged image.

Following R2CNet, only a few other methods(e.g., UAT (Wu et al., 2025) and CIRCOD (Gupta et al., 2025)) emerged for the Ref-COD task. While UAT advanced the state of the art, it relies on heavy transformer-based encoder–decoder stacks. For instance, UAT employs a relatively deep transformer encoder–decoder architecture for reference feature integration and uncertainty modeling. Such designs yield high performance but come with substantial trainable and total parameter footprints, which can hinder deployment in resource-constrained environments or latency-sensitive applications. This exposes an important gap: there is a need for Ref-COD frameworks that retain competitive performance while being substantially more parameter-efficient.

Towards addressing this gap, this paper presents a novel framework, dubbed 'Cross Attention Fusion Network for Referring Camouflaged Object Detection' (CAReFuseNet). The framework is designed to extract reference-conditioned feature representations of the input camouflaged image, which are subsequently used to generate a binary foreground map to segment the target camouflaged object in the image. It independently extracts feature representations of the reference image(s) and of the camouflaged image, and then fuses them to align the camouflaged image representations with the reference target object. To perform this reference feature fusion, a novel cross-attention based reference feature fusion module, dubbed 'CAReFuse', is proposed. Specifically, two submodules are designed, namely the Cross-Attention Fusion (CAF) and the Multi-Scale Fusion (MSF) modules. The CAF module fuses the reference features with the multi-scale features extracted from the camouflaged image by performing cross-attention between them, thereby enabling global interactions between reference and camouflaged features. The MSF module then combines the reference-conditioned multi-scale features (the output of the CAF module) into final feature representations. Crucially, all fusion operations are carried out in a unified low-dimensional feature space, and a lightweight convolutional decoder is used for segmentation, leading to a compact overall architecture.

Extensive empirical evaluation is conducted to demonstrate both the effectiveness and the parameter efficiency of the proposed framework. With only 25% as many trainable parameters and 52% as many total inference parameters as SOTA Uncertainty-Aware Transformer (UAT) (27.22M vs 111.60M trainable; 89.30M vs 173.64M total; see Table 1), the PVT-v2 instantiation of CAReFuseNet (Ours-P) surpasses UAT in terms of standard Ref-COD metrics. These results indicate that a carefully designed, low-dimensional cross-attention fusion module, coupled with a lightweight decoder (Sections 3.3.1 and 3.2), can yield superior performance to heavier transformer-based architectures at a substantially reduced parameter budget.

Contributions. To summarize, the contributions of this work are as follows:

- A novel architecture for the Referring Camouflaged Object Detection task, dubbed CAReFuseNet, is proposed that leverages cross-attention for effective reference feature fusion while explicitly targeting parameter efficiency.

- A new feature fusion module, composed of cross-attention reference fusion and multi-scale fusion and dubbed 'CAReFuse', is introduced to extract reference-conditioned feature representations of camouflaged images in a unified low-dimensional feature space.
- Extensive experiments showing that CAReFuseNet (Ours-P) achieves superior performance to state-of-the-art UAT, while using only one quarter as many trainable parameters and roughly half as many total inference parameters (see Table 1), establishing CAReFuseNet as a strong parameter-efficient baseline for Ref-COD and demonstrating its generality across multiple backbone architectures.

2 Related Work

We first review the advances in Camouflaged Object Detection (COD), in Section 2.1, highlighting the notable methods that have been proposed. Next, we discuss the emerging novel task setting called Referring Camouflaged Object Detection (Ref-COD) in Section 2.2, which is also the task studied in our work. Finally, we explore the cross-attention based feature fusion methods, in Section 2.3, as it pertains to reference-guided vision tasks.

2.1 Camouflaged Object Detection

One of the first works that attempted to solve the challenging problem of identifying camouflaged objects in confusing and deceptive scenes was proposed by Fan et al. (2020a), which paved the way for increasing interest in COD. Their work published a large-scale dataset, named COD10K, which contains images of camouflaged objects and their annotations, and a framework, called SINet (Fan et al., 2020a), to predict the binary segmentation mask for the input camouflaged image. Subsequently, several deep learning strategies emerged for COD, including multi-scale-context based strategies such as CamoFormer (Yin et al., 2024) and FSPNet (Huang et al., 2023), mechanism simulation based strategies such as ZoomNet (Pang et al., 2022) and PreyNet (Zhang et al., 2022), and multi-source information fusion strategies such as FEMNet (Zhong et al., 2022) and FEDER (He et al., 2023). A survey paper by Xiao et al. (2024) extensively presents all the works related to COD. Despite various strategies and methods proposed, COD still remains a formidable problem. Owing to the increased attention towards COD due to several real-world applications (Fan et al., 2020b;c; Tabernik et al., 2020; Le et al., 2020), some novel task settings of COD have also emerged. One novel setting of COD is Referring Camouflaged Object Detection.

2.2 Referring Camouflaged Object Detection

The Referring Camouflaged Object Detection (Ref-COD) task, proposed by Zhang et al. (2025), requires detecting the camouflaged objects of only a particular user-specified category rather than detecting all the camouflaged objects in the input image. Thus, Ref-COD requires reference images containing salient target objects, along with the input camouflaged image, to guide the detection of the target camouflaged object. This makes Ref-COD a more constrained and more practically useful task setting of COD. Zhang et al. (2025) assembled a large-scale dataset, called R2C7K (Zhang et al., 2025), containing 7K images (including camouflaged and reference images) covering 64 object categories in real-world scenarios. R2C7K remains the only large-scale dataset containing salient reference images along with camouflaged images, unlike several COD datasets that contain camouflaged images alone. Zhang et al. (2025) also proposed a framework, dubbed R2CNet (Zhang et al., 2025), which fuses reference image features with camouflaged image features in two stages using a Referring Mask Generation module followed by a Referring Feature Enrichment module. Our proposed framework achieves the reference feature fusion in a single stage through our novel feature fusion module, called the CAReFuse module.

Recent methods further improved state-of-the-art (SOTA) in referring camouflaged object detection. Wu et al. (2025) introduced Uncertainty-Aware Transformer (UAT) for Referring Camouflaged Object Detection, aggregating reference features with cross-attention and modelling token uncertainties via a probabilistic decoder. While effective, UAT relies on a relatively heavy transformer encoder–decoder stack for visual reference integration and uncertainty modeling, whereas our method seeks to achieve better performance with a substantially more compact fusion and decoding design. Another method named CIRCOD (Gupta

et al., 2025) was proposed, which adopted a co-saliency-inspired approach to perform the Ref-COD task. However, CIRCOD uses slightly different settings than R2CNet and UAT: a higher image size (*i.e.* 512×512) and a single reference image (*i.e.* $K = 1$), versus 352×352 and $K = 5$ for the latter methods.

2.3 Cross-Attention based Feature Fusion

Since the success of transformer architecture (composed of attention operations) for vision tasks (Dosovitskiy et al., 2021), cross attention is also being used for feature fusion in vision tasks. CrossViT (Chen et al., 2021) was one of the early works that used cross-attention to fuse image feature representations extracted at two different granularities. Shen et al. (2024) have used iterative cross-attention guided feature fusion for multi-spectral object detection. Li & Wu (2024) and Sun et al. (2025) have used cross-attention mechanism for cross-modal feature fusion between infrared and visible images. Cross-attention fusion has also been used for cross-domain feature fusion by Tripathi et al. (2020; 2024), where feature representations of hand-drawn sketches are fused with image features for the sketch-guided object localization task that detects and localizes the target objects as specified by a sketch query. While these applications demonstrate cross-attention's versatility across modalities and scales, its effectiveness in fusing salient target object features with camouflaged image features remains underexplored. Our work addresses this by proposing a cross-attention-based feature fusion method for Ref-COD, which demonstrably surpasses the existing Ref-COD approaches.

