CamoDiffusion: Camouflaged Object Detection via Conditional Diffusion Models

Zhongxi Chen, Ke Sun, Xianming Lin*

Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China
{chenzhongxi, skjack}@stu.xmu.edu.cn, linxm@xmu.edu.cn

Abstract

Camouflaged Object Detection (COD) is a challenging task in computer vision due to the high similarity between camouflaged objects and their surroundings. Existing COD methods struggle with nuanced object boundaries and overconfident incorrect predictions. In response, we propose a new paradigm that treats COD as a conditional mask-generation task leveraging diffusion models. Our method, dubbed CamoDiffusion, employs the denoising process to progressively refine predictions while incorporating image conditions. Due to the stochastic sampling process of diffusion, our model is capable of sampling multiple possible predictions, avoiding the problem of overconfident point estimation. Moreover, we develop specialized network architecture, training, and sampling strategies, to enhance the model's expressive power, refinement capabilities and suppress overconfident mis-segmentations, thus aptly tailoring the diffusion model to the demands of COD. Extensive experiments on three COD datasets attest to the superior performance of our model compared to existing state-of-the-art methods, particularly on the most challenging COD10K dataset, where our approach achieves **0.019** in terms of MAE. Codes and models are available at https://github.com/Rapisurazurite/CamoDiffusion.

Introduction

Camouflage, a pervasive defense strategy in nature, endows organisms with the capacity to meld seamlessly into their surroundings, thereby allowing them to elude predators or approach prey surreptitiously (Fan et al. 2021a). As a result, Camouflaged Object Detection (COD) has materialized as a rapidly expanding research field, focusing on the identification of hidden objects or organisms in their natural habitats. This area boasts applications across a multitude of sectors, including species conservation (Nafus et al. 2015), medical image segmentation (Dong et al. 2021), and industrial flaw detection (Bhajantri and Nagabhushan 2006).

To address the camouflaged properties of the foreground object, numerous strategies have been proposed for this task, predominantly encompassing three perspectives (Fan et al. 2023): 1) Multi-stream frameworks (Pang et al. 2022), which exploit multiple input streams to explicitly learn multi-source representations. 2) Bottom-up and top-down frameworks (Fan et al. 2021a) that harness deeper features to progressively enhance shallower ones in a single feed-forward pass. 3) Branched frameworks (Ji et al. 2023), constituting a single-input multiple-output architecture, comprising both segmentation and auxiliary task branches. Such techniques mainly build upon the foundational semantic segmentation paradigm, which capitalizes on a learning-based backbone for feature extraction, subsequently employing a decoder head to yield a segmentation mask. However, in the absence of intricately designed strategies, such a paradigm is prone to confuse the subtle deviation between boundaries and surroundings of camouflaged objects due to body outline disguising (Sun et al. 2022). Furthermore, identifying the main subject in COD is challenging, which can lead to overconfident but incorrect predictions based on pixel probabilities, which amplifies the potential for misjudgments. Examples of these situations are presented in Fig. 1.

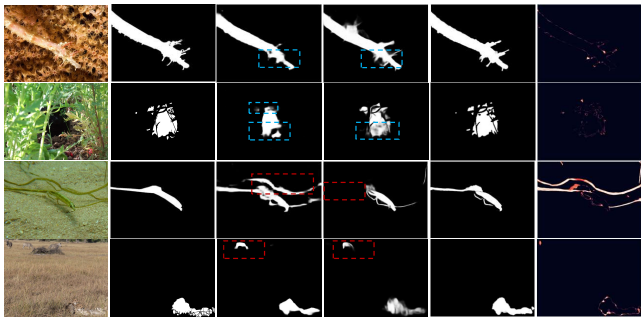

Figure 1: Current COD methods blur the boundary-context distinction of camouflaged objects (rows 1, 2) and occasionally yield overconfident mis-segmentations (rows 3, 4). We introduce diffusion models into COD, which enhance the edge discrimination, suppress overconfident mis-segmentations (Ours) and enable uncertainty assessment through variance calculation (Variance).

Given such specific challenges posed by COD, we adopt the diffusion model paradigm (Ho, Jain, and Abbeel 2020) as a fitting solution. Diffusion models possess impressive generative capabilities with a good awareness of conditions. Their iterative noise reduction mechanism obviates

*Corresponding Author.

the need for intricate refinement modules within prevailing COD models. This mechanism enables a progressive differentiation between boundary nuances and contextual surroundings, fostering a more holistic comprehension. And the inherent stochastic sampling process enables generating multiple predictions and evaluating segmentation uncertainty, which mitigates the risk of erroneous confidence in the model's results. However, the direct application of diffusion models to COD revealed inadequacies due to the distinct nature of the task. To overcome the inherent limitations of conventional diffusion models for COD, characterized by limited discriminative ability and inadequate mask refinement, we have tailored our approach. Our novel framework, named CamoDiffusion treats COD as a conditional mask-generation task based on diffusion models. In essence, CamoDiffusion leverages the denoising process of diffusion models to progressively rectify the discrepancy between initial noise and ground truth and incorporate the image as an auxiliary condition. Specifically, we propose an Adaptive Transformer Conditional Network to enhance the expressive power by introducing a guidance cue, which plays a pivotal role in distinguishing camouflaged objects featuring intricate boundaries. Furthermore, we design specialized learning strategies during the training and sampling processes. These encompass an SNR-based variance schedule, Structure Corruption, and Consensus Time Ensemble, designed to foster comprehensive image feature exploration, bolster corrective capabilities, and optimally harness the inherent attributes of the diffusion model to suppress overconfident incorrect segmentation (shown in Fig. 1). Compared to previous COD methods, our method has the following advantages: 1) CamoDiffusion can significantly enhance the capacity to handle subtle details of the foreground. 2) CamoDiffusion allows for a more comprehensive understanding of the intrinsic differences between camouflaged objects and their surroundings, leading to better generalization and less mis-segmentation. 3) CamoDiffusion can sample multiple possible predictions, avoiding the problem of overconfident point estimation.

