Camouflaged Instance Segmentation via Explicit De-camouflaging