3 Methodology

We first describe the formulation of the problem in Section 3.1. Next, we explain the overall architecture of our proposed framework in Section 3.2. Subsequently, we present our novel CAReFuse module in Section 3.3, comprising of the Cross-Attention Fusion module in Section 3.3.1 and the Multi-Scale Fusion module in Section 3.3.2.

3.1 Problem Formulation

Ref-COD is a reference-guided foreground map prediction task and is formally defined as follows. For a given image containing camouflaged objects, termed $I^{camo} \in \mathbb{R}^{3\times H\times W}$, and a given set of referring images containing a salient example of target object category c, denoted $\{I_k^{ref}\}_{k=1}^{K}, I_k^{ref} \in \mathbb{R}^{3\times H\times W}$, the output of Ref-COD is a binary segmentation mask $M^{seg} \in \{0,1\}^{1\times H\times W}$ for the camouflaged objects of category c in I^{camo}. Here, H and W represent height and width of the image, respectively. $M_{ij}^{seg} = 0$ indicates the pixel at position (i, j) belongs to background, while $M_{ij}^{seg} = 1$ represents the pixel (i, j) is part of the camouflaged object.

3.2 Overall Architecture

Our proposed CAReFuseNet, as illustrated in Figure 2, follows a dual-branch architecture. The first branch extracts the salient target object features, denoted as E, from K reference images, $\{I_k^{ref}\}_{k=1}^{K}$, through the Ref-Image Encoder followed by the Ref-Feature Combiner. The second branch extracts multi-scale features, denoted as $\{F_j\}_{j=1}^{4}$, from the camouflaged image, I^{camo}, using the Camo-Image Encoder. Our novel cross-attention based reference feature fusion module, dubbed CAReFuse, fuses the reference features, E, with the image features, $\{F_j\}_{j=1}^{4}$, to form the reference-conditioned feature representations of the input camouflaged image. This process helps to better align the features of the camouflaged image with the referring salient object. These strongly aligned image feature representations are subsequently used by the segmentation decoder to generate a binary segmentation mask for the target camouflaged objects as specified by the referring image.

Ref-Image Encoder. Following Zhang et al. (2025), we adopt the encoder from pre-trained ICON (Zhuge et al., 2023) model (with Pyramid Vision Transformer (Wang et al., 2022) backbone as default) as Ref-Image Encoder to extract the respective feature representations $\{F_k^{ref}\}_{k=1}^{K}$, where $F_k^{ref} \in \mathbb{R}^{c_r\times h\times w}$, from the reference images $\{I_k^{ref}\}_{k=1}^{K}$, where $I_k^{ref} \in \mathbb{R}^{3\times H\times W}$ (see Figure 2). Here, $h \times w$ represents the spatial

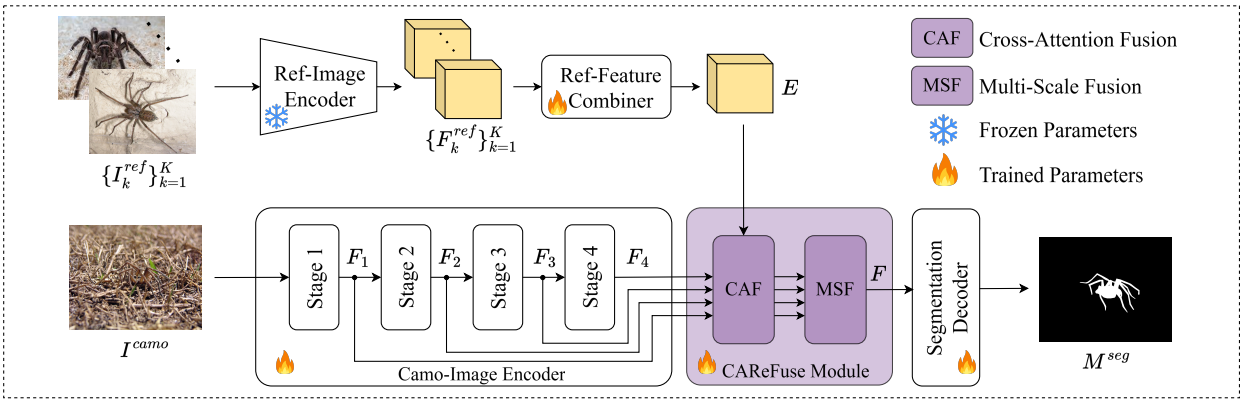

Figure 2: Overall architecture of our **CAReFuseNet** framework (Section 3.2). **Ref-Image Encoder** extracts features from reference images containing salient target objects. **Ref-Feature Combiner** combines individual reference image features to form a common feature representation, E, of the target object. **Camo-Image Encoder** extracts multi-scale features from input camouflaged image I^{camo}. **CAReFuse Module** fuses feature representations of salient target object with multi-scale features of I^{camo}. **Segmentation Decoder** generates a segmentation mask using the reference-conditioned feature representations.

dimension of the reference feature volume, while c_r is its channel dimension. These individual reference feature representations are subsequently combined by Ref-Feature Combiner.

Ref-Feature Combiner. Ref-Feature Combiner combines K reference feature representations, $\{F_k^{ref}\}_{k=1}^{K}$, where $F_k^{ref} \in \mathbb{R}^{c_r \times h \times w}$, into a single common salient target object representation, $E \in \mathbb{R}^{c_r \times h \times w}$ (see Figure 2), in a learnable way using spatial attention, channel attention followed by channel-wise weighted sum. Here, each feature volume comprises c_r feature planes of $h \times w$ spatial dimension each. In addition to the learnable parameters in the spatial attention and channel attention modules, a weight matrix $W \in \mathbb{R}^{K \times c_r}$ is learned to compute E, which is formulated as:

$$\begin{aligned} \tilde{F}_k^{ref} &= M_c(F_k^{ref}) \odot F_k^{ref}, \\ \hat{F}_k^{ref} &= M_s(\tilde{F}_k^{ref}) \odot \tilde{F}_k^{ref}, \\ E[c] &= \sum_{k=1}^{K} \alpha_{k,c} \cdot \hat{F}_k^{ref}[c], \text{ for } c = 1, 2, ..., c_r \end{aligned} \tag{1}$$

$$\text{where } \alpha_{k,c} = \frac{exp(W_{k,c})}{\sum_{j=1}^{K} exp(W_{j,c})}, \; M_c(.) \text{ and } M_s(.) \text{ are given by equations 2 and 3.}$$

$E[c]$ is the c^{th} feature plane of E, and $\hat{F}_k^{ref}[c]$ is the c^{th} feature plane from the k^{th} feature volume $\hat{F}_k^{ref}$.