Our main contributions can be summarized as follows:

1. We are the first to treat the COD task as a mask generation paradigm and use a conditional diffusion framework to form predictions.
2. We propose a novel and effective framework called CamoDiffusion, which uses a specially designed network structure and learning strategies to generate more accurate and generalized results for the COD task.
3. Our CamoDiffusion achieves state-of-the-art (SOTA) performance on three COD datasets, demonstrating its superior effectiveness in completing the COD task.

Related Work

Camouflaged Object Detection

In recent years, CNN-based methods have achieved remarkable advancements in the COD task by employing intricate strategies. For instance, SINet-V2 (Fan et al. 2022) and BASNet (Qin et al. 2019) employ multi-stage approaches for preliminary rough segmentations, succeeded by refinement techniques. Multi-task learning-based approaches, such as LSR (Lv et al. 2021), and DGNet (Ji et al. 2023), merge diverse detection tasks like ranking, localization, and gradient generation to assist segmentation and enhance COD performance. ZoomNet (Pang et al. 2022) and MFFN (Zheng et al. 2023) employ multiple augmented views as inputs to extract features and fuse them to obtain an improved feature map. Conversely, Vision Transformer has recently demonstrated exceptional performance in COD by using self-attention mechanisms to model long-range dependencies. Notable works in this area include DTINet (Liu et al. 2022), FSPNet (Huang et al. 2023), and HitNet (Hu et al. 2023), which adopt a dual-task interactive Transformer, non-local mechanisms, and iterative refinement of low-resolution representations through feedback, respectively. Despite these advancements, many COD models are still limited to point estimations and struggle with overconfident mis-segmentation. Furthermore, the intricate architecture of these models demands meticulous tuning.

Diffusion Models for Image Segmentation

Diffusion models (Ho, Jain, and Abbeel 2020; Song et al. 2020; Song and Ermon 2019) are parameterized Markov chains that gradually denoise data samples starting from random noise. Although originally applied in fields without definitive ground truth (Dhariwal and Nichol 2021), recent research has demonstrated their effectiveness for problems with unique ground truths, such as super-resolution (Li et al. 2022; Wang et al. 2021), deblurring (Whang et al. 2022; Lee et al. 2022), and image segmentation (Baranchuk et al. 2021; Amit et al. 2021; Chen et al. 2022). This work highlights the potential of diffusion models in various segmentation tasks, such as remote sensing change detection (Bandara, Nair, and Patel 2022) and medical image segmentation (Wolleb et al. 2022; Wu et al. 2022; Rahman et al. 2023). However, these diffusion-based methods still lag behind universal image segmentation approaches, particularly in the context of COD, which presents a unique predicament where conventional diffusion methods tend to yield image features of limited discriminative prowess and fail to refine masks.

Method

In this section, we introduce our CamoDiffusion framework, which progressively generates predictions by conditioning each subsequent step with the image prior. This approach is enhanced through three key aspects: the network architecture, training and sampling strategy. As depicted in Fig. 2, our model comprises an Adaptive Transformer Conditional Network (ATCN) and a Denoising Network (DN). We first introduce the basic background and notation. Then we discuss the architectural details of the ATCN and DN. The details of the specially designed learning strategies are then elaborated upon.

Background and Notation

Our CamoDiffusion is based on diffusion models, which include a forward process, where a mask is gradually noised,

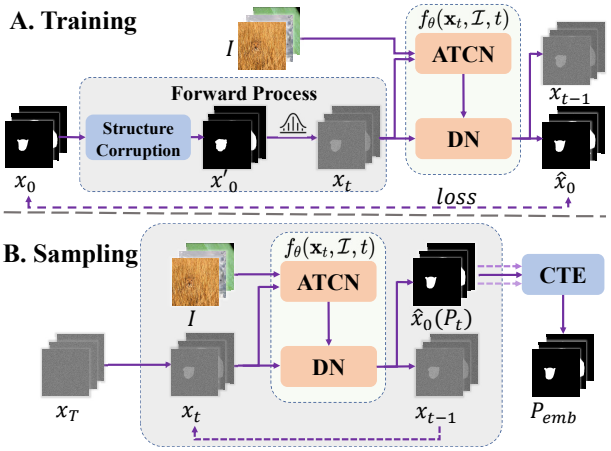

Figure 2: The overall framework of our model $f_\theta(\mathbf{x}_t, \mathcal{I}, t)$, which includes an Adaptive Transformer Conditional Network (ATCN), and a Denoising Network (DN). In Training, the ground truth (gt) $\mathbf{x}_0$ undergoes Structure Corruption and Gaussian noise to generate the noised mask $\mathbf{x}_t$ at time t. Subsequently, our model predicts the denoised mask $\hat{\mathbf{x}}_0$ and is trained by minimizing the loss between the prediction and the gt. During Sampling, our model denoises the random noise $\mathbf{x}_T$ for T steps. All predictions $\{P_t\}_{t=1}^T$ generated in the reverse process are aggregated by the Consensus Time Ensemble (CTE), resulting in a more reliable outcome P_{emb}.

and a reverse process, where noise is transformed back to the target distribution. Given a training sample $\mathbf{x}_0 \sim q(\mathbf{x}_0)$, the noised version $\{\mathbf{x}_t\}_{t=1}^T$ are obtained according to the following Markov process:

$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

where t runs from 1 to T, and variance is controlled by noise schedule $\beta_t \in (0, 1)$. The marginal distribution of $\mathbf{x}_t$ can be described as:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}), \tag{2}$$

where $\bar{\alpha}_t = \prod_{i=1}^T \alpha_t, \alpha_t = 1 - \beta_t$. Starting from $p(\mathbf{x}_T) = \mathcal{N}(\mathbf{x}_T; \mathbf{0}, \mathbf{I})$, the reverse process uses a neural network f_θ to create a sequence of incremental denoising operations to obtain back the clean mask. The network learns the reverse distribution:

$$p(\mathbf{x}_{t-1}|\mathbf{x}_t) := \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \Sigma_\theta(\mathbf{x}_t, t)). \tag{3}$$

