Unveiling Camouflage: A Learnable Fourier-based Augmentation for Camouflaged Object Detection and Instance Segmentation

Minh-Quan Le[1, 2, 3*]**, Minh-Triet Tran**[1, 2, 4*]**, Trung-Nghia Le**[1, 2]**, Tam V. Nguyen**[5]**, Thanh-Toan Do**[6]

[1] University of Science, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Stony Brook University, USA
[4] John von Neumann Institute, Ho Chi Minh City, Vietnam
[5] University of Dayton, USA
[6] Monash University, Australia

Abstract

Camouflaged object detection (COD) and camouflaged instance segmentation (CIS) aim to recognize and segment objects that are blended into their surroundings, respectively. While several deep neural network models have been proposed to tackle those tasks, augmentation methods for COD and CIS have not been thoroughly explored. Augmentation strategies can help improve the performance of models by increasing the size and diversity of the training data and exposing the model to a wider range of variations in the data. Besides, we aim to automatically learn transformations that help to reveal the underlying structure of camouflaged objects and allow the model to learn to better identify and segment camouflaged objects. To achieve this, we propose a learnable augmentation method in the frequency domain for COD and CIS via Fourier transform approach, dubbed **CamoFourier**. Our method leverages a conditional generative adversarial network and cross-attention mechanism to generate a reference image and an adaptive hybrid swapping with parameters to mix the low-frequency component of the reference image and the high-frequency component of the input image. This approach aims to make camouflaged objects more visible for detection and segmentation models. Without bells and whistles, our proposed augmentation method boosts the performance of camouflaged object detectors and camouflaged instance segmenters by large margins.

Introduction

Camouflage refers to the use of any combination of materials, coloration, or illumination for concealment, either by making animals or objects hard to see or by disguising them as something else. In the context of computer vision, camouflaged object detection (COD) and camouflaged instance segmentation (CIS) are tasks that aim to recognize and segment objects that are integrated into their surroundings. COD is the task of identifying objects that are "seamlessly" embedded in their surroundings (Le et al. 2019). The high intrinsic similarities between the target object and the background make COD far more challenging than traditional object detection tasks(Fan et al. 2020a). CIS is a related task that involves not only detecting camouflaged objects but also segmenting them at the pixel level (Le et al. 2022). Both COD and CIS require
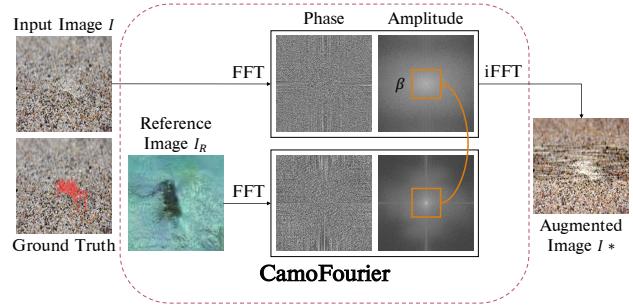
---

*These authors contributed equally.

Figure 1: Our CamoFourier not only preserves the spatial structure and resolution of an image but also highlights the underlying structure of camouflaged objects for better identification and segmentation.

sophisticated computer vision algorithms that can distinguish between the target object and its background despite their high visual similarity. These tasks have many potential applications, including wildlife monitoring, search and rescue operations (Le et al. 2019), medical diagnosis (polyp segmentation (Fan et al. 2020b); COVID-19 infection identification from lung x-rays (Das et al. 2022)).

The performance of COD has recently been elevated by convolutional neural networks (CNNs)-based approaches. Current cutting-edge methods concentrate on designing detector architectures based on attention (Pang et al. 2022; Fan et al. 2020a; Jia et al. 2022), and specific features of camouflaged objects (Fan et al. 2022; Zhai et al. 2021). Moreover, OSFormer (Pei et al. 2022) leverages ViT (Dosovitskiy et al. 2021) to propose the first architecture for the CIS task. While several architectures have been designed to address COD and CIS tasks, augmentation strategies have not been completely investigated even though they can help improve the model's ability to generalize to new data and increase its robustness to variations in the input. Therefore, this work explores the data-centric approach to COD and CIS problems by proposing a learnable augmentation method in the frequency domain, named CamoFourier, which allows deep learning models to learn how to augment their own training data. In this way, our learnable augmentation aims to reveal the underlying structure of camouflaged objects and allow the model to learn to better identify and segment camouflaged objects.

Although several augmentation methods have been proposed for generic object detection (Zhang et al. 2018; Yun et al. 2019; Zoph et al. 2020) and instance segmentation (Ghiasi et al. 2021)) tasks, these works could not be directly applied to COD and CIS tasks as these methods introduce occlusions and deformations to input images, which may make camouflaged objects even harder to detect and segment (Zhang et al. 2018). Moreover, these methods do not consider the intrinsic similarities between the target object and the background (Ghiasi et al. 2021), the main challenges for COD and CIS. Therefore, we introduce a learnable augmentation strategy in the frequency domain via Fourier transform for COD and CIS. The Fourier transform of an image can encode semantic information and appearance information of the image in its phase and amplitude components, respectively. Changing the amplitude information while keeping the phase information of the Fourier transform of an image can change the appearance of the image, but it is still recognizable. Additionally, Fourier transform can also preserve the spatial structure and resolution of an image, see Fig. 1.

Motivated by the above properties of Fourier transform, we propose a learnable augmentation approach that aims to manipulate the amplitude information of the Fourier transform of an input image to enhance the visibility of camouflaged objects in the image. In particular, our method leverages a conditional generative adversarial network to synthesize a generated image $\mathcal{I}_G$ and cross-attention mechanism to learn the spatial correspondence and alignment between the input image $\mathcal{I}$ and the generated one $\mathcal{I}_G$, which outputs a reference image $\mathcal{I}_R$. Furthermore, we propose a hybrid swapping with an adaptive parameter to mix the low-frequency component of the reference image and the high-frequency component of the input image to control the amount of texture and color information that is transferred from the reference image $\mathcal{I}_R$ to the input image $\mathcal{I}$. To the best of our knowledge, our work is the first one that designs a trainable Fourier-based augmentation specifically for both COD and CIS tasks. In summary, our main contributions are three-fold.

- We propose a novel learnable Fourier-based augmentation for camouflaged object detection and instance segmentation, highlighting the camouflaged object of interest from the background and making it more visible for deep models. The augmentation framework is flexible and can be adapted to different segmentation or detection algorithms.

- We introduce an adaptive hybrid swapping approach that can control the amount of texture and color information that is transferred from the reference image to the input image. We leverage a simple yet effective cross-attention mechanism between the input and generated image so that the reference image is able to transfer its texture and color information to the input image in a more attentive manner.

- Extensive experimental results on a variety of camouflaged object detection and instance segmentation methods on different datasets show that the proposed method significantly improves the performance of existing models.

Related Work

Camouflaged object detection. Fan et al. (2020a) present Search and Identification Net (SINet), a strong baseline for the COD task. There is also an enhanced version of SINet called SINetV2 (Fan et al. 2022) with two well-elaborated sub-modules: neighbor connection decoder (NCD) and group-reversal attention (GRA). To well involve the frequency clues into the CNN models, Zhong et al. (2022) introduce the frequency domain as an additional clue to better detect camouflaged objects from backgrounds. Several sophisticated architectures have been proposed for COD, our CamoFourier is a learnable augmentation strategy for COD that can be plugged into these existing works and improve their performance in an efficient and easy implementation.