$$M_c(.) = \sigma\Big(MLP\big(AvgPool_s(.)\big) + MLP\big(MaxPool_s(.)\big)\Big) \tag{2}$$

$$M_s(.) = \sigma\Big(\mathcal{F}_{conv7\times7}\big([AvgPool_c(.); MaxPool_c(.)]\big)\Big) \tag{3}$$

Camo-Image Encoder. To extract multi-scale features from $I^{camo} \in \mathbb{R}^{3 \times H \times W}$, the pre-trained Pyramid Vision Transformer (PVT) (Wang et al., 2022) is chosen as the default Camo-Image Encoder. The features $\{F_j\}_{j=1}^{4}$, where $F_j \in \mathbb{R}^{c_j \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}}$, are extracted from stage 1, stage 2, stage 3 and stage 4 of PVT, respectively (see Figure 2). The varying channel dimension, c_j, of the multi-scale features is adjusted to a common channel dimension, c_d, through the convolution operation. Subsequently, these channel-adjusted multi-scale features $\{F_j\}_{j=1}^{4}$, where $F_j \in \mathbb{R}^{c_d \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}}$ are passed to the CAReFuse module for feature fusion. Importantly, all downstream fusion in CAReFuse module (cross-attention and multi-scale fusion)

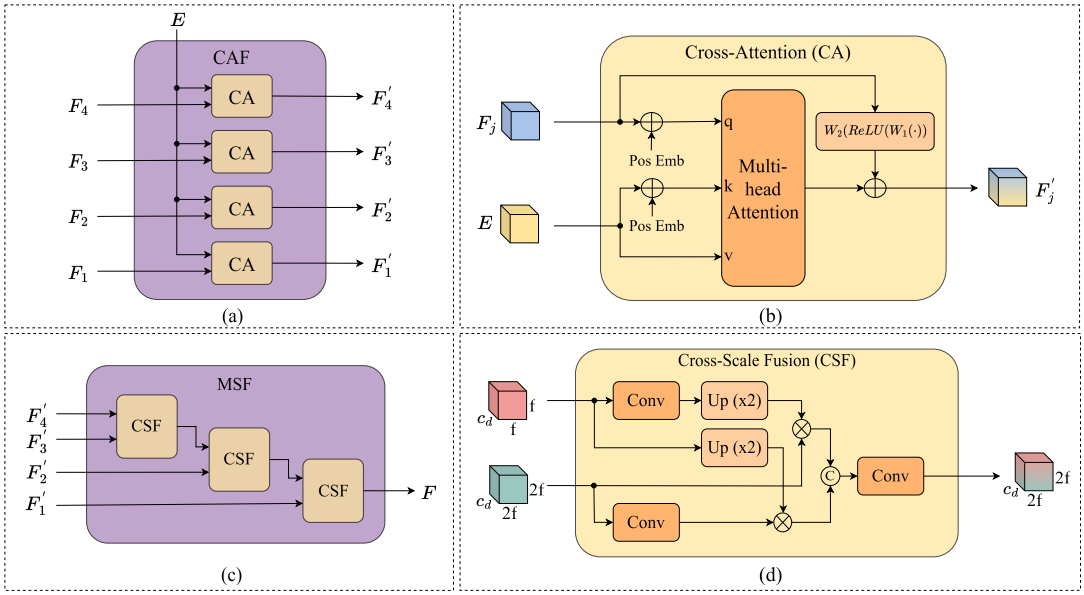

Figure 3: **CAReFuse Module** (Section 3.3): cross-attention based feature fusion module that fuses reference image features with multi-scale camouflaged image features. (a) **Cross-Attention Fusion (CAF) Module**: fuses reference image features E with multi-scale camouflaged image features F_1, F_2, F_3 and F_4 individually. (b) **Cross-Attention (CA) Module**: performs multi-head cross-attention using camuoflaged image features for *queries* and reference image features for *keys* and *values*. (c) **Multi-Scale Fusion (MSF) Module**: fuses reference-conditioned multi-scale features of camouflaged image into a single feature volume through cross-scale fusion. (d) **Cross-Scale Fusion (CSF) Module:** fuses a feature volume of lower spatial dimension with a feature volume of higher spatial dimension through convolution and up-sampling operations. $\oplus$ - element-wise sum, $\otimes$ - element-wise multiplication, and ⓒ - concatenate operation.

operates on this unified c_d-dimensional space. This contrasts with prior work such as UAT, which performs cross-attention and probabilistic decoding in a substantially higher-dimensional feature space, leading to a much larger number of trainable parameters.

CAReFuse module, described in detail in Section 3.3, performs feature fusion between E and $\{F_j\}_{j=1}^{4}$ using cross-attention fusion followed by multi-scale fusion. The final feature representations produced by the CAReFuse module are denoted as $F \in \mathbb{R}^{c_d \times \frac{H}{4} \times \frac{W}{4}}$.

Segmentation Decoder. Similar to R2CNet (Zhang et al., 2025), we have used a convolution head as the segmentation decoder. The convolution head maps the final reference-conditioned feature representations F to the segmentation mask $M^{seg} \in \{0,1\}^{1 \times H \times W}$ through a series of convolution blocks.

3.3 CAReFuse Module

To segment the camouflaged objects of category c present in the image I^{camo}, using reference images $\{I_k^{ref}\}_{k=1}^{K}$ containing salient examples of object category c, reference-conditioned feature representations of I^{camo} must be extracted. To achieve this effectively through reference feature fusion, we propose a novel feature fusion module based on cross-attention. We denote these reference-conditioned camouflaged image features as F.

Our proposed CAReFuse module contains two submodules, namely, the Cross-Attention Fusion (CAF) and the Multi-Scale Fusion (MSF) modules.

The multi-scale features of the camouflaged image, $\{F_j\}_{j=1}^{4}$, are extracted at different spatial scales. CAF module fuses the referring salient target object features, E, with $\{F_j\}_{j=1}^{4}$ individually to form $\{F_j^{'}\}_{j=1}^{4}$

respectively. This fusion operation can be formulated as:

$$\{F_j^{'}\}_{j=1}^{4} = \text{CAF}(\{F_j\}_{j=1}^{4}, E). \tag{4}$$

To integrate information from different spatial scales, the reference-fused multi-scale features, $\{F_j^{'}\}_{j=1}^{4}$, are fused together using the MSF module. The resultant feature representations, F, can be expressed as:

$$F = \text{MSF}(\{F_j^{'}\}_{j=1}^{4}). \tag{5}$$

3.3.1 Cross-Attention Fusion Module

The Cross-Attention Fusion module contains four Cross-Attention (CA) blocks, each of which fuses E with $\{F_j\}_{j=1}^{4}$ respectively. Figure 3(a) illustrates this process.