While it is feasible to model $\mathbf{x}_{t-1}$ directly, Ho, Jain, and Abbeel (2020) suggested that a consistent output space for the network leads to enhanced performance. In our proposed CamoDiffusion, we choose to train a network $f_\theta(\mathbf{x}_t, \mathcal{I}, t)$ to predict the denoised mask $\hat{\mathbf{x}}_0$ conditional on image $\mathcal{I}$. In practice, $\Sigma_\theta(\mathbf{x}_t, t)$ is set to $\sigma_t^2 = \frac{1-\bar{\alpha}_{t-1}}{1-\bar{\alpha}_t}\beta_t$, and $\mu_\theta(\mathbf{x}_t, t)$ can be expressed as:

$$\mu_\theta(\mathbf{x}_t, t) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\hat{\mathbf{x}}_0, \tag{4}$$

where $\hat{\mathbf{x}}_0$ is predicted by our model $f_\theta(\mathbf{x}_t, I, t)$, which consists of ATCN and DN. Then we optimize $\mathcal{L}$ to train our model, Formally,

$$\mathcal{L} = \mathcal{L}_{IoU}^w(\hat{\mathbf{x}}_0, \mathbf{x}_0) + \mathcal{L}_{BCE}^w(\hat{\mathbf{x}}_0, \mathbf{x}_0), \tag{5}$$

where $\mathcal{L}_{IoU}^w$ and $\mathcal{L}_{BCE}^w$ represent weighted intersection-over-union (IoU) loss and weighted binary cross entropy (BCE) loss.

Architecture Design

In contrast to the conventional segmentation paradigm, our proposed model employs conditional diffusion models to generate predictions. Specifically, as illustrated in Fig. 3, we utilize the ATCN to extract hierarchical image features as conditions, which are then integrated with the downstream DN. We discuss the design details of these networks in the following sections.

Adaptive Transformer Conditional Network (ATCN) Within our framework, the role of the ATCN is to enable the downstream denoising network at each step to adequately discern the camouflaged image, thereby distinguishing the camouflaged target. We identify two primary challenges in the design of a conditional network for COD: 1) Extracting more discriminative image features. 2) Adaptively supplying conditional features in accordance with the denoising step.

The inherent concealment of camouflaged objects poses a substantial challenge in extracting discriminative image features, leading to compromised mask decoder performance. To tackle this, we incorporate a coarsely predicted mask from the previous step as a guiding cue, enabling the network to selectively concentrate on specific regions. This approach facilitates the revelation of intricate details and contours of camouflaged objects. Furthermore, to enhance the adaptability of the extracted features across denoising steps, we introduce t into the conditional network. These innovations distinguish our model from conventional diffusion models, where prior research (Amit et al. 2021; Chen et al. 2022) employed a conditional network that solely relied on image inputs, utilizing the same image features in reverse processes. Through the proposed ATCN, we achieve dynamic and comprehensive extraction of discerning image features. Specifically, as depicted in Fig. 3A, the ATCN comprises Pyramid Vision Transformer (PVT) (Wang et al. 2022b) layers that extract multi-scale features $\{\mathbf{F}_i\}_{i=1}^4$ from image $\mathcal{I}$ considering both the previous segmentation $\mathbf{x}_t$ and the current denoising step t. To achieve this, we design **Zero Overlapping Embedding** and **Time Token Concatenation** to inject them adaptively.

- **Zero Overlapping Embedding (ZOE):** To incorporate the noise mask $\mathbf{x}_t$ into the PVT without destroying the original transformer structure and pre-training parameters, we propose the ZOE in the first layer of ATCN instead of the original position encoding module Overlapping Embedding (OE). Specifically, ZOE uses an extra convolution layer initialized with zeros, which gradually introduces $\mathbf{x}_t$ in a controlled manner without affecting the position encoding during initialization. The mathematical representation of the embedding of the i-th layer is given by:

$$\mathbf{emb_i} = \begin{cases} \mathrm{LN}(\mathbb{R}(\mathrm{Conv}(\mathcal{I}) + \mathrm{Conv}_z(\mathbf{x}_t))), & i = 1, \\ \mathrm{LN}(\mathbb{R}(\mathrm{Conv}(\mathbf{F}_{i-1}))), & i \neq 1. \end{cases} \tag{6}$$