Camouflaged instance segmentation. Le et al. (2022) investigate the interesting yet challenging problem of camouflaged instance segmentation. To promote the new task of camouflaged instance segmentation of in-the-wild images, Le et al. (2022) introduce a dataset, dubbed CAMO++, that extends the preliminary CAMO dataset (camouflaged object segmentation) in terms of quantity and diversity. To detect and segment the whole scope of a camouflaged object, camouflaged object detection is introduced as a binary segmentation task, with the binary ground truth camouflage map indicating the exact regions of the camouflaged objects. Pei et al. present OSFormer (Pei et al. 2022), the first one-stage transformer framework for camouflaged instance segmentation. As the CIS task has not been deeply explored, our CamoFourier is an efficient plug-and-play tool, which helps improve the performance of CIS model and is valuable for researchers working on CIS.

Data augmentation. Data augmentation is a critical component of training deep learning models for object detection and segmentation tasks (Zhang et al. 2018; Yun et al. 2019; Zoph et al. 2020; Ghiasi et al. 2021; Luo et al. 2023). Data augmentation has been shown to significantly improve generic object detection and instance segmentation, but its potential has not been thoroughly investigated for camouflaged object detection and camouflaged instance segmentation. The existing works are augmentation methods in the spatial domain. When images are augmented in the spatial domain, it is possible to introduce occlusions, deformations, or noise to the input images, which may make the camouflaged objects even harder to detect and segment. Moreover, these methods do not consider the intrinsic similarities between the target object and the background, which are the main challenges for camouflaged object detection and camouflaged instance segmentation. Our CamoFourier is an augmentation method in the frequency domain that is specifically designed for COD and CIS and can help to avoid these artifacts by preserving the spatial information in the image.

It is worth noting that Fourier transform has been leveraged for domain generalization (Xu et al. 2021), and domain adaptation (Yang and Soatto 2020). Different from those works, in this paper, we explore Fourier transform to propose a learnable augmentation method for COD and CIS tasks.
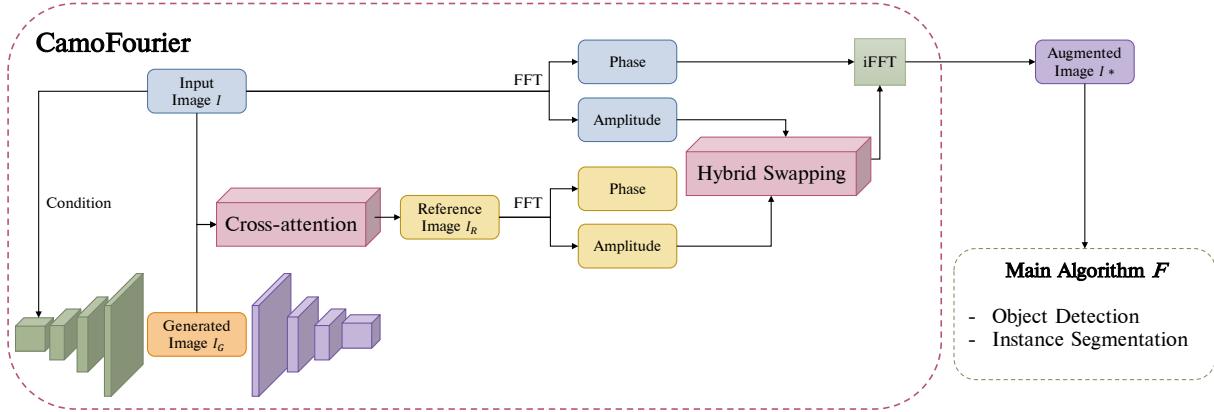
Figure 2: Overview of the proposed CamoFourier. Our method leverages a conditional generative adversarial network and cross-attention mechanism to generate a reference image and an adaptive hybrid swapping with parameters to mix the low-frequency component of the reference image and the high-frequency component of the input image.

Proposed Method

Overview Framework

We propose a general framework (Fig. 2) that can be used to provide an effective and highly accurate solution for any segmentation or detection algorithm for camouflage objects. Specifically, we propose a learnable data augmentation module that can be integrated to be *end-to-end training* with any existing object detection and segmentation methods. In this framework, the original input image $\mathcal{I}$ is transformed to synthesize a new input image $\mathcal{I}*$ that can highlight the camouflaged object of interest from the background and is more suitable for the main processing algorithm $\mathcal{F}$, including segmentation and detection. It is worth noting that the proposed module is trained together with the main algorithm $\mathcal{F}$ in an end-to-end fashion.

As shown in Fig. 2, we aim to synthesize the transformed image $\mathcal{I}*$ from an original input image $\mathcal{I}$ by replacing the amplitude in the FFT of $\mathcal{I}$ with the amplitude of the FFT of a context-aware reference image $\mathcal{I}_R$. The input image $\mathcal{I}$ is used as the conditional information for a generative model to generate an RGB image $\mathcal{I}_G$. Then, both the generated image $\mathcal{I}_G$ and original input image $\mathcal{I}$ are fed into a Cross Attention module to produce a context-aware reference image $\mathcal{I}_R$ which captures information of both $\mathcal{I}$ and $\mathcal{I}_G$.

Our intention is to synthesize $\mathcal{I}_R$ so that its amplitude component is more appropriate than that of the original image $\mathcal{I}$ to differentiate the camouflaged object and its surrounding area. Then, we blend the original image $\mathcal{I}$ with the context-aware reference image $\mathcal{I}_R$ by swapping the amplitude component from the FFT of $\mathcal{I}_R$ with the one of the original image $\mathcal{I}$. We keep the phase component from the FFT of I because we want to preserve the spatial structure and the semantic information of the original input image. We are inspired by the idea of swapping the amplitude component for style transfer (Yang and Soatto 2020) and we adopt this idea into our proposed framework. However, in our solution, instead of using a target style for reference as in (Yang and Soatto 2020), we train our framework with data to generate a reference image $\mathcal{I}_R$ that provides a better FFT amplitude component

for camouflaged object detection or segmentation.

We do not simply swap entirely the amplitude from a reference image $\mathcal{I}_R$ to the original image $\mathcal{I}$ but we propose an adaptive hybrid swapping approach. We reuse the amplitude at high frequencies in the FFT of the original image $\mathcal{I}$ because the high-frequency components are responsible for the texture and detail of an image and adopt the amplitude at low frequencies in the FFT of the reference image $\mathcal{I}_R$ for transferring the texture and color information from the reference image.

Furthermore, we do not set a fixed threshold frequency for the hybrid swapping step but we decide to train an adaptive threshold for hybrid swapping. In this way, the amplitude swapping step can exploit adaptively to the original input image $\mathcal{I}$.