Cross-Attention Block. This block is illustrated in Figure 3(b). The reference feature volume $E \in \mathbb{R}^{c_r \times h \times w}$ is reshaped as a sequence of tokens denoted as $\phi_E \in \mathbb{R}^{hw \times c_r}$. Similarly, the feature volume $F_j \in \mathbb{R}^{c_d \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}}$, extracted from the camouflaged image, is also reshaped as a sequence of tokens denoted as $\phi_{F_j} \in \mathbb{R}^{(\frac{H}{2^{j+1}} \frac{W}{2^{j+1}}) \times c_d}$. Now, multi-headed cross-attention (Vaswani et al., 2017) is performed between the two sequences, using ϕ_{F_j} as *queries* and ϕ_E as *keys* and *values*, as formulated in equation 6. To induce positional information, we also add learnable positional embeddings to query and key tokens before computing the multi-headed attention.

$$\phi_{F_j^{'}} = \mathcal{F}_{softmax}\left(\frac{(\phi_{F_j} W^Q)(\phi_E W^K)^T}{\sqrt{d}}\right)\phi_E W^V \tag{6}$$

Here, d is the dimension of the projected query and key vectors, while $W^Q \in \mathbb{R}^{c_d \times d}$, $W^K \in \mathbb{R}^{c_r \times d}$ and $W^V \in \mathbb{R}^{c_r \times c_d}$ are the projection matrices for query, key, and value vectors, respectively. For brevity, equation 6 shows the computation of cross-attention only with one attention head, though we use multiple attention heads in practice. The multi-headed cross-attention output, $\phi_{F_j^{'}} \in \mathbb{R}^{(\frac{H}{2^{j+1}} \frac{W}{2^{j+1}}) \times c_d}$, has the same dimension as that of the input ϕ_{F_j}. Inspired by the cross-attention fusion process in Sketch-guided Vision Transformer Encoder (Tripathi et al., 2024), we nonlinearly project ϕ_{F_j} and add it to the output of the multi-head attention. This process can be formulated as:

$$\phi_{F_j^{'}} = W_2 \cdot \mathcal{F}_{ReLU}(W_1 \cdot \phi_{F_j}) + \phi_{F_j^{'}}, \tag{7}$$

where, $W_1 \in \mathbb{R}^{d^{'} \times (\frac{H}{2^{j+1}} \frac{W}{2^{j+1}})}$ and $W_2 \in \mathbb{R}^{(\frac{H}{2^{j+1}} \frac{W}{2^{j+1}}) \times d^{'}}$ are the trainable projection matrices.

The output sequence of tokens $\phi_{F_j^{'}} \in \mathbb{R}^{(\frac{H}{2^{j+1}} \frac{W}{2^{j+1}}) \times c_d}$ are reshaped into the feature volume $F_j^{'} \in \mathbb{R}^{c_d \times \frac{H}{2^{j+1}} \times \frac{W}{2^{j+1}}}$. The reference-fused multi-scale features, $\{F_j^{'}\}_{j=1}^{4}$, are then fused together using the MSF module.

Note that at each scale, only a single cross-attention block is applied between the camouflaged image features and the reference features, followed by a lightweight non-linear projection. Together with the convolution-only multi-scale fusion (see Section 3.3.2) and segmentation decoder, this design keeps the fusion head compact while still enabling rich global interactions between camouflaged and reference features.

3.3.2 Multi-Scale Fusion Module

The Multi-Scale Fusion module progressively fuses the reference-fused multi-scale features, $\{F_j^{'}\}_{j=1}^{4}$, to form a single feature volume F, as illustrated in Figure 3(c). This is performed using three Cross-Scale Fusion (CSF) blocks, each of which fuses a feature volume of the lower spatial dimension with that of the higher spatial dimension, through convolution and up-sampling operations.

$$F = \text{CSF}(\text{CSF}(\text{CSF}(F_4^{'}, F_3^{'}), F_2^{'}), F_1^{'}) \tag{8}$$

Table 1: Comparision of our CAReFuseNet (with different backbone choices) with Ref-COD models. '-Ref': R2CNet's referring framework applied to the existing COD model (results taken from (Zhang et al., 2025)). 'Params': Number of model parameters in million. 'Macs': Multiply-accumulate operations in billion. '↑': the higher the better. '↓': the lower the better. Settings: image size 352×352, number of reference images $K = 5$. The best two values are shown in **bold** and underline, respectively.

Model	Backbone	Trainable Params(M)	Total Params(M)	Macs(G)	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
R2CNet (Zhang et al., 2025)	ResNet-50	27.15	60.23	23.23	0.805	0.879	0.669	0.036
PFNet-Ref	ResNet-50	57.58	90.66	59.59	0.811	0.885	0.687	0.036
PreyNet-Ref	ResNet-50	38.70	71.78	117.60	0.817	0.900	0.704	0.032
DGNet-Ref	EfficientNet-B5	20.10	53.18	7.24	0.821	0.891	0.696	0.032
SINetV2-Ref	Res2Net-50	27.70	60.78	26.01	0.823	0.888	0.700	0.033
BSANet-Ref	Res2Net-50	33.07	66.15	66.08	0.830	0.912	0.727	0.030
ZoomNet-Ref	Tripple ResNet-50	33.30	66.38	218.24	0.834	0.886	0.720	0.029
BGNet-Ref	Res2Net-50	151.06	184.14	171.03	0.840	0.909	0.738	0.029
UAT (Wu et al., 2025)	PVT-v2	111.60	173.64	61.70	<u>0.855</u>	0.912	<u>0.757</u>	<u>0.026</u>
Ours-R	ResNet-50	27.21	50.72	13.79	0.824	0.898	0.707	0.032
Ours-E	EfficientNet-B5	31.65	55.16	11.34	0.830	0.904	0.718	0.029
Ours-S	Swin-Tiny	21.93	108.61	29.30	0.841	0.912	0.735	0.028
Ours-C	ConvNeXt-Tiny	30.89	54.40	13.91	0.847	<u>0.920</u>	0.751	0.027
Ours-P	PVT-v2	27.22	89.30	20.15	**0.861**	**0.928**	**0.778**	**0.024**

Cross-Scale Fusion Block. The process of cross-scale fusion, inspired by Zheng et al. (2023), is illustrated in Figure 3(d). The multi-scale features, $\{F_j^{'}\}_{j=1}^{4}$, have their spatial dimensions halved for $j = 1$ to 4. Therefore, the fusion between the feature volumes of two different spatial scales, say, $U \in \mathbb{R}^{c_d \times f \times f}$ and $V \in \mathbb{R}^{c_d \times 2f \times 2f}$, is formulated as:

$$\begin{aligned} u &= \mathcal{F}_{up}(\mathcal{F}_{conv1\times1}(U)) \odot V, \\ v &= \mathcal{F}_{up}(U) \odot \mathcal{F}_{conv1\times1}(V), \\ \text{CSF}(U, V) &= \mathcal{F}_{conv1\times1}(\mathcal{F}_{concate}(u, v)), \end{aligned} \tag{9}$$

where, $\mathcal{F}_{up}(.)$ indicates up-sampling, $\mathcal{F}_{conv1\times1}(.)$ is a 1×1 convolution operation, and $\otimes$ represents element-wise multiplication. Here, the up-sampling operation is performed to match the spatial dimensions.

4 Experiments and Results

We explain the training and evaluation setup in Section 4.1. Next, we report the results of the quantitative evaluation and the ablation studies in Section 4.2 and Section 4.3, respectively. Finally, Section 4.4 discusses results of qualitative evaluation.

4.1 Training and Evaluation Setup

Following the existing Ref-COD framework, R2CNet (Zhang et al., 2025), we choose structure loss (Wei et al., 2020) as our training objective. This function is formulated as:

$$\mathcal{L}(P, G) = \sum_{i=1}^{4} \mathcal{L}_{bce}(P_i, G) + \mathcal{L}_{iou}(P_i, G), \tag{10}$$

where $P = \{M^{seg}, M_4^{scale}, M_3^{scale}, M_2^{scale}\}$ refers to the predicted segmentation masks and G represents the ground truth. Here, M^{seg} is the final predicted segmentation mask, while $M_4^{scale}, M_3^{scale}, M_2^{scale}$ are the mask predictions using the multi-scale features at the output of the Cross-Attention Fusion module.