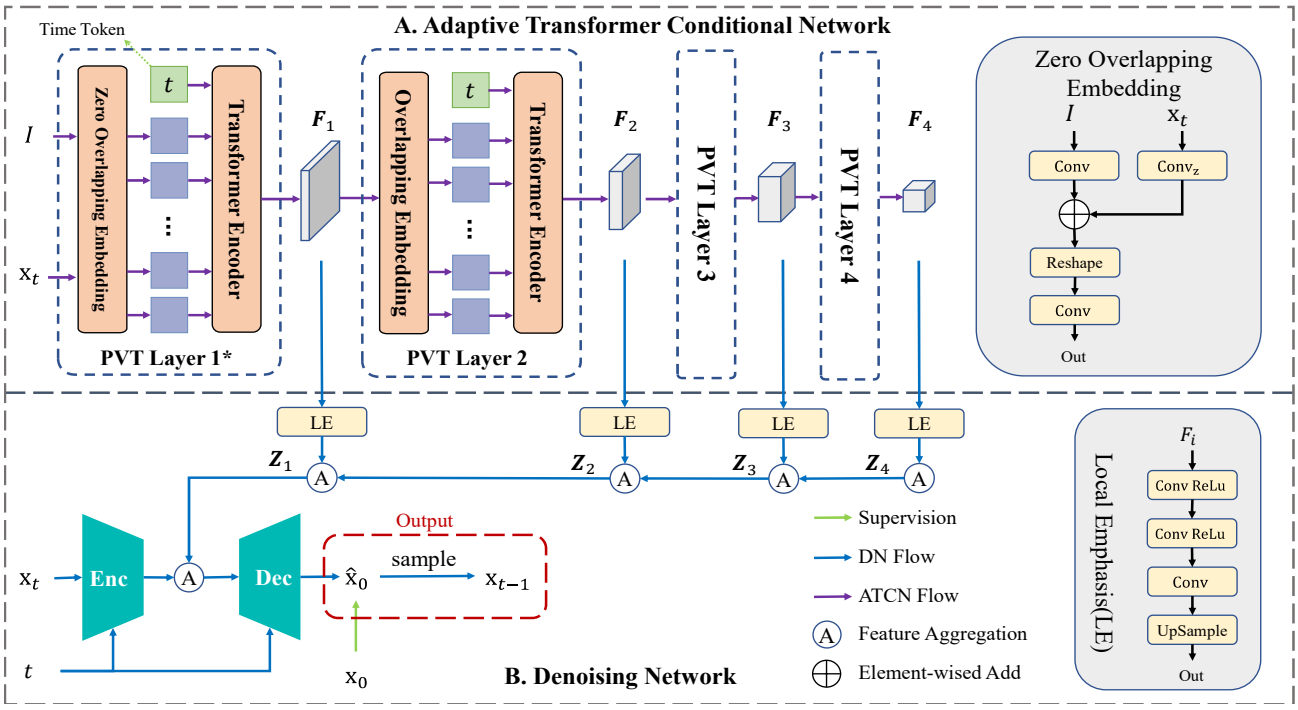

Figure 3: The architecture design of our CamoDiffusion. It consists of an Adaptive Transformer Conditional Network (ATCN) for extracting multi-scale features as conditions, and a Denoising Network (DN) for recovering clear mask predictions from the noised mask. "*" represents that the PVT layer is modified by our proposed Zero Overlapping Embedding (ZOE).

Here, $\mathrm{Conv}(\cdot)$ refers to a convolution layer, $\mathrm{Conv}_z(\cdot)$ denotes a convolution layer with weight and bias initialized as zeros, $\mathtt{R}(\cdot)$ converts the feature map obtained from the convolution to tokens, and $\mathrm{LN}(\cdot)$ denotes layer normalization.

• **Time Token Concatenation (TTC):** Beyond the noisy mask, we anticipate our ATCN to autonomously adjust the characteristics of the condition according to the temporal stage. To address it, we propose TTC:

$$\mathbf{F}_i = \mathtt{R}^{-1}(\mathrm{FFN}(\mathrm{MHA}([\mathbf{t}; \mathbf{emb_i}]))), \tag{7}$$

where $\mathbf{t}$ represent the time token, $[\cdots]$ refers to the concatenation operation, and $\mathtt{R}^{-1}$ convert tokens to feature map. MHA and FFN represent the multi-head self-attention and the feed-forward neural network, respectively.

Denoising Network (DN) The Denoising Network aims to decode the denoised mask prediction $\hat{\mathbf{x}}_0$ and $\mathbf{x}_{t-1}$, based on the diffusion paradigm. Instead of designing complex hierarchically refined decoders, we achieved satisfactory results by utilizing the simple structure due to the iterative denoising process of diffusion. Specifically, $\{\mathbf{F}_i\}_{i=1}^4$ are upsample to the same size using Local Emphasis (LE) module (Wang et al. 2022a):

$$\mathrm{LE}(\mathbf{F}_i) = \mathrm{Up}(\mathrm{CR}(\mathrm{CR}(\mathbf{F}_i))), \tag{8}$$

where $\mathrm{Up}(\cdot)$ is the bilinear interpolation and $\mathrm{CR}(\cdot)$ represents the combination of convolution and ReLU. The features are then aggregation gradually:

$$\mathbf{Z}_i = \mathrm{Conv}([\mathbf{Z_{i+1}}, \mathrm{LE}(\mathbf{F}_i)]), \quad i \in \{3, 2, 1\}, \tag{9}$$

where $\mathbf{Z_4} = \mathrm{LE}(\mathbf{F}_4)$. Finally, a lightweight encoder and decoder is used to denoise $\mathbf{x}_t$ under the guidance of condition $\mathbf{Z}_1$ and t.

Training Strategy

During Training, we initiate a diffusion process from the GT to the noisy mask and train our model to reverse this process. Despite successful training, certain challenges arise. Due to the intricacies of COD, the model struggles to recover a clear mask from a low Signal-to-Noise Ratio (SNR) mask according to image features. To address this, an **SNR-based variance schedule** is adopted to improve the model's efficacy. The second challenge stems from the model's lack of corrective capabilities with regard to the subtle contours of camouflaged objects, and we address it by utilizing **Structure Corruption** to facilitate the model's ability to learn structure-level denoising.

• **SNR-based Variance Schedule:** It was observed that the masks have an excessive SNR in training, which poses a challenge for the model to recover the mask from low SNR inputs (Chen et al. 2022). This predicament arises due to the model's proclivity to gravitate towards the path of least resistance, relying on the more conspicuous noised mask instead of leveraging image features. In contrast to other image segmentation, COD presents a tougher challenge due to intricate contexts and higher resolutions which leads to more easy cases in training. To address this, we adopt an SNR-based variance schedule (Hoogeboom, Heek, and Salimans 2023) during training. Specifically, we introduce an offset