Learnable Fourier-based Augmentation

GAN module. Let G be the generator network and D be the discriminator network. The generator G takes an input image $\mathcal{I}$ and a random noise vector $\mathbf{z}$ as inputs and outputs $\hat{\mathcal{I}}_G = G(\mathcal{I}, \mathbf{z})$, which is an image that is supposed to look realistic given $\mathcal{I}$. The discriminator D takes the concatenation of $(\mathcal{I}, \mathcal{I}_G)$ as input and outputs $D(\mathcal{I}, \mathcal{I}_G)$, which is a scalar value indicating how likely $\mathcal{I}_G$ is to be a realistic image given $\mathcal{I}$. The objective function of the conditional generative adversarial network (cGAN) is given by:

$$\begin{aligned}\mathcal{L}_{cGAN}(G, D) &= \mathbb{E}_{\mathcal{I} \sim p_{\text{data}}(\mathcal{I})}[\log D(\mathcal{I}, \mathcal{I})] \\ &+ \mathbb{E}_{\mathcal{I} \sim p_{\text{data}}(\mathcal{I}), \mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}\left[\log(1 - D(\mathcal{I}, G(\mathcal{I}, \mathbf{z})))\right].\end{aligned} \quad (1)$$

Similar to the cycle-consistency loss used in some image-to-image translation models (Isola et al. 2017), we use a conditional generative adversarial network with an additional $L1$ loss term to measure the similarity between the generated image $\mathcal{I}_G$ and the input image $\mathcal{I}$. $L1$ loss can be written as:

$$\mathcal{L}_{L1}(G) = \mathbb{E}_{\mathcal{I} \sim p_{\text{data}}(\mathcal{I}), \mathbf{z} \sim p_{\mathbf{z}}(\mathbf{z})}\left[\|\mathcal{I} - G(\mathcal{I}, \mathbf{z})\|_1\right], \quad (2)$$

where $\|\cdot\|_1$ denotes the $L1$ norm. The total objective function of your model can be written as:

$$\min_{G} \max_{D} \mathcal{L}_{cGAN}(G, D) + \lambda \mathcal{L}_{L1}(G), \quad (3)$$

where λ is a hyperparameter that controls the weight of the $L1$ loss. By adding this term, we encourage the generator to produce images that are not only realistic but also close to the input in the spatial domain (RGB domain).

Basic swapping of amplitude. After transforming both input image $\mathcal{I}$ and reference image $\mathcal{I}_R$ (produced the Cross Attention module) to the frequency domain via FFT, we obtain the phase and amplitude of each image. Then we swap the amplitude of both images. The phase of the input image together with the amplitude of the reference image is transformed back to the spatial domain via inverse FFT (iFFT) to get $\mathcal{I}*$. By swapping the amplitude of the input and reference images, we transfer the texture and color information from the reference image to the input image, while preserving the shape and structure information from the input image.

Let $\mathcal{F}(\mathcal{I})$ and $\mathcal{F}(\mathcal{I}_R)$ be the Fourier transforms of $\mathcal{I}$ and $\mathcal{I}_R$, respectively. We define $\mathcal{A}(\mathcal{F}(\mathcal{I}))$ and $\phi(\mathcal{F}(\mathcal{I}))$ to be the amplitude and phase of $\mathcal{F}(\mathcal{I})$. Similarly, the amplitude and phase of $\mathcal{F}(\mathcal{I}_R)$ are denoted by $\mathcal{A}(\mathcal{F}(\mathcal{I}_R))$ and $\phi(\mathcal{F}(\mathcal{I}_R))$. The inverse Fourier transform of the amplitude-swapped image is given by:

$$\mathcal{I}* = \mathcal{F}^{-1}(\mathcal{A}(\mathcal{F}(\mathcal{I}_R)), \phi(\mathcal{F}(\mathcal{I}))). \tag{4}$$

The image $\mathcal{I}*$ is the output of the amplitude swapping operation. It is the input image $\mathcal{I}$ with the texture and color information from the reference image $\mathcal{I}_R$.

Cross-Attention Module

We hypothesize that the reference image $\mathcal{I}_R$ must capture information from the input image $\mathcal{I}$ so that the amplitude of the reference image can pay attention to specific regions in the input data. Cross-attention (Chen, Hsieh, and Liu 2021; Rombach et al. 2022) between an input image and a generated image can help to model the spatial correspondence of different source images. This can help to extract appropriate features and achieve adaptive and balanced fusion. It can dynamically learn the spatial correspondence to derive better alignment of essential details from the two images $\mathcal{I}$ and $\mathcal{I}_G$.

The input to the Cross Attention module is an input image $\mathcal{I}$ and a generated image $\mathcal{I}_G$, both of size $H \times W \times C$, where $H \times W$ is the spatial dimension and C is the number of channels (e.g., 3 for RGB images). Inspired by Vision Transformer (Dosovitskiy et al. 2021), first, we split both images into patches of size $P \times P \times C$, resulting in $(H/P) \times (W/P)$ patches for each image. Each patch is then flattened into a vector of size P^2C and linearly projected into a token of size D using a learnable projection matrix W_e of size $P^2C \times D$. This can be expressed mathematically as:

$$\mathcal{I}^t = I^p * W_e \text{ and } \mathcal{I}_G^t = \mathcal{I}_G^p * W_e, \tag{5}$$

where $\mathcal{I}^t$ and $\mathcal{I}_G^t$ are the token representations of the input and generated images; I^p and $\mathcal{I}_G^p$ are the patch representations of the input and generated images, respectively.

Next, we apply cross-attention between the two sets of tokens to allow the reference image to pay attention to specific regions of the input image. This can be expressed mathematically using the scaled dot-product attention mechanism:

$$A = \text{softmax}((\mathcal{I}^t * W_q) * (\mathcal{I}_G^t * W_k)^T / \sqrt{D})$$
$$O = A * (\mathcal{I}_G^t * W_v), \tag{6}$$

where W_q, W_k, and W_v are learnable projection matrices of size $D \times D$ for the query, key, and value representations, respectively, and A is the attention matrix that represents the attention paid by the generated image to each patch of the input image. The output O is then linearly projected via a learnable projection matrix W_d of size $D \times P^2C$ and reshaped into an image of size $H \times W \times C$ in the RGB domain. To ensure that the output O remains in the RGB domain with pixel values in the range $[0, 255]$, we apply a scaling and clipping operation to the output before reshaping it into an image. This can be done by first normalizing the output to have zero mean and unit variance, then scaling it by a factor of 255, and finally clipping the values to the range $[0, 255]$. The final result of the cross-attention module is a reference image $\mathcal{I}_R$.

Adaptive Hybrid Swapping of Amplitude

Inspired by the hybrid image (Oliva, Torralba, and Schyns 2006) which mixes the low-spatial frequencies of one image with the high spatial frequencies of another image, we propose an adaptive hybrid swapping the amplitude of the input image $\mathcal{I}$ and the reference image $\mathcal{I}_R$. We transform the input image $\mathcal{I}$ and the reference image $\mathcal{I}_R$ via FFT. The low-frequency part of the amplitude of the input image is adaptively replaced by that of the reference image. We denote $\mathbb{M}_\beta$ as a binary mask matrix, whose value is one in the center region and zero in the rest where $\beta \in (0,1)$:

$$\mathbb{M}_\beta(h,w) = \mathbb{1}_{(h,w) \in [-\beta H : \beta H, -\beta W : \beta W]}. \tag{7}$$

The inverse Fourier transform of the adaptive hybrid swapping of amplitude is given by:

$$\mathcal{I}* = \mathcal{F}^{-1}\Big(\big[\mathbb{M}_\beta \odot \mathcal{A}(\mathcal{F}(\mathcal{I}_R)) + (1 - \mathbb{M}_\beta) \odot \mathcal{A}(\mathcal{F}(\mathcal{I})) \big], \phi(\mathcal{F}(\mathcal{I})) \Big). \tag{8}$$

The parameter β is optimized via Bayesian optimization. As a consequence, the augmented image $\mathcal{I}*$ is the output of the adaptive hybrid swapping operation. The benefit of adaptive hybrid swapping is that it is able to control the amount of texture and color information that is transferred from the reference image $\mathcal{I}_R$ to the input image $\mathcal{I}$. The augmented image $\mathcal{I}*$ is following fed as input to camouflaged object detection and instance segmentation models for training with specific objective functions of COD and CIS tasks.