Implementation Details. We adopt Adam optimizer (Kingma & Ba, 2015) to train our network for 100 epochs with a batch size of 32. We choose the initial learning rate of 5e-4 that eventually decays according to cosine annealing (Loshchilov & Hutter, 2017). During training, the Ref-Image Encoder parameters are

Table 2: Comparison of our CAReFuseNet (with different backbone choices) with CIRCOD (under Ref-COD setting). Settings: image size 512×512, number of ref images $K = 1$. The best two values are shown in **bold** and underline, respectively.

Model	Backbone	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
CIRCOD (Gupta et al., 2025)	PVT-v2	0.848	0.918	0.756	0.026
Ours-R	ResNet-50	0.835	0.909	0.735	0.030
Ours-E	EfficientNet-B5	0.859	0.927	0.775	<u>0.024</u>
Ours-S	Swin-Tiny	<u>0.861</u>	0.924	0.773	0.025
Ours-C	ConvNeXt-Tiny	<u>0.861</u>	<u>0.931</u>	<u>0.782</u>	<u>0.024</u>
Ours-P	PVT-v2	**0.879**	**0.939**	**0.812**	**0.021**

frozen while the rest of the network parameters are updated. The parameters of the pre-trained Camo-Image Encoder are updated using 0.1 times the original learning rate used for updating the rest of the trainable parameters. During both training and evaluation, we resize the images to the default size of 352×352. The experiments are implemented using PyTorch (Paszke et al., 2019) framework and run on NVIDIA RTX 6000 Ada GPU.

Evaluation Metrics. To evaluate our model, we adopt the four metrics commonly used in foreground map evaluation, including structure-measure (S_m) (Fan et al., 2017), adaptive E-measure (αE) (Fan et al., 2018), weighted F-measure (ωF) (Margolin et al., 2014) and mean absolute error (M) (Perazzi et al., 2012).

4.2 Quantitative Evaluation

We compare the performance of our CAReFuseNet with CIRCOD (Gupta et al., 2025) and with remaining Ref-COD methods separately, since CIRCOD uses different settings. Table 1 compares our CAReFuseNet with R2CNet (and all other Ref-COD extended models using R2CNet's framework) (Zhang et al., 2025) and UAT (Wu et al., 2025), which use image size 352×352 and the number of reference images $K = 5$. Table 2 compares with CIRCOD that uses image size 512×512 and $K = 1$. To verify the generality of our method, we evaluated our CAReFuseNet with several other backbone architectures in addition to Pyramid Vision Transformer (PVT) Wang et al. (2022), including Swin-Tiny Transformer (Sw-T) (Liu et al., 2021), ResNet-50 (R-50) (He et al., 2016), ConvNeXt-Tiny (Cxt-T) (Liu et al., 2022) and EfficientNet-B5 (E-B5) (Tan & Le, 2019). When using Sw-T backbone for Camo-Image Encoder, we use ICON (Zhuge et al., 2023) with Swin backbone as Ref-Image Encoder. Whereas, when using R-50, E-B5, Cxt-T for Camo-Image Encoder, we use ICON with ResNet-50 backbone as Ref-Image Encoder.

As can be observed in Table 1, with only 27% as many trainable parameters and 53% as many total inference parameters, our method (Ours-P) outperforms UAT. This reduction arises from several architectural choices. First, CAReFuseNet uses the ICON encoder as a frozen Ref-Image Encoder, so a large fraction of the total parameters do not contribute to the trainable budget. Second, all multi-scale camouflaged features are projected to a unified channel dimension $c_d = 256$, and both the Cross-Attention Fusion and Multi-Scale Fusion modules operate at this reduced dimensionality. Third, reference–image fusion is implemented with a single cross-attention block per scale followed by shallow 1×1 convolutions for cross-scale fusion and a lightweight convolutional decoder. Table 2 shows that our method also outperforms CIRCOD in all metrics under E-B5, Sw-T, CxT-T and PVT backbones.

4.3 Ablative Studies

Ablation on c_d. As the channel dimension of the multi-scale camouflaged image features (*i.e.* c_d) impacts the number of parameters of the model, we conduct an ablation study on c_d to assess its effect on performance. We vary the channel dimension from 64 to 512 in powers of two, training and evaluating CAReFuseNet for each configuration. As shown in Table 3, an increase in c_d from 64 to 256 improves performance, while the performance gain saturates with a further increase. Therefore, for a better trade-off between performance and computational efficiency, and to avoid over-parameterization, we select 256 as the default channel dimension. This design, together with the parameter-efficient fusion and decoding modules, also

Table 3: Ablation on channel dimension (*i.e.* c_d) of our CAReFuseNet with different backbone choices. 'Param': Number of model parameters in million. 256 is chosen as the default channel dimension, striking a balance between performance and computational efficiency.

c_d	CAReFuseNet-R (Ours-R)				CAReFuseNet-E (Ours-E)				CAReFuseNet-S (Ours-S)				CAReFuseNet-C (Ours-C)				CAReFuseNet-P (Ours-P)			
	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
64	0.819	0.894	0.698	0.033	0.826	0.901	0.712	0.030	0.841	0.909	0.734	0.029	0.844	0.915	0.744	0.028	0.857	0.923	0.770	0.025
128	0.818	0.899	0.706	**0.032**	0.829	0.902	0.716	0.030	**0.843**	**0.912**	**0.740**	**0.028**	0.845	**0.920**	0.749	**0.027**	0.858	0.921	0.772	**0.024**
256	**0.824**	0.898	0.707	**0.032**	**0.830**	0.904	0.718	**0.029**	0.841	**0.912**	0.735	**0.028**	**0.847**	**0.920**	**0.751**	**0.027**	**0.861**	**0.928**	0.778	**0.024**
512	**0.824**	**0.902**	**0.709**	**0.032**	**0.830**	**0.905**	**0.719**	**0.029**	**0.843**	0.911	0.739	**0.028**	0.846	**0.920**	0.749	**0.027**	0.860	**0.928**	**0.778**	**0.024**

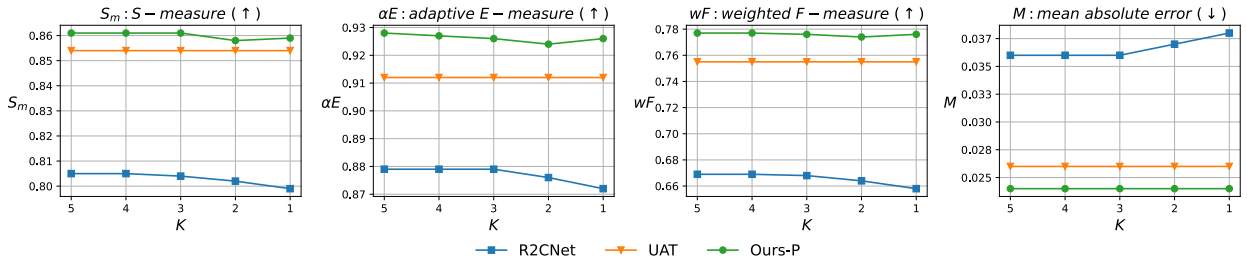

Figure 4: Comparison among CAReFuseNet, UAT and R2CNet w.r.t ablation experiment on number of reference images (*i.e.*, K).

contributes to the substantial parameter savings observed in Table 1 when comparing Ours-P to UAT, which relies on higher-dimensional transformer blocks for both reference feature fusion and decoding.