Baseline Models	Backbone	CAMO(250)				COD10K(2026)				NC4K(4121)			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Salient / Camouflaged Object Detection													
EGNet (Zhao et al. 2019)	ResNet-50	0.662	0.683	0.495	0.125	0.733	0.761	0.519	0.055	0.767	0.793	0.626	0.077
SINet (Fan et al. 2020)	ResNet-50	0.745	0.804	0.644	0.092	0.776	0.864	0.631	0.043	0.808	0.871	0.723	0.058
MGL (Zhai et al. 2021)	ResNet-50	0.775	0.812	0.673	0.088	0.814	0.852	0.666	0.035	0.833	0.867	0.740	0.052
PFNet (Mei et al. 2021)	ResNet-50	0.782	0.841	0.695	0.085	0.800	0.877	0.660	0.040	0.829	0.887	0.745	0.053
UGTR (Yang et al. 2021)	ResNet-50	0.785	0.823	0.686	0.086	0.818	0.853	0.667	0.035	0.839	0.874	0.747	0.052
LSR (Lv et al. 2021)	ResNet-50	0.787	0.838	0.696	0.080	0.804	0.880	0.673	0.037	0.840	0.895	0.766	0.048
PreyNet (Zhang et al. 2022)	ResNet-50	0.790	0.842	0.708	0.077	0.813	0.881	0.697	0.034	0.834	0.887	0.763	0.050
SINet-V2 (Fan et al. 2021a)	Res2Net-50	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.770	0.048
ZoomNet (Pang et al. 2022)	ResNet-50	0.820	0.877	0.752	0.066	0.838	0.888	0.729	0.029	0.853	0.896	0.784	0.043
DTINet (Liu et al. 2022)	MiT-B5	0.856	0.916	0.796	0.050	0.824	0.896	0.695	0.034	0.863	0.917	0.792	0.041
FEDER-R50 (He et al. 2023)	ResNet-50	0.802	0.867	0.738	0.071	0.822	0.900	0.716	0.032	0.847	0.907	0.789	0.044
DGNet (Ji et al. 2023)	EffNet-B4	0.839	0.901	0.769	0.057	0.822	0.896	0.693	0.033	0.857	0.911	0.784	0.042
HitNet (Hu et al. 2023)	PVTv2-B2	0.849	0.906	0.809	0.055	0.871	0.935	0.806	0.023	0.875	0.926	0.834	0.037
FSPNet (Huang et al. 2023)	ViT-B	0.856	0.899	0.799	0.050	0.851	0.895	0.735	0.026	0.879	0.915	0.816	0.035
Diffusion Models for Image Segmentation													
EnsemDiff (Wolleb et al. 2022)	N/A	0.716	0.751	0.568	0.113	0.712	0.748	0.508	0.069	0.758	0.791	0.601	0.089
MedSegDiff (Wu et al. 2022)	N/A	0.735	0.801	0.641	0.098	0.731	0.812	0.548	0.047	0.785	0.824	0.664	0.069
CamoDiffusion(Ours)	PVTv2-B4	**0.878**	**0.940**	**0.853**	**0.042**	0.881	**0.944**	0.814	0.020	0.893	**0.942**	0.859	0.029
CamoDiffusion-E(Ours)	PVTv2-B4	**0.878**	0.936	0.851	**0.042**	**0.883**	0.943	**0.817**	**0.019**	**0.895**	**0.942**	**0.861**	**0.028**

Table 1: Quantitative results of our method and other state-of-the-art methods on three benchmark datasets.

to the variance schedule on the log scale: $\mathrm{SNR}_{\text{shift}}(t) = \exp(\log \mathrm{SNR}(t) + \text{shift})$, where $\mathrm{SNR}(t) = \bar{\alpha}_t/(1-\bar{\alpha}_t)$. The essence of this approach lies in intentionally diminishing the SNR of input masks, a methodical decision aimed at heightening the complexity of the training process, encouraging better feature exploration and adapting the model to the unique complexities of COD.

• **Structure Corruption:** Existing diffusion models employ pixel-level corruption to generate the noised mask directly from the GT, leading the model to incorrectly assume that the restored contour from the noised mask is accurate, leading to limited corrective capabilities. However, the difficulty in distinguishing camouflaged object boundaries often results in substantial errors in the model's initial segmentation predictions. To address this issue, we propose Structure Corruption during forward diffusion, where we randomly destroy the contour of the GT and then add Gaussian noise. This improves the model's ability to correct biases in the previous predictions, which is particularly crucial given the indistinct boundary contours characteristic of camouflaged objects.

Sampling Strategy

Our denoising model applies incremental denoising to a sample $\mathbf{x}_T$ drawn from a standard normal distribution over T steps. This iterative denoising process steadily mitigates the divergence between the predicted mask and the ground truth, culminating in a more precise outcome. In COD, the difficulty of identifying the main subject often leads to overconfident incorrect segmentations in existing models. In response, we introduce a **Consensus Time Ensemble** (CTE) approach premised on the concept that predictions generated during the denoising process carry valuable insights. This technique amalgamates predictions from each denoising step, augmenting both the precision and reliability of the resultant output.

• **Consensus Time Ensemble (CTE):** Inspired by the annotation procedure of saliency detection (Zhang et al. 2021), we propose CTE strategy to combine predictions from different sampling steps without incurring additional computational costs. Specifically, for each Sampling stage at time t, the denoised image $\hat{\mathbf{x}}_0$ is denoted as P_t. Given multiple predictions $\{P_t\}_{t=1}^T$, the binary masks $\{P_t^b\}_{t=1}^T$ are first calculated through adaptive thresholding. These predictions $\{P_t^b\}_{t=1}^T$ vote on the position of each point to generate a candidate mask, and the probability value of the selected point is the mean of all predictions. Mathematically,

$$P_{emb} = \left\lfloor \frac{\sum_{t=1}^{T} P_t^b}{T} + \frac{1}{2} \right\rfloor * \mathrm{mean}(P_t). \tag{10}$$

In addition, our model can generate multiple predictions by sampling from the mask distribution. This allows us to improve mask accuracy through ensemble techniques or assess uncertainty by calculating variance, as shown in Fig. 1. In practice, we sample the mask thrice and apply CTE to combine $3T$ predictions. The performance of this approach is presented as 'CamoDiffusion-E' in our evaluation, and the visual analysis of CTE strategy for suppressing overconfident mis-segmentation is discussed in *supp*.