Experiments

Experimental Settings

Datasets. For COD, we use 3 popular benchmarks: CAMO (Le et al. 2019), COD10K (Fan et al. 2020a), and NC4K (Lv et al. 2021). Among them, CAMO consists of 2,500 images, half with camouflaged objects and half without. COD10K (Fan et al. 2020a) contains 5,066 camouflaged, 3,000 background, and 1,934 non-camouflaged images. NC4K is a large-scale COD dataset with 4,121 images. We follow prior studies (Fan et al. 2020a; Lv et al. 2021) and use 1,000 images from CAMO (Le et al. 2019), 3,040 images from COD10K(Fan et al. 2020a), and all images from NC4K (Lv et al. 2021) for testing.

For CIS, there are few task-specific datasets since this is a new and challenging task. Fan et al. (2020a) provided a COD

Table 1: The effectiveness of our CamoFourier in improving camouflaged object detection models on three benchmarks: CAMO (Le et al. 2019), COD10K (Fan et al. 2020a), and NC4K (Lv et al. 2021). Our CamoFourier drastically increases the performance of cutting-edge methods of COD, improved results are highlighted in blue color.

Method	COD10K-Test				NC4K-Test				CAMO-Test			
	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
SINet (Fan et al. 2020a)	0.808	0.883	0.723	0.058	0.776	0.867	0.631	0.043	0.745	0.825	0.644	0.092
SINet (Fan et al. 2020a) + **CamoFourier**	0.820	0.894	0.746	0.055	0.782	0.881	0.652	0.040	0.773	0.861	0.665	0.087
SINetV2 (Fan et al. 2022)	0.847	0.898	0.770	0.048	0.815	0.863	0.680	0.037	0.820	0.875	0.743	0.070
SINetV2 (Fan et al. 2022) + **CamoFourier**	0.859	0.914	0.778	0.043	0.826	0.890	0.704	0.032	0.858	0.893	0.752	0.066
SegMaR (Jia et al. 2022)	0.841	0.905	0.781	0.046	0.833	0.895	0.724	0.033	0.815	0.872	0.742	0.071
SegMaR (Jia et al. 2022) + **CamoFourier**	0.859	0.927	0.803	0.045	0.862	0.916	0.744	0.029	0.828	0.894	0.771	0.068
ZoomNet (Pang et al. 2022)	0.853	0.907	0.784	0.043	0.838	0.893	0.729	0.029	0.820	0.883	0.752	0.066
ZoomNet (Pang et al. 2022) + **CamoFourier**	0.872	0.923	0.801	0.037	0.864	0.911	0.740	0.025	0.852	0.906	0.774	0.060
FDNet (Zhong et al. 2022)	0.834	0.895	0.750	0.052	0.837	0.897	0.731	0.030	0.844	0.903	0.778	0.062
FDNet (Zhong et al. 2022) + **CamoFourier**	0.857	0.926	0.794	0.048	0.840	0.918	0.752	0.029	0.863	0.927	0.790	0.055

Table 2: Effectiveness of our CamoFourier in improving CIS model OSFormer (Pei et al. 2022) and generic instance segmenters on two benchmarks: COD10K (Fan et al. 2020a) and NC4K (Lv et al. 2021). Our CamoFourier significantly boosts the performance of state-of-the-art methods and improved results are highlighted in blue color.

Method	COD10K-Test			NC4K-Test		
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
Mask R-CNN (He et al. 2017)	25.00	55.50	20.40	27.70	58.60	22.70
Mask R-CNN (He et al. 2017) + **CamoFourier**	28.74	59.13	22.86	30.46	62.72	25.09
Cascade R-CNN (Cai and Vasconcelos 2018)	25.30	56.10	21.30	29.50	60.80	24.80
Cascade R-CNN (Cai and Vasconcelos 2018) + **CamoFourier**	28.65	58.90	22.93	31.25	63.14	26.38
YOLACT (Bolya et al. 2019)	24.30	53.30	19.70	32.10	65.30	27.90
YOLACT (Bolya et al. 2019) + **CamoFourier**	27.65	57.40	22.17	34.97	69.22	30.84
SOTR (Guo et al. 2021)	27.90	58.70	24.10	29.30	61.00	25.60
SOTR (Guo et al. 2021) + **CamoFourier**	30.23	61.57	25.93	32.64	65.11	28.87
OSFormer (Pei et al. 2022)	41.00	71.10	40.80	42.50	72.50	42.30
OSFormer (Pei et al. 2022) + **CamoFourier**	43.52	74.84	42.65	44.95	75.67	44.28

Table 3: Comparisons between our CamoFourier and existing generic augmentation methods in camouflaged object detection task. Cutting-edge augmentations show ineffectiveness when applying to COD while our CamoFourier boosts the performance of SINetV2 (Fan et al. 2022) significantly.

Method	Augmentation	COD10K-Test				NC4K-Test				CAMO-Test			
		$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
SINetV2 (Fan et al. 2022)	-	0.847	0.898	0.770	0.048	0.815	0.863	0.680	0.037	0.820	0.875	0.743	0.070
	Mixup (Zhang et al. 2018)	0.813	0.856	0.745	0.067	0.764	0.829	0.636	0.042	0.784	0.817	0.694	0.088
	Cutmix (Yun et al. 2019)	0.806	0.838	0.723	0.071	0.752	0.816	0.630	0.056	0.771	0.805	0.683	0.079
	AutoAugment (Zoph et al. 2020)	0.839	0.903	0.766	0.051	0.812	0.858	0.675	0.040	0.823	0.872	0.746	0.074
	Copy-Paste (Ghiasi et al. 2021)	0.820	0.861	0.754	0.055	0.806	0.849	0.673	0.042	0.811	0.864	0.727	0.080
	CamDiff (Luo et al. 2023)	0.851	0.895	0.772	0.047	0.819	0.866	0.678	0.037	0.831	0.882	0.745	0.073
	CamoFourier	**0.859**	**0.914**	**0.778**	**0.043**	**0.826**	**0.890**	**0.704**	**0.032**	**0.858**	**0.893**	**0.752**	**0.066**

dataset called COD10K that also has instance-level annotations for CIS models. COD10K has 2,026 images for testing and 3,040 camouflaged images with instance-level labels for training. Recently, Le et al. (2022) released the CAMO++ dataset, a larger CIS dataset with 5,500 samples with hierarchical pixel-wise annotation. Lyu et al. (2021) introduced the NC4K test set for CIS, which has 4,121 images.

Evaluation metrics. For COD, we use four common metrics: Structure-measure (S_m) (Fan et al. 2017), mean absolute error (M) (Perazzi et al. 2012), weighted F-measure (wF) (Margolin, Zelnik-Manor, and Tal 2014), and adaptive E-measure (αE) (Fan et al. 2018). S_m measures the structural similarity between predictions and ground truth, taking into account both region- and object-awareness. M measures the pixel-level accuracy between a predicted map and ground truth and is widely used in salient object detection (SOD)

tasks. wF measures the balance between recall and precision. αE measures the human visual perception-based alignment between predictions and ground truth at both pixel and image levels and is suitable for evaluating the global and local accuracy of camouflaged object detection results. Regarding CIS, we adopt COCO-style (Lin et al. 2014) evaluation metrics such as AP_{50}, AP_{75}, and AP. Unlike the mAP metric used in instance segmentation, we do not consider the class labels of camouflaged instances, since they are class-agnostic. We only need to consider the existence of camouflaged instances and ignore the class mean value.

Implementation Details

In our CamoFourier, the GAN architecture consists of generator and discriminator components. We adopt a simple UNet encoder-decoder architecture for the generator. However, we

Table 4: Comparisons between our CamoFourier and existing generic augmentation methods in CIS task. State-of-the-art generic augmentations downgrade the effectiveness of OSFormer (Pei et al. 2022) while our CamoFourier improves the performance of CIS model drastically.