Ablation on K. We trained and evaluated our CAReFuseNet by varying the number of reference images, $5 \geq K \geq 1$, used to guide camouflaged object detection. Our results are compared with those of UAT and R2CNet in Figure 4. The comparison demonstrates that our CAReFuseNet outperforms both UAT and R2CNet, even when utilizing only one reference image, while others using five.

Ablation on Ref-Feature Combiner. As formulated in Equations 1, 2 and 3, our Ref-Feature Combiner performs channel attention (CA) followed by spatial attention (SA) on the individual reference feature volumes, $\{F_k^{ref}\}_{k=1}^{K} \in \mathbb{R}^{c_r \times h \times w}$, extracted by Ref-Image Encoder. Subsequently, a learnable channel-wise weighted sum is performed on these K feature volumes. We conduct an ablation study on various operations involved in the Ref-Feature Combiner. As shown in Table 4, our chosen strategy gives better results among all choices.

Ablation on CAReFuse Module Components. To establish the importance of each component of the CAReFuse module, we conducted ablation studies, with results reported in Table 5. Both the proposed modules boost performance. The CAF module yields improvements of 0.6%, 0.4%, 1.1% and 0.2%, while MSF module delivers gains of 4.8%, 7.1%, 11.7% and 1.4%, in metrics S_m, αE, ωF and M, respectively. When combined, the two modules together achieve gains of 5.7%, 7.4%, 13% and 1.5% in the respective metrics.

Table 4: Ablation study on the Ref-Feat Combiner module. 'CA': Channel Attention, 'SA': Spatial Attention, 'WS': Weighted Sum, 'CWS': Channel-wise Weighted Sum.

CA	SA	WS	CWS	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
		✓		0.859	0.925	0.775	0.024
			✓	0.859	0.927	0.776	0.024
	✓		✓	0.860	0.927	0.777	0.024
✓			✓	0.860	0.927	0.777	0.024
✓	✓		✓	0.861	0.928	0.778	0.024

Table 5: Ablation study on the components of the CAReFuse module. 'CAF': Cross Attention Fusion, 'MSF': Multi Scale Fusion.

CAF	MSF	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
✗	✗	0.804	0.854	0.647	0.039
✓	✗	0.810	0.858	0.658	0.037
✗	✓	0.852	0.925	0.764	0.025
✓	✓	0.861	0.928	0.778	0.024

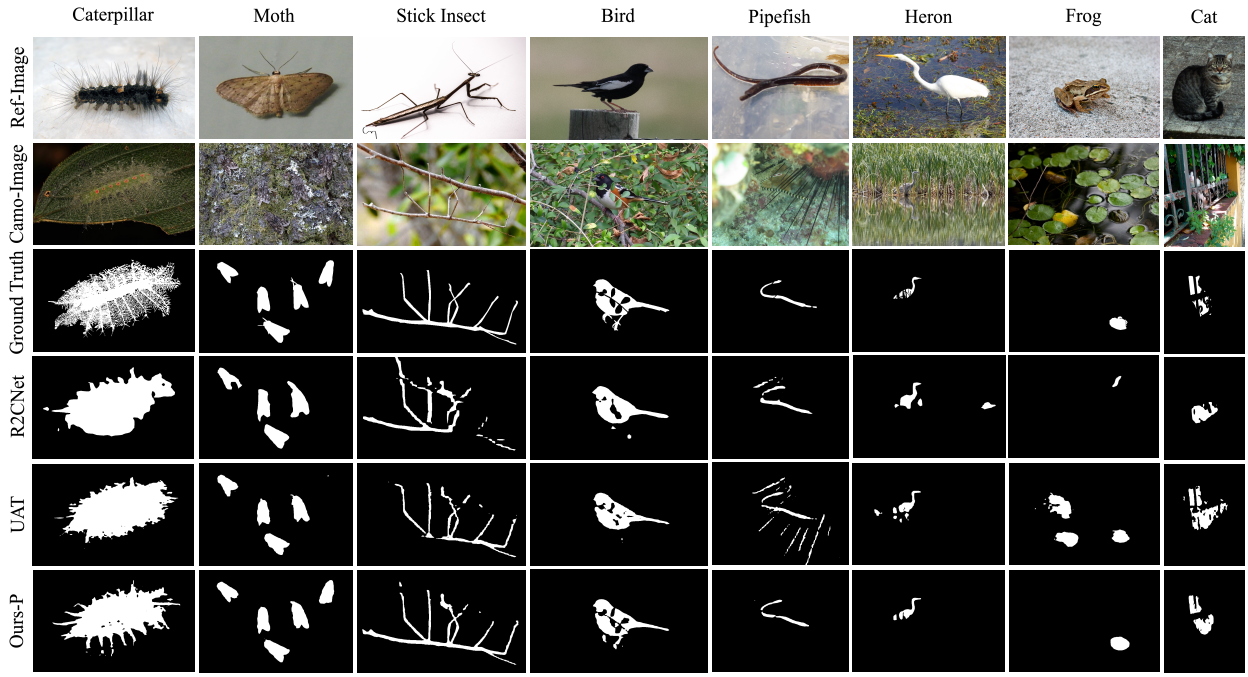

Figure 5: Visual comparison between the predictions of a state-of-the-art Ref-COD methods (*i.e.* R2CNet, UAT) and our CAReFuseNet-P. Our method evidently outperforms state-of-the-art.

4.4 Qualitative Evaluation

We present a qualitative comparison of the mask predictions of CAReFuseNet (Ours-P) with other Ref-COD methods in Figure 5. As can be observed from the visualization, the predictions of our method are substantially superior compared to those of the other methods in highly challenging scenarios such as very small or narrow objects, occluded objects, and objects with complex boundaries. As can be noticed in case of *Moth*, *Heron* and *Frog*, R2CNet and UAT either miss some smaller objects or detect false positives in the camouflaged images, while our method segments them accurately. In case of *Stick Insect* and *Pipefish*, it can be observed that our method detects very narrow parts of the objects more effectively, while other methods fail. The case of *Bird* and *Cat* proves the effectiveness of our CAReFuseNet in detecting and accurately segmenting objects under severe occlusion, where R2CNet and UAT struggle. The predicted masks for *Caterpillar* demonstrate that our CAReFuseNet segmented the objects with complex boundaries more precisely than other methods.