Experiments

Experiment Settings

Datasets. Our CamoDiffusion is evaluated on three widely-used COD datasets: CAMO (Le et al. 2019), COD10K (Fan et al. 2021a), and NC4K (Lv et al. 2021). The CAMO dataset comprises 1,250 camouflaged images and 1,250 non-camouflaged images. COD10K con-

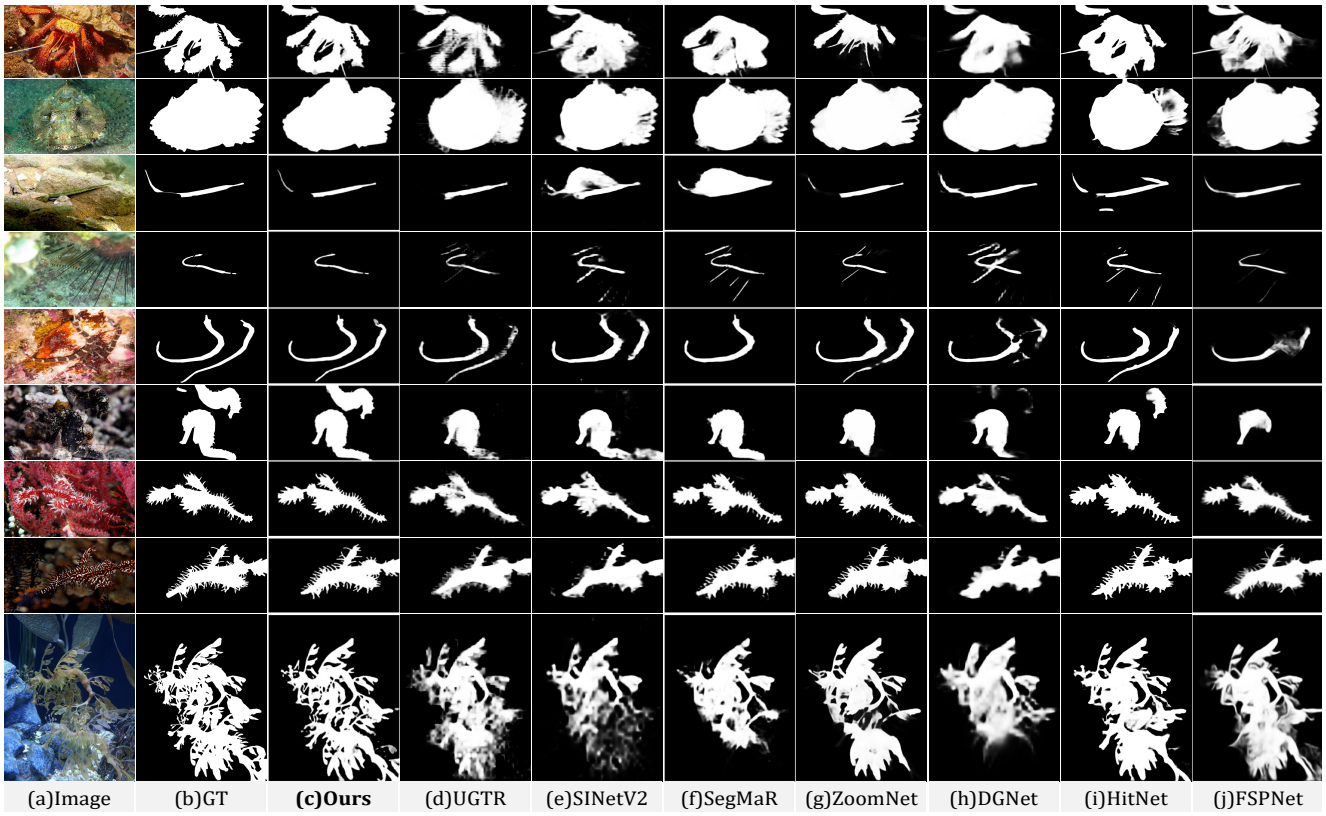

Figure 4: Visual comparisons with recent SOTA models in typical and challenging scenarios.

tains 5,066 camouflaged, 3,000 background, and 1,934 non-camouflaged images. NC4K is a large-scale COD dataset consisting of 4,121 images and is often used to evaluate the generalization ability of models.

Evaluation Metrics. In order to evaluate the performance of our proposed method, we adopt four commonly used metrics: mean absolute error (M) (Perazzi et al. 2012), S-measure (S_α) (Fan et al. 2017), weighted F-measure (F_β^ω) (Margolin, Zelnik-Manor, and Tal 2014), and mean E-measure (E_ϕ) (Fan et al. 2021b).

Implementation Details. We implemented our model based on PyTorch using an NVIDIA A100 for both training and inference. ATCN is initialized using PVTv2-B4, and input images are resized to 384×384. For optimization, the AdamW was utilized along with a batch size set to 32. To adjust the learning rate, we implemented the cosine strategy with an initial learning rate of 0.001 for 170 epochs. Acknowledging the randomness inherent to our sampling method, we evaluated it thrice with different seeds and reported the average of the metrics for CamoDiffusion. Additionally, we introduced CamoDiffusion-E, which enhances reliability by CTE strategy through triple inference passes. Notably, the default configuration employs $T = 10$ for sampling.

Quantitative and Qualitative Evaluation Tab. 1 summarizes the quantitative results of our proposed method against 14 competitors on three COD benchmark datasets. Impressively, our model outperforms all rivals on these evaluations, dramatically reducing the MAE error by 20.9% and increasing F_β^w by 7.7% compared to the second-best performer, FSPNet. The enhanced performance results from the iterative denoising process of diffusion models, further bolstered by the CTE strategy. Notably, our method stands out among diffusion models due to its tailored enhancements that align adeptly with the demands of the COD task. Fig. 4 presents the visual comparisons of CamoDiffusion with several recent models. The examples illustrated in the first six rows exhibit typical scenarios, including objects of large or small sizes and multiple objects. Our proposed CamoDiffusion excels and exhibits reduced overconfident mis-segmentation in these instances. Rows 7-9 present examples with intricate topological structures and detailed edges, which pose significant challenges to current COD models. However, our model adeptly delineates clear boundaries in these cases, leveraging its iterative denoising paradigm.

Ablation Studies

Ablation of Model Components. We conduct an ablation study on the individual components of CamoDiffusion, with detailed results provided in Tab. 2. In cases where both ZOE and TTC are marked as '×', it indicates that we did not employ ATCN, thus utilizing the denoising process based on the same image features. Employing ATCN allows us to extract more discriminative and dynamic features, leading to enhanced predictive accuracy. Additionally, our discoveries

	ZOE	TTC	SNR	SC	$S_\alpha \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Model Components	×	×	✓		0.867	0.783	0.022
	×	✓			0.872	0.795	0.022
	✓	×			0.877	0.806	**0.020**
	✓	✓			**0.881**	**0.814**	**0.020**
Training Strategies	✓		×	×	0.821	0.713	0.033
			✓	×	0.868	0.789	0.022
			×	✓	0.855	0.767	0.024
			✓	✓	**0.881**	**0.814**	**0.020**

Table 2: Ablation of model components and training strategies on COD10K dataset.