Method	Augmentation	COD10K-Test			NC4K-Test		
		AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
OSFormer (Pei et al. 2022)	-	41.00	71.10	40.80	42.50	72.50	42.30
	Mixup (Zhang et al. 2018)	39.65	68.47	38.59	39.21	68.10	38.72
	Cutmix (Yun et al. 2019)	37.85	66.32	37.14	38.96	67.18	38.25
	AutoAugment (Zoph et al. 2020)	41.30	71.48	39.26	42.74	72.31	42.55
	Copy-Paste (Ghiasi et al. 2021)	39.82	71.25	38.63	41.90	71.58	41.43
	CamDiff (Luo et al. 2023)	42.24	72.57	41.39	43.25	72.86	42.18
	CamoFourier	**43.52**	**74.84**	**42.65**	**44.95**	**75.67**	**44.28**

Table 5: Our ablation study in the task of camouflaged object detection. CamoFourier with a cross-attention mechanism and adaptive hybrid swapping achieves the best performance when plugged into SINetV2 (Fan et al. 2022).

Method	CamoFourier		COD10K-Test				NC4K-Test			
	Cross-attention	Hybrid swapping	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
SINetV2 (Fan et al. 2022)	✗	✗	0.852	0.908	0.773	0.045	0.821	0.882	0.695	0.034
	✓	✗	0.857	0.911	0.776	0.044	0.824	0.888	0.697	0.033
	✗	✓	0.853	0.910	0.774	0.045	0.825	0.884	0.699	0.033
	✓	✓	**0.859**	**0.914**	**0.778**	**0.043**	**0.826**	**0.890**	**0.704**	**0.032**

Table 6: Our ablation study in the task of camouflaged instance segmentation. CamoFourier with cross-attention mechanism and adaptive hybrid swapping achieves the best performance when plugged into OSFormer (Pei et al. 2022).

Method	CamoFourier		COD10K-Test			NC4K-Test		
	Cross-attention	Hybrid swapping	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
OSFormer (Pei et al. 2022)	✗	✗	42.41	72.86	42.04	43.87	73.66	42.95
	✓	✗	42.96	74.15	42.39	44.58	75.21	44.01
	✗	✓	42.75	73.84	42.26	44.13	74.28	43.72
	✓	✓	**43.52**	**74.84**	**42.65**	**44.95**	**75.67**	**44.28**

remove all skip connections between the encoder and the decoder. Regarding the discriminator, we leverage 70×70 PatchGAN (Isola et al. 2017) to classify if each $N \times N$ patch in an image is real or fake. The Fast Fourier Transform (FFT) and iFFF are implemented using the PyTorch FFT package. Our CamoFourier is plugged into several camouflaged object detectors and instance segmenters for end-to-end training. The parameter β in adaptive hybrid swapping is optimized by Bayesian optimization in range $(0, 1)$ with step size 0.01. Original images are resized to 512×512 and transformed using CamoFourier then fed into specific architectures of COD and CIS. We adopt the Adam optimizer with a learning rate of 10^{-4}, $\beta_1 = 0.5$, and $\beta_2 = 0.999$. All variants of CamoFourier are trained with a batch size of 8 on a single NVIDIA RTX 3090 GPU.

Comparisons with State-of-the-art Methods

We validate the effectiveness of our learnable augmentation in the frequency domain by integrating our augmentation strategy into several camouflaged object detection and camouflaged instance segmentation models.

Camouflaged object detection. We integrate the proposed CamoFourier into 5 different state-of-the-art camouflaged object detection models including: SINet (Fan et al. 2020a), SINetV2 (Fan et al. 2022), SegMaR (Jia et al. 2022), Zoom-Net (Pang et al. 2022), and FDNet (Zhong et al. 2022). The experimental results presented in Table 1 show that our Camo-Fourier improves the performance of the 5 methods on all considered metrics.

Camouflaged instance segmentation. The CIS task has not been well-studied as only one CIS model (OSFormer (Pei et al. 2022)) has been proposed to date. Therefore, we also conduct an experiment on several popular generic instance segmentation models for more comprehensive evaluation. Following the experiments of OSFormer (Pei et al. 2022), we train those instance segmentation models on the instance-level COD10K training set and test them on the COD10K (Fan et al. 2020a) and NC4K (Lv et al. 2021) test sets. Table 2 shows that our learnable augmentation improves the performance of the instance segmentation models in the CIS task on all AP metrics. Specifically, our method increases the AP score of OSFormer (Pei et al. 2022) by a significant margin of 2.52. For the generic instance segmentation model Mask R-CNN (He et al. 2017), our augmentation method boosts the AP score by an even higher margin of 3.74.

Comparisons with state-of-the-art augmentation methods. We compare our CamoFourier with CamDiff (Luo et al. 2023) and cutting-edge augmentation methods for generic object detection and instance segmentation including Mixup (Zhang et al. 2018), CutMix (Yun et al. 2019), AutoAugment (Zoph et al. 2020), and Copy-Paste (Ghiasi et al. 2021) on COD and CIS tasks. In particular, we plug the mentioned augmentation methods into SINetV2 (Fan et al. 2022) and OSFormer (Pei et al. 2022) for COD and CIS tasks, respectively. Table 3 and Table 4 indicate that generic augmentation methods are obviously not suitable for camouflaged object detection and instance segmentation; they even make the performance of
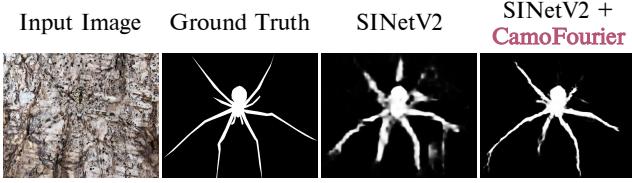
Figure 3: Qualitative comparison of SINetV2 with and without our proposed CamoFourier in COD task.
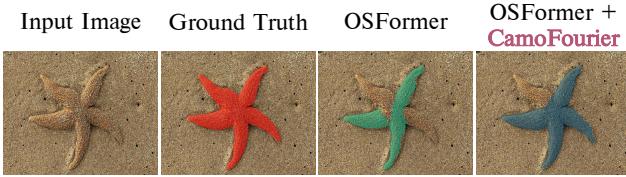

Figure 4: Qualitative comparison of OSFormer with and without our proposed CamoFourier in CIS task.

SINetV2 (Fan et al. 2022) and OSFormer (Pei et al. 2022) deteriorate compared to the ones without augmentation. On the contrary, our proposed CamoFourier helps improve the effectiveness of SINetV2 (Fan et al. 2022) and OSFormer (Pei et al. 2022) by large margins. Our CamoFourier also outperforms CamDiff (Luo et al. 2023), the current state-of-the-art method for camouflaged objects on both COD and CIS tasks.

Qualitative Results

Qualitative evaluation. Figures 3 and 4 visualize the segmentation results of SINetV2 (Fan et al. 2022) and OS-Former (Pei et al. 2022) with and without our CamoFourier, respectively. The results demonstrate that SINetV2 and OS-Former equipped with CamoFourier are able to delineate the camouflaged object in a more accurate way. More qualitative results for COD and CIS tasks are presented in the supplementary materials.

Visualization of augmented images. In Fig. 5, we illustrate the outputs of our CamoFourier, which are augmented images $\mathcal{I}*$. In the CamoFourier framework, the conditional GAN module and a cross-attention mechanism synthesize a reference image $\mathcal{I}_R$. After that, we transform an input image and a reference image into the frequency domain and perform basic swapping or adaptive hybrid swapping of these two amplitudes. We also compare the qualitative results of our CamoFourier with and without adaptive hybrid swapping. Figure 5 shows that our adaptive hybrid swapping is able to control the amount of texture and color information that is transferred from the reference image $\mathcal{I}_R$ to the input image $\mathcal{I}$ thanks to the parameter β.