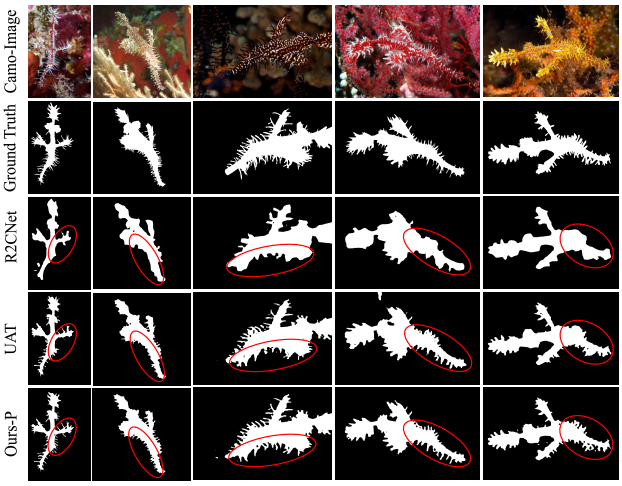

Figure 6: Visual comparision between the masks produced by state-of-the-art R2CNet and CAReFuseNet (Ours-P) for camouflaged objects with high frequency edges. Our method produces superior quality masks, especially evidenced by the finer details being better captured. (Best viewed zoomed in)

Another notable strength of our method is its superior ability to generate high-quality segmentation masks for camouflaged objects, particularly in regions with high-frequency edges. As illustrated in Figure 6, the segmentation masks produced by our CAReFuseNet capture object boundaries with greater accuracy and

preserve fine edge details more effectively than those generated by R2CNet. Remarkably, with significantly less number of trainable and total inference parameters (see Table 1), our method still narrowly outperforms UAT in producing high quality edges. This qualitative comparison clearly demonstrates the effectiveness of our CAReFuseNet (Ours-P) in handling challenging edge-rich regions, underscoring its overall superiority in segmenting camouflaged objects.

5 Conclusion

Given the difficult nature of the COD problem in general and the practical utility of the Ref-COD task in particular, this paper proposes a novel cross-attention based reference feature fusion network (CAReFuseNet) for segmenting camouflaged objects in an image under the guidance of reference images containing the salient target object. The method introduces a cross-attention based feature fusion module (CAReFuse module) to extract the reference-conditioned feature representations of camouflaged images, combining a Cross-Attention Fusion (CAF) submodule with a Multi-Scale Fusion (MSF) submodule that operates in a unified low-dimensional feature space. By freezing the Ref-Image Encoder, projecting all camouflaged image features to a shared reduced channel dimension, and employing a single cross-attention block per scale followed by lightweight convolutional fusion and decoding, the overall architecture is explicitly designed to be parameter-efficient.

Extensive experiments on the R2C7K dataset demonstrate that CAReFuseNet achieves strong quantitative performance while being substantially more compact than prior approaches. In particular, the PVT-v2 instantiation (Ours-P) surpasses Uncertainty-Aware Transformer (UAT), despite using only about one-quarter of UAT's trainable parameters and roughly half of the total inference parameters (Table 1). Under the alternative setting adopted by CIRCOD, CAReFuseNet also attains superior performance across multiple backbones (Table 2). Qualitative results further show that the proposed model yields marked improvements in challenging scenarios such as very small or narrow objects, heavily occluded objects, and objects with complex boundaries, while still preserving fine edge details.

Overall, this work highlights that explicit attention to parameter efficiency in the design of reference feature fusion can yield Ref-COD models that are both accurate and lightweight. The proposed CAReFuseNet provides a practical, parameter-efficient baseline for future research on Ref-COD.

Limitations and Future Directions

Like other Ref-COD methods (e.g., R2CNet (Zhang et al., 2025) and UAT (Wu et al., 2025)), our method does not distinguish the scenarios where the camouflaged image to be segmented does not contain the target object specified by the reference images; it assumes the camouflaged image to always contain the specified target objects. It would be valuable to extend our parameter-efficient Ref-COD method to accommodate more generalized scenes without this restricted assumption. Furthermore, while accepting multiple reference images, our method tries to segment a single target object category (specified by the set of reference images) at a time. It will also be promising to extend the Ref-COD methods to accept multi-category references, enabling them to segment the target objects of multiple categories in the camouflaged image.

References

Chun-Fu Richard Chen, Quanfu Fan, and Rameswar Panda. CrossViT: Cross-Attention Multi-Scale Vision Transformer for Image Classification . In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 347–356, Los Alamitos, CA, USA, October 2021. IEEE Computer Society. doi: 10.1109/ICCV48922.2021.00041. URL `https://doi.ieeecomputersociety.org/10.1109/ICCV48922.2021.00041`.

Xuehan Chen, Guangyu Ren, Tianhong Dai, Tania Stathaki, and Hengyan Liu. Enhancing prompt generation with adaptive refinement for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20672–20682, October 2025.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations (ICLR)*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy. Oral.

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)*, October 2017.

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*, pp. 698–704. International Joint Conferences on Artificial Intelligence Organization, July 2018. doi: 10.24963/ijcai.2018/97. URL https://doi.org/10.24963/ijcai.2018/97.

Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020a.

Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In Anne L. Martel, Purang Abolmaesumi, Danail Stoyanov, Diana Mateus, Maria A. Zuluaga, S. Kevin Zhou, Daniel Racoceanu, and Leo Joskowicz (eds.), *Medical Image Computing and Computer Assisted Intervention – MICCAI 2020*, pp. 263–273, Cham, 2020b. Springer International Publishing. ISBN 978-3-030-59725-2.

Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020c. doi: 10.1109/TMI.2020.2996645.

Avi Gupta, Koteswar Rao Jerripothula, and Tammam Tillo. Circod: Co-saliency inspired referring camouflaged object discovery. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 8313–8323, 2025. doi: 10.1109/WACV61041.2025.00806.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22046–22055, 2023. doi: 10.1109/CVPR52729.2023.02111.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

Ronghang Hu, Marcus Rohrbach, and Trevor Darrell. Segmentation from natural language expressions. In Bastian Leibe, Jiri Matas, Nicu Sebe, and Max Welling (eds.), *Computer Vision – ECCV 2016*, pp. 108–124, Cham, 2016. Springer International Publishing. ISBN 978-3-319-46448-0.

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 *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5557–5566, Los Alamitos, CA, USA, June 2023. IEEE Computer Society. doi: 10.1109/CVPR52729.2023.00538. URL https://doi.ieeecomputersociety.org/10.1109/CVPR52729.2023.00538.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In Yoshua Bengio and Yann LeCun (eds.), *ICLR (Poster)*, 2015. URL http://dblp.uni-trier.de/db/conf/iclr/iclr2015.html#KingmaB14.

Minh-Quan Le, Minh-Triet Tran, Trung-Nghia Le, Tam V. Nguyen, and Thanh-Toan Do. CamoFA: A Learnable Fourier-Based Augmentation for Camouflage Segmentation . In *2025 IEEE/CVF Winter Conference*

on Applications of Computer Vision (WACV), pp. 3427–3436, Los Alamitos, CA, USA, March 2025. IEEE Computer Society. doi: 10.1109/WACV61041.2025.00338. URL https://doi.ieeecomputersociety.org/10.1109/WACV61041.2025.00338.

Xinyi Le, Junhui Mei, Haodong Zhang, Boyu Zhou, and Juntong Xi. A learning-based approach for surface defect detection using small image datasets. *Neurocomputing*, 408:112–120, 2020. ISSN 0925-2312. doi: https://doi.org/10.1016/j.neucom.2019.09.107. URL https://www.sciencedirect.com/science/article/pii/S0925231220303386.

Minhyun Lee, Seungho Lee, Song Park, Dongyoon Han, Byeongho Heo, and Hyunjung Shim. MaskRIS: Semantic distortion-aware data augmentation for referring image segmentation. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=EtK4madHmc.