Methods	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	OIP	TIP	%
w/o CTE	0.877	0.941	0.803	0.021	9419	10394	91%
Ours	0.881	**0.944**	0.814	0.020	7644	8957	85%
Ours-E	**0.883**	0.943	**0.817**	**0.019**	**6649**	**8698**	**76%**

Table 3: Ablation of our CTE strategy on COD10K, where 'OIP' stands for Overconfident Incorrect Pixels and 'TIP' is Total Incorrect Pixels.

indicate that the application of ZOE yields the most significant enhancement in performance and the exclusion of TTC results in performance degradation.

Ablation of Training Strategies. We evaluate the effectiveness of the SNR-based variance schedule and Structure Corruption employed in our CamoDiffusion, marked as 'SNR' and 'SC' in Tab. 2. These training strategies contribute to improved feature exploration and corrective capabilities, through increased training complexity. The outcomes indicate a notable performance improvement when incorporating both the SNR-based schedule and Structure Corruption. This combined approach significantly enhances the efficacy of our model for COD.

Ablation of Sampling Strategy. As detailed in Tab. 3, we explored the impact of our CTE approach. In terms of evaluation metrics, our model's performance exhibits notable improvement with the integration of CTE, and further enhancement is achieved by ensembling additional predictions (CamoDiffusion-E). To evaluate the effect of mitigating overconfident mis-segmentation, we calculate misclassified pixels where the confidence was extreme and dilate to filter out the impact of inadequate model precision. The results presented in Tab. 3 indicate a significant reduction in overconfident incorrect pixels with the CTE strategy. The proportion of such pixels in total incorrect ones decreased from 91% to 76%, highlighting improved model reliability. Additional implementation details are available in the *supp*.

Analysis

As evidenced in Tab. 4, we undertook an evaluation to discern the impact of the diffusion paradigm. By substituting the backbone with the lightweight PVTv2-B2, we conducted a comparative analysis against models that share the same backbone. Specifically, the 'baseline' model adheres

Model	CAMO M	COD10K M	NC4K M	Avg. Imp.
Baseline	0.057	0.031	0.042	-
DGNet*	0.051	0.029	0.037	10.64%
HitNet	0.055	0.023	0.037	14.83%
Ours*	**0.048**	**0.022**	**0.032**	23.83%

Table 4: Evaluating models with same backbone (PVTv2-B2) across three datasets using the MAE metric. '*' means which utilizing backbones distinct from those in Tab. 1

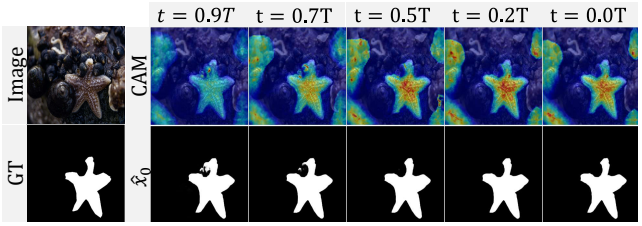

Figure 5: The mask predictions and feature maps at different sample steps.

to the segmentation paradigm, equipped with a PVTv2-B2 backbone along with several convolutional layers, and the 'DGNet*' implementation is available through their official open-source code. Recent models exhibit better performance due to tailored designs addressing COD's unique challenges. Notably, our method consistently outperforms its counterparts, enhancing the baseline by an impressive 23.83%. This result underscores the substantial performance improvement achieved through the integration of the diffusion framework.

To illustrate our model's ability to diminish noise and progressively focus on intricate details, Fig. 5 showcases prediction results captured at different sampling stages. Initially, the model produces a coarse mask with a high degree of uncertainty in specific regions where borders are unclear. However, as the number of sampling steps increases, the model progressively concentrates on the concealed object and refines the mask, establishing deterministic boundaries based on the subtle details of the foreground. The iterative nature of this refinement is empowered by the integration of diffusion models, and obviates the need for intricate refinement modules within prevailing COD models.

Conclusion

In this study, we present a diffusion-based COD model, named CamoDiffusion. By utilizing the iterative nature of the diffusion model, we effectively address two pivotal challenges in COD: the need for intricate strategies to delineate distinct boundaries and overconfident incorrect predictions. In line with the unique attributes of COD, we improved the network architecture, training strategy, and sampling strategy in three steps. This augmentation enhances the features' discriminative abilities and refinement capabilities, and mitigates overconfident mis-segmentation tendencies, enabling the entire framework to achieve SOTA performance.

Acknowledgments

This work was supported by National Key R&D Program of China (No.2022ZD0118202), the National Science Fund for Distinguished Young Scholars (No.62025603), the National Natural Science Foundation of China (No. U21B2037, No. U22B2051, No. 62176222, No. 62176223, No. 62176226, No. 62072386, No. 62072387, No. 62072389, No. 62002305 and No. 62272401), and the Natural Science Foundation of Fujian Province of China (No.2021J01002, No.2022J06001).

References

Amit, T.; Nachmani, E.; Shaharbany, T.; and Wolf, L. 2021. Segdiff: Image segmentation with diffusion probabilistic models. *arXiv preprint arXiv:2112.00390*.

Bandara, W. G. C.; Nair, N. G.; and Patel, V. M. 2022. DDPM-CD: Remote Sensing Change Detection using Denoising Diffusion Probabilistic Models. *arXiv preprint arXiv:2206.11892*.

Baranchuk, D.; Voynov, A.; Rubachev, I.; Khrulkov, V.; and Babenko, A. 2021. Label-Efficient Semantic Segmentation with Diffusion Models. In *International Conference on Learning Representations*.

Bhajantri, N. U.; and Nagabhushan, P. 2006. Camouflage defect identification: a novel approach. In *9th International Conference on Information Technology (ICIT'06)*, 145–148. IEEE.