Ablation Study

Effectiveness of cross-attention module. We examine the effectiveness of our cross-attention module in the Camo-Fourier framework by comparing the performance of COD (SINetV2 (Fan et al. 2022)) and CIS (OSFormer (Pei et al. 2022)) models with and without the cross-attention mechanism. Without the cross-attention mechanism, we directly transform the input image $\mathcal{I}$ and the generated image $\mathcal{I}_G$ from the GAN module into the frequency domain and perform am-
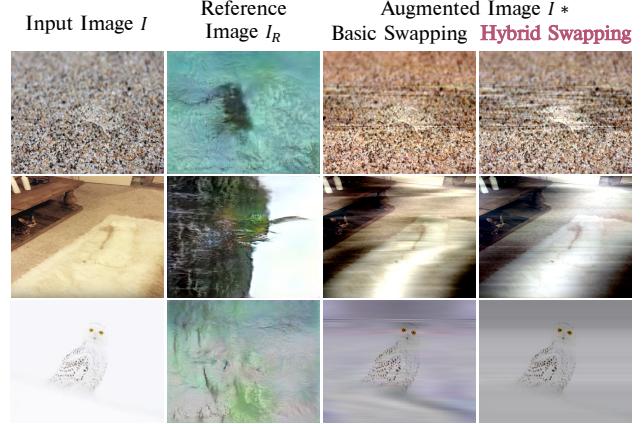

Figure 5: Visualization of augmented images by our Camo-Fourier. Our proposed augmentation highlights the underlying structure of camouflaged objects for better identification and segmentation. We also compare the transformed results of our method with and without adaptive hybrid swapping. Our adaptive hybrid swapping is able to control the amount of texture and color information that is transferred from the reference image to the input image.

plitude swapping. Table 5 shows that the cross-attention module in CamoFourier improves the performance of COD on the COD10K (Fan et al. 2020a) and NC4K (Lv et al. 2021) test sets on four metrics. Table 6 shows that our cross-attention module in CamoFourier also enhances the performance of CIS on the COD10K and NC4K test sets.

Significance of adaptive hybrid swapping. We evaluate the significance of the proposed adaptive hybrid swapping in the CamoFourier augmentation. We also perform experiments on COD and CIS tasks with 2 models: SINetV2 (Fan et al. 2022) for COD, OSFormer (Pei et al. 2022) for CIS. We apply our CamoFourier to these architectures with and without adaptive hybrid swapping. Without using the proposed hybrid swapping, we use the basic swapping of amplitude as described in Section . Tables 5 and 6 show that our adaptive hybrid swapping helps the CamoFourier augmentation to boost the performance of COD and CIS models. More ablation studies on the effectiveness of the cross-attention mechanism and adaptive hybrid swapping on other COD and CIS models are presented in the supplementary material.

Conclusion

In this paper, we propose CamoFourier, a novel learnable Fourier-based augmentation method for camouflaged object detection and instance segmentation. Our method incorporates a generative model and cross-attention mechanism to create a reference image and a Fourier transform to mix the low and high-frequency components of the input and reference images. This makes the camouflaged objects more visible and easier to be detected and segmented. We conduct extensive experiments on various camouflaged object detection and instance segmentation models on different datasets. The experimental results show that our method significantly improves the existing state-of-the-art COD and CIS methods by a large margin.

References

Bolya, D.; Zhou, C.; Xiao, F.; and Lee, Y. J. 2019. YOLACT: Real-Time Instance Segmentation. In *ICCV*.

Cai, Z.; and Vasconcelos, N. 2018. Cascade R-CNN: Delving Into High Quality Object Detection. In *CVPR*.

Chen, D.-J.; Hsieh, H.-Y.; and Liu, T.-L. 2021. Adaptive Image Transformer for One-Shot Object Detection. In *CVPR*, 12247–12256.

Das, N. N.; Kumar, N.; Kaur, M.; Kumar, V.; and Singh, D. 2022. Automated deep transfer learning-based approach for detection of COVID-19 infection in chest X-rays. *Irbm*, 43(2): 114–119.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *ICLR*.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-Measure: A New Way to Evaluate Foreground Maps. In *ICCV*.

Fan, D.-P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.-M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*, 698–704.

Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2022. Concealed Object Detection. *IEEE T-PAMI*, 44(10): 6024–6042.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged Object Detection. In *CVPR*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. PraNet: Parallel Reverse Attention Network for Polyp Segmentation. In Martel, A. L.; Abolmaesumi, P.; Stoyanov, D.; Mateus, D.; Zuluaga, M. A.; Zhou, S. K.; Racoceanu, D.; and Joskowicz, L., eds., *MICCAI*, 263–273. Cham.

Ghiasi, G.; Cui, Y.; Srinivas, A.; Qian, R.; Lin, T.-Y.; Cubuk, E. D.; Le, Q. V.; and Zoph, B. 2021. Simple Copy-Paste Is a Strong Data Augmentation Method for Instance Segmentation. In *CVPR*, 2918–2928.

Guo, R.; Niu, D.; Qu, L.; and Li, Z. 2021. SOTR: Segmenting Objects With Transformers. In *ICCV*, 7157–7166.

He, K.; Gkioxari, G.; Dollar, P.; and Girshick, R. 2017. Mask R-CNN. In *ICCV*.

Isola, P.; Zhu, J.-Y.; Zhou, T.; and Efros, A. A. 2017. Image-To-Image Translation With Conditional Adversarial Networks. In *CVPR*.

Jia, Q.; Yao, S.; Liu, Y.; Fan, X.; Liu, R.; and Luo, Z. 2022. Segment, Magnify and Reiterate: Detecting Camouflaged Objects the Hard Way. In *CVPR*, 4713–4722.

Le, T.-N.; Cao, Y.; Nguyen, T.-C.; Le, M.-Q.; Nguyen, K.-D.; Do, T.-T.; Tran, M.-T.; and Nguyen, T. V. 2022. Camouflaged Instance Segmentation In-The-Wild: Dataset, Method, and Benchmark Suite. *IEEE T-IP*, 31: 287–300.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *CVIU*, 184: 45–56.

Lin, T.-Y.; Maire, M.; Belongie, S. J.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft COCO: Common Objects in Context. In *ECCV*.

Luo, X.-J.; Wang, S.; Wu, Z.; Sakaridis, C.; Cheng, Y.; Fan, D.-P.; and Gool, L. V. 2023. CamDiff: Camouflage Image Augmentation via Diffusion Model. arXiv:2304.05469.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *CVPR*, 11591–11601.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to Evaluate Foreground Maps. In *CVPR*, 248–255.

Oliva, A.; Torralba, A.; and Schyns, P. G. 2006. Hybrid Images. In *ACM SIGGRAPH*, 527–532. ISBN 1595933646.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022. Zoom in and Out: A Mixed-Scale Triplet Network for Camouflaged Object Detection. In *CVPR*, 2160–2170.

Pei, J.; Cheng, T.; Fan, D.-P.; Tang, H.; Chen, C.; and Van Gool, L. 2022. OSFormer: One-Stage Camouflaged Instance Segmentation with Transformers. In *ECCV*.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 733–740.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-Resolution Image Synthesis With Latent Diffusion Models. In *CVPR*.

Xu, Q.; Zhang, R.; Zhang, Y.; Wang, Y.; and Tian, Q. 2021. A Fourier-Based Framework for Domain Generalization. In *CVPR*, 14383–14392.

Yang, Y.; and Soatto, S. 2020. FDA: Fourier Domain Adaptation for Semantic Segmentation. In *CVPR*.

Yun, S.; Han, D.; Oh, S. J.; Chun, S.; Choe, J.; and Yoo, Y. 2019. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *ICCV*, 6023–6032.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D.-P. 2021. Mutual Graph Learning for Camouflaged Object Detection. In *CVPR*, 12997–13007.

Zhang, H.; Cisse, M.; Dauphin, Y. N.; and Lopez-Paz, D. 2018. mixup: Beyond Empirical Risk Minimization. In *ICLR*.

Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; and Ding, S. 2022. Detecting Camouflaged Object in Frequency Domain. In *CVPR*, 4504–4513.

Zoph, B.; Cubuk, E. D.; Ghiasi, G.; Lin, T.-Y.; Shlens, J.; and Le, Q. V. 2020. Learning Data Augmentation Strategies for Object Detection. In Vedaldi, A.; Bischof, H.; Brox, T.; and Frahm, J.-M., eds., *ECCV*, 566–583. Cham.

Unveiling Camouflage: A Learnable Fourier-based Augmentation for Camouflaged Object Detection and Instance Segmentation
—Supplementary Material—

Minh-Quan Le[1, 2, 3*]**, Minh-Triet Tran**[1, 2, 4*]**, Trung-Nghia Le**[1, 2]**, Tam V. Nguyen**[5]**, Thanh-Toan Do**[6]

[1] University of Science, Ho Chi Minh City, Vietnam
[2] Vietnam National University, Ho Chi Minh City, Vietnam
[3] Stony Brook University, USA
[4] John von Neumann Institute, Ho Chi Minh City, Vietnam
[5] University of Dayton, USA
[6] Monash University, Australia

Comprehensive Experiments

Ablation Study

Table 1: Our ablation study in the task of camouflaged object detection. CamoFourier with a cross-attention mechanism and adaptive hybrid swapping achieves the best performance when plugged into SINetV2 (**?**) and ZoomNet (**?**).

Method	CamoFourier		COD10K-Test				NC4K-Test			
	Cross-attention	Hybrid swapping	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$	$S_m \uparrow$	$\alpha E \uparrow$	$wF \uparrow$	$M \downarrow$
SINetV2 (**?**)	✗	✗	0.852	0.908	0.773	0.045	0.821	0.882	0.695	0.034
	✓	✗	0.857	0.911	0.776	0.044	0.824	0.888	0.697	0.033
	✗	✓	0.853	0.910	0.774	0.045	0.825	0.884	0.699	0.033
	✓	✓	**0.859**	**0.914**	**0.778**	**0.043**	**0.826**	**0.890**	**0.704**	**0.032**
ZoomNet (**?**)	✗	✗	0.868	0.915	0.792	0.041	0.851	0.907	0.735	0.027
	✓	✗	0.871	0.920	0.798	0.039	0.860	0.910	0.738	0.026
	✗	✓	0.869	0.917	0.793	0.037	0.855	0.910	0.737	0.027
	✓	✓	**0.872**	**0.923**	**0.801**	**0.037**	**0.864**	**0.911**	**0.740**	**0.025**

Table 2: Our ablation study in the task of camouflaged instance segmentation. CamoFourier with cross-attention mechanism and adaptive hybrid swapping achieves the best performance when plugged into OSFormer (**?**) and Mask R-CNN (**?**).

Method	CamoFourier		COD10K-Test			NC4K-Test		
	Cross-attention	Hybrid swapping	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}
OSFormer (**?**)	✗	✗	42.41	72.86	42.04	43.87	73.66	42.95
	✓	✗	42.96	74.15	42.39	44.58	75.21	44.01
	✗	✓	42.75	73.84	42.26	44.13	74.28	43.72
	✓	✓	**43.52**	**74.84**	**42.65**	**44.95**	**75.67**	**44.28**
Mask R-CNN (**?**)	✗	✗	27.63	57.82	21.55	29.16	61.59	24.83
	✓	✗	28.27	58.64	22.16	30.08	62.35	24.94
	✗	✓	27.80	58.27	21.85	29.73	62.08	24.86
	✓	✓	**28.74**	**59.13**	**22.86**	**30.46**	**62.72**	**25.09**

Effectiveness of cross-attention module. We examine the effectiveness of our cross-attention module in the CamoFourier framework by comparing the performance of COD (SINetV2 (**?**), ZoomNet (**?**)) and CIS (OSFormer (**?**), Mask R-CNN (**?**))

---

*These authors contributed equally.

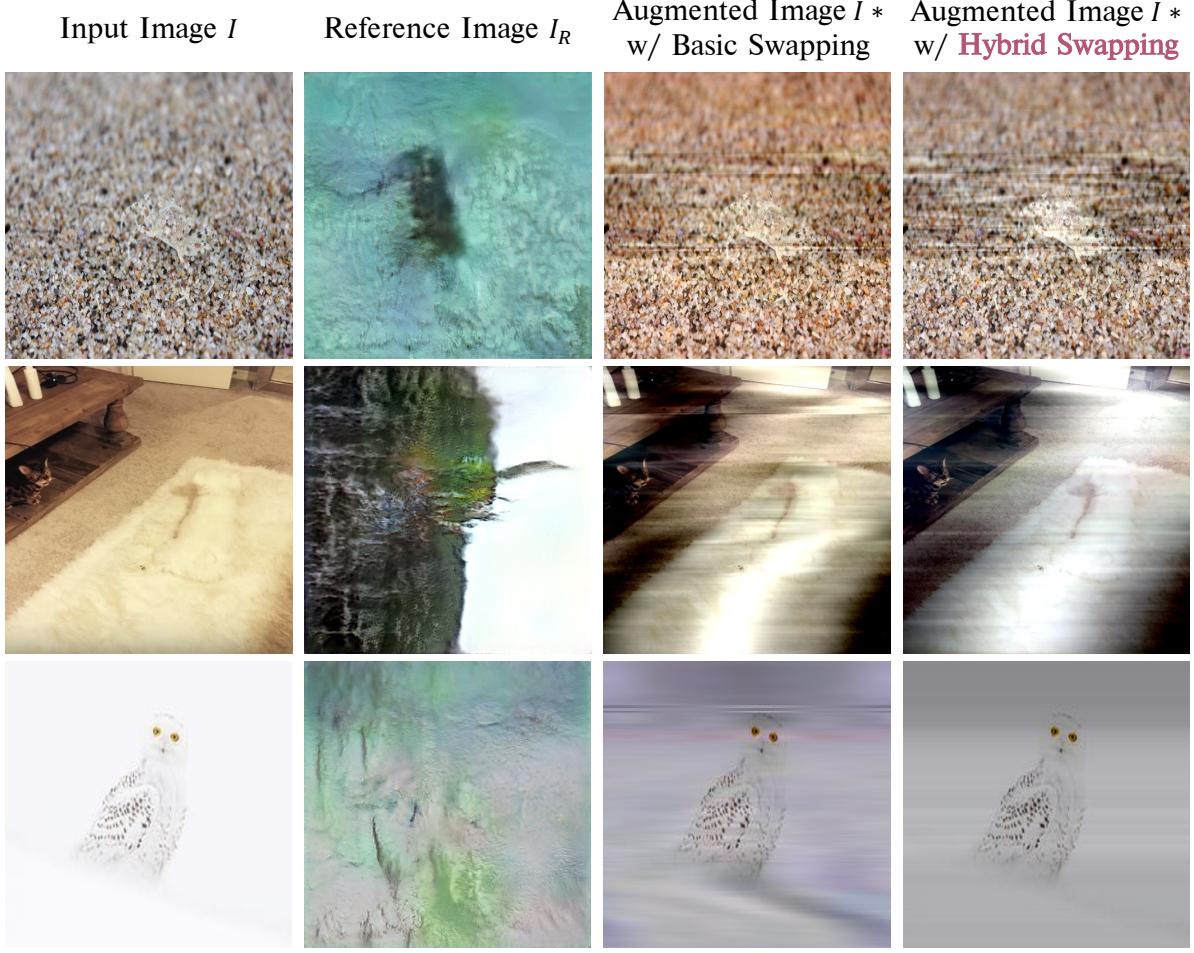| Input Image I | Reference Image I_R | Augmented Image $I*$ w/ Basic Swapping | Augmented Image $I*$ w/ Hybrid Swapping |

Figure 1: Visualization of augmented images by our CamoFourier. Our proposed augmentation highlights the underlying structure of camouflaged objects for better identification and segmentation. We also compare the transformed results of our method with and without adaptive hybrid swapping. Our adaptive hybrid swapping is able to control the amount of texture and color information that is transferred from the reference image to the input image.

models with and without the cross-attention mechanism. Without the cross-attention mechanism, we directly transform the input image $\mathcal{I}$ and the generated image $\mathcal{I}_G$ from the GAN module into the frequency domain and perform amplitude swapping. Table 1 shows that the cross-attention module in CamoFourier improves the performance of COD on the COD10K (**?**) and NC4K (**?**) test sets on four metrics. Table 2 shows that our cross-attention module in CamoFourier also enhances the performance of CIS on the COD10K and NC4K test sets.

Significance of adaptive hybrid swapping. We evaluate the significance of the proposed adaptive hybrid swapping in the CamoFourier augmentation. We also perform experiments on COD and CIS tasks with 4 models: SINetV2 (**?**), ZoomNet (**?**) for COD, OSFormer (**?**), and Mask R-CNN (**?**) for CIS. We apply our CamoFourier to these architectures with and without adaptive hybrid swapping. Without using the proposed hybrid swapping, we use the basic swapping of amplitude. Table 1 and Table 2 show that our adaptive hybrid swapping helps the CamoFourier augmentation to boost the performance of COD and CIS models.

Qualitative Results

Visualization of augmented images. In Fig. 1, we illustrate the outputs of our CamoFourier, which are augmented images $\mathcal{I}*$. In the CamoFourier framework, the conditional GAN module and a cross-attention mechanism synthesize a reference image $\mathcal{I}_R$. After that, we transform an input image and a reference image into the frequency domain and perform basic swapping or adaptive hybrid swapping of these two amplitudes. We also compare the qualitative results of our CamoFourier with and without adaptive hybrid swapping. Figure 1 shows that our adaptive hybrid swapping is able to control the amount of texture and color information that is transferred from the reference image $\mathcal{I}_R$ to the input image $\mathcal{I}$ thanks to the parameter β.

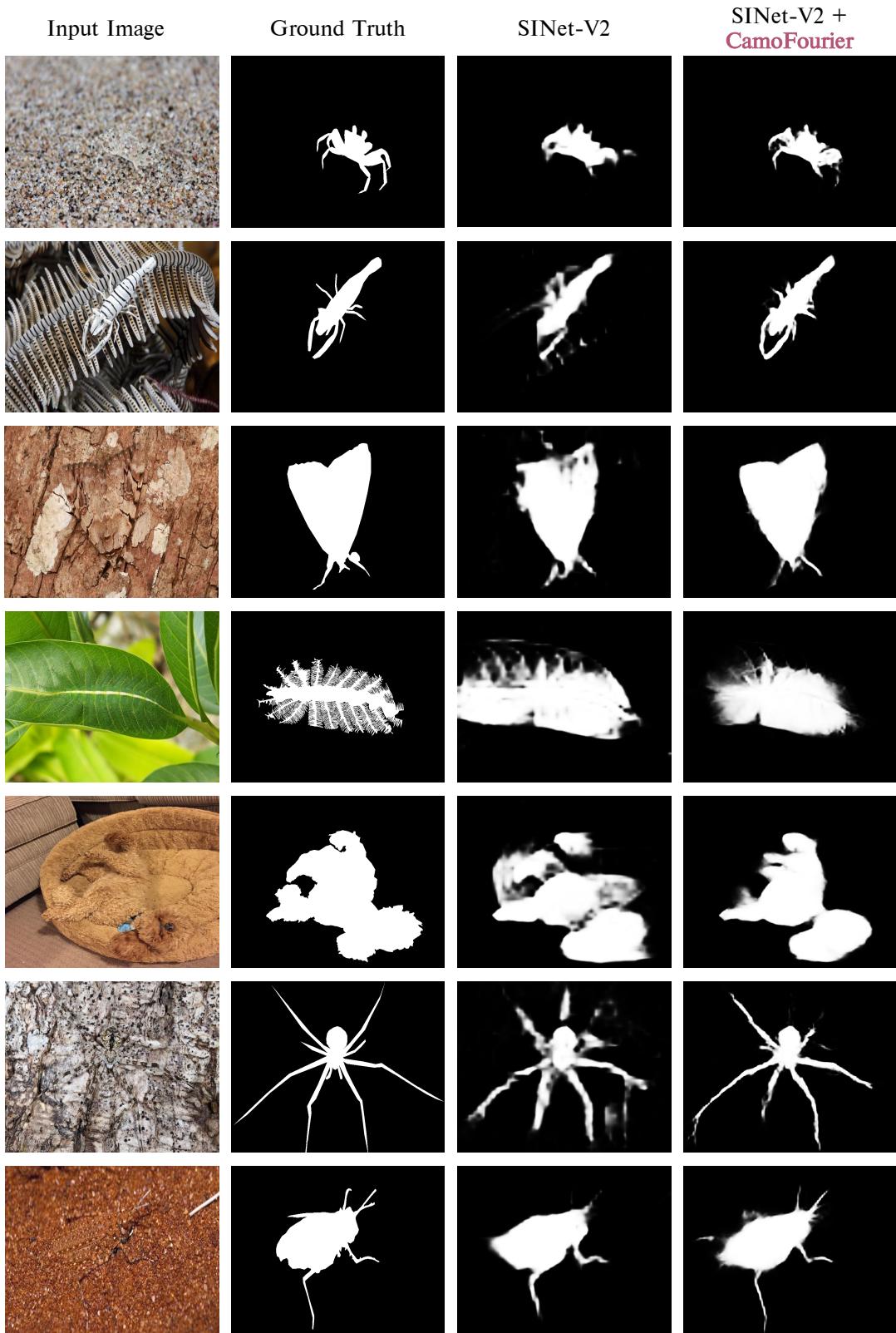| Input Image	Ground Truth	SINet-V2	SINet-V2 + CamoFourier

Figure 2: Qualitative comparison of SINetV2 in the COD task with and without our proposed CamoFourier (best view in color and zoom-in). The segmentation results without CamoFourier are incorrect.

Figure 3: Qualitative comparison of OSFormer in the CIS task with and without our proposed CamoFourier (best view in color and zoom-in). The segmentation results without CamoFourier are incorrect.

Qualitative comparisons.   Figure 2 and Figure 3 visualize the segmentation results of SINetV2 (**?**) and OSFormer (**?**) with and without our CamoFourier in COD and CIS tasks, respectively. The results demonstrate that SINetV2 and OSFormer equipped with CamoFourier are able to delineate camouflaged objects in a more accurate way. Moreover, our CamoFourier can alleviate noisy segmentation as seen in SINetV2's and OSFormer's examples.