Cheng Lei, Jie Fan, Xinran Li, Tian-Zhu Xiang, Ao Li, Ce Zhu, and Le Zhang. Towards Real Zero-Shot Camouflaged Object Segmentation Without Camouflaged Annotations . *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 47(12):11990–12004, December 2025. ISSN 1939-3539. doi: 10.1109/TPAMI.2025.3600461. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2025.3600461.

Hui Li and Xiao-Jun Wu. Crossfuse: A novel cross attention mechanism based infrared and visible image fusion approach. *Information Fusion*, 103:102147, 2024. ISSN 1566-2535. doi: https://doi.org/10.1016/j.inffus.2023.102147. URL https://www.sciencedirect.com/science/article/pii/S1566253523004633.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9992–10002, 2021. doi: 10.1109/ICCV48922.2021.00986.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11966–11976, 2022. doi: 10.1109/CVPR52688.2022.01167.

Ilya Loshchilov and Frank Hutter. SGDR: stochastic gradient descent with warm restarts. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. URL https://openreview.net/forum?id=Skq89Scxx.

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.

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)*, pp. 2160–2170, June 2022.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 733–740, 2012. doi: 10.1109/CVPR.2012.6247743.

Guangyu Ren, Hengyan Liu, Michalis Lazarou, and Tania Stathaki. Multi-modal segment anything model for camouflaged scene segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 19882–19892, October 2025.

Jifeng Shen, Yifei Chen, Yue Liu, Xin Zuo, Heng Fan, and Wankou Yang. Icafusion: Iterative cross-attention guided feature fusion for multispectral object detection. *Pattern Recognition*, 145:109913, 2024. ISSN 0031-3203. doi: https://doi.org/10.1016/j.patcog.2023.109913. URL https://www.sciencedirect.com/science/article/pii/S0031320323006118.

Xicheng Sun, Fu Lv, Yongan Feng, and Xu Zhang. Dmcm: Dwo-branch multilevel feature fusion with cross-attention mechanism for infrared and visible image fusion. *PLOS ONE*, 20(3):1–25, 03 2025. doi: 10.1371/journal.pone.0318931. URL https://doi.org/10.1371/journal.pone.0318931.

Domen Tabernik, Samo Šela, Jure Skvarč, and Danijel Skočaj. Segmentation-based deep-learning approach for surface-defect detection. *Journal of Intelligent Manufacturing*, 31(3):759–776, March 2020. doi: 10.1007/s10845-019-01476-x. URL https://ideas.repec.org/a/spr/joinma/v31y2020i3d10.1007_s10845-019-01476-x.html.

Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 6105–6114. PMLR, 2019. URL http://proceedings.mlr.press/v97/tan19a.html.

Aditay Tripathi, Rajath R. Dani, Anand Mishra, and Anirban Chakraborty. Sketch-guided object localization in natural images. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm (eds.), *Computer Vision – ECCV 2020*, pp. 532–547, Cham, 2020. Springer International Publishing. ISBN 978-3-030-58539-6.

Aditay Tripathi, Anand Mishra, and Anirban Chakraborty. Query-guided attention in vision transformers for localizing objects using a single sketch. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pp. 1072–1081, 2024. doi: 10.1109/WACV57701.2024.00112.

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 (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

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. doi: 10.1007/s41095-022-0274-8.

Jun Wei, Shuhui Wang, and Qingming Huang. F^{3}net: Fusion, feedback and focus for salient object detection. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(07):12321–12328, Apr. 2020. doi: 10.1609/aaai.v34i07.6916. URL https://ojs.aaai.org/index.php/AAAI/article/view/6916.

Ranwan Wu, Tian-Zhu Xiang, Guo-Sen Xie, Rongrong Gao, Xiangbo Shu, Fang Zhao, and Ling Shao. Uncertainty-aware transformer for referring camouflaged object detection. *IEEE Transactions on Image Processing*, 34:5341–5354, 2025. doi: 10.1109/TIP.2025.3587579.

Yu-Huan Wu, Shang-Hua Gao, Jie Mei, Jun Xu, Deng-Ping Fan, Rong-Guo Zhang, and Ming-Ming Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *IEEE Transactions on Image Processing*, 30:3113–3126, 2021. doi: 10.1109/TIP.2021.3058783.

Fengyang Xiao, Sujie Hu, Yuqi Shen, Chengyu Fang, Jinfa Huang, Longxiang Tang, Ziyun Yang, Xiu Li, and Chunming He. A survey of camouflaged object detection and beyond. *CAAI Artificial Intelligence Research*, 3:9150044, 2024. doi: 10.26599/AIR.2024.9150044. URL https://www.sciopen.com/article/10.26599/AIR.2024.9150044.

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 & Machine Intelligence*, 46(12):10362–10374, December 2024. ISSN 1939-3539. doi: 10.1109/TPAMI.2024.3438565. URL https://doi.ieeecomputersociety.org/10.1109/TPAMI.2024.3438565.

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, pp. 5323–5332, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3548178. URL https://doi.org/10.1145/3503161.3548178.

Xuying Zhang, Bowen Yin, Zheng Lin, Qibin Hou, Deng-Ping Fan, and Ming-Ming Cheng. Referring camouflaged object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5): 3597–3610, 2025. doi: 10.1109/TPAMI.2025.3532440.

Jianwei Zheng, Hao Liu, Yuchao Feng, Jinshan Xu, and Liang Zhao. CASF-Net: Cross-attention and cross-scale fusion network for medical image segmentation. *Computer Methods and Programs in Biomedicine*, 229:107307, 2023. ISSN 0169-2607. doi: https://doi.org/10.1016/j.cmpb.2022.107307. URL https://www.sciencedirect.com/science/article/pii/S0169260722006885.

Yijie Zhong, Bo Li, Lv Tang, Senyun Kuang, Shuang Wu, and Shouhong Ding. Detecting camouflaged object in frequency domain. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4494–4503, 2022. doi: 10.1109/CVPR52688.2022.00446.

Mingchen Zhuge, Deng-Ping Fan, Nian Liu, Dingwen Zhang, Dong Xu, and Ling Shao. Salient Object Detection via Integrity Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45 (3):3738–3752, 2023. doi: 10.1109/TPAMI.2022.3179526.

A Additional Ablative Studies

In this section, we present the additional ablation studies on the supervision strategy followed.

A.1 Ablation on Supervision

Table 6: Ablation study on supervision strategy.

No.	Supervision Strategy	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
1	Single Supervision (Eq. 11)	0.860	0.927	0.775	0.024
2	Multi-Stage Supervision (Eq. 10)	0.861	0.928	0.778	0.024

Equation 10 shows our supervision strategy, where supervision is applied to mask predictions computed at multiple scales using multiscale features from multiple stages of the Camo-Image Encoder. We conducted an ablation study on the supervision to compare our chosen strategy with an alternative strategy, where supervision is applied only to the final segmentation mask as described below:

$$\mathcal{L}(P, G) = \mathcal{L}_{bce}(P, G) + \mathcal{L}_{iou}(P, G), \tag{11}$$

where P refers to the predicted segmentation mask M^{seg}, and G is the ground truth mask. Table 6 shows that our chosen supervision strategy yields a marginal improvement over the alternative strategy.