Chen, T.; Li, L.; Saxena, S.; Hinton, G.; and Fleet, D. J. 2022. A generalist framework for panoptic segmentation of images and videos. *arXiv preprint arXiv:2210.06366*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-PVT: Polyp Segmentation with PyramidVision Transformers. *arXiv preprint arXiv:2108.06932*.

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.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2021a. Concealed Object Detection. *IEEE TPAMI*.

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

Fan, D.-P.; Ji, G.-P.; Qin, X.; and Cheng, M.-M. 2021b. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 6.

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

Fan, D.-P.; Ji, G.-P.; Xu, P.; Cheng, M.-M.; Sakaridis, C.; and Van Gool, L. 2023. Advances in Deep Concealed Scene Understanding. *arXiv preprint arXiv:2304.11234*.

He, C.; Li, K.; Zhang, Y.; Tang, L.; Zhang, Y.; Guo, Z.; and Li, X. 2023. Camouflaged Object Detection with Feature Decomposition and Edge Reconstruction. In *CVPR*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33: 6840–6851.

Hoogeboom, E.; Heek, J.; and Salimans, T. 2023. simple diffusion: End-to-end diffusion for high resolution images. *arXiv preprint arXiv:2301.11093*.

Hu, X.; Fan, D.-P.; Qin, X.; Dai, H.; Ren, W.; Tai, Y.; Wang, C.; and Shao, L. 2023. High-resolution Iterative Feedback Network for Camouflaged Object Detection. In *AAAI*.

Huang, Z.; Dai, H.; Xiang, T.-Z.; Wang, S.; Chen, H.-X.; Qin, J.; and Xiong, H. 2023. Feature Shrinkage Pyramid for Camouflaged Object Detection with Transformers. In *CVPR*.

Ji, G.-P.; Fan, D.-P.; Chou, Y.-C.; Dai, D.; Liniger, A.; and Van Gool, L. 2023. Deep Gradient Learning for Efficient Camouflaged Object Detection. *MIR*, 20: 92–108.

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.

Lee, S.; Chung, H.; Kim, J.; and Ye, J. C. 2022. Progressive Deblurring of Diffusion Models for Coarse-to-Fine Image Synthesis. In *NeurIPS 2022 Workshop on Score-Based Methods*.

Li, H.; Yang, Y.; Chang, M.; Chen, S.; Feng, H.; Xu, Z.; Li, Q.; and Chen, Y. 2022. Srdiff: Single image super-resolution with diffusion probabilistic models. *Neurocomputing*, 479: 47–59.

Liu, Z.; Zhang, Z.; Tan, Y.; and Wu, W. 2022. Boosting camouflaged object detection with dual-task interactive transformer. In *ICPR*.

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*.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *CVPR*.

Mei, H.; Ji, G.-P.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D.-P. 2021. Camouflaged Object Segmentation with Distraction Mining. In *CVPR*.

Nafus, M. G.; Germano, J. M.; Perry, J. A.; Todd, B. D.; Walsh, A.; and Swaisgood, R. R. 2015. Hiding in plain sight: a study on camouflage and habitat selection in a slow-moving desert herbivore. *Behavioral Ecology*, 26(5): 1389–1394.

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*.

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

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. BASNet: Boundary-Aware Salient Object Detection. In *CVPR*.

Rahman, A.; Valanarasu, J. M. J.; Hacihaliloglu, I.; and Patel, V. M. 2023. Ambiguous Medical Image Segmentation using Diffusion Models. In *CVPR*.

Song, Y.; and Ermon, S. 2019. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32.

Song, Y.; Sohl-Dickstein, J.; Kingma, D. P.; Kumar, A.; Ermon, S.; and Poole, B. 2020. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*.

Sun, Y.; Wang, S.; Chen, C.; and Xiang, T.-Z. 2022. Boundary-guided camouflaged object detection. In *IJCAI*.

Wang, C.; Yeo, K.; Jin, X.; Duarte, A. C.; Klein, L.; and Elmegreen, B. 2021. S3RP: Self-Supervised Super-Resolution and Prediction for Advection-Diffusion Process. In *Annual Conference on Neural Information Processing Systems*.

Wang, J.; Huang, Q.; Tang, F.; Meng, J.; Su, J.; and Song, S. 2022a. Stepwise feature fusion: Local guides global. In *Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part III*, 110–120. Springer.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2022b. Pvt v2: Improved baselines with pyramid vision transformer. *CVMJ*, 8(3): 415–424.

Whang, J.; Delbracio, M.; Talebi, H.; Saharia, C.; Dimakis, A. G.; and Milanfar, P. 2022. Deblurring via stochastic refinement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 16293–16303.

Wolleb, J.; Sandkühler, R.; Bieder, F.; Valmaggia, P.; and Cattin, P. C. 2022. Diffusion models for implicit image segmentation ensembles. In *International Conference on Medical Imaging with Deep Learning*, 1336–1348. PMLR.

Wu, J.; Fang, H.; Zhang, Y.; Yang, Y.; and Xu, Y. 2022. MedSegDiff: Medical Image Segmentation with Diffusion Probabilistic Model. *arXiv preprint arXiv:2211.00611*.

Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; and Fan, D.-P. 2021. Uncertainty-guided transformer reasoning for camouflaged object detection. In *ICCV*.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D.-P. 2021. Mutual graph learning for camouflaged object detection. In *CVPR*.

Zhang, J.; Fan, D.-P.; Dai, Y.; Anwar, S.; Saleh, F.; Aliakbarian, S.; and Barnes, N. 2021. Uncertainty inspired RGB-D saliency detection. *IEEE transactions on pattern analysis and machine intelligence*, 44(9): 5761–5779.

Zhang, M.; Xu, S.; Piao, Y.; Shi, D.; Lin, S.; and Lu, H. 2022. Preynet: Preying on camouflaged objects. In *ACM MM*.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet:Edge Guidance Network for Salient Object Detection. In *ICCV*.

Zheng, D.; Zheng, X.; Yang, L. T.; Gao, Y.; Zhu, C.; and Ruan, Y. 2023. MFFN: Multi-view Feature Fusion Network for Camouflaged Object Detection. In *WACV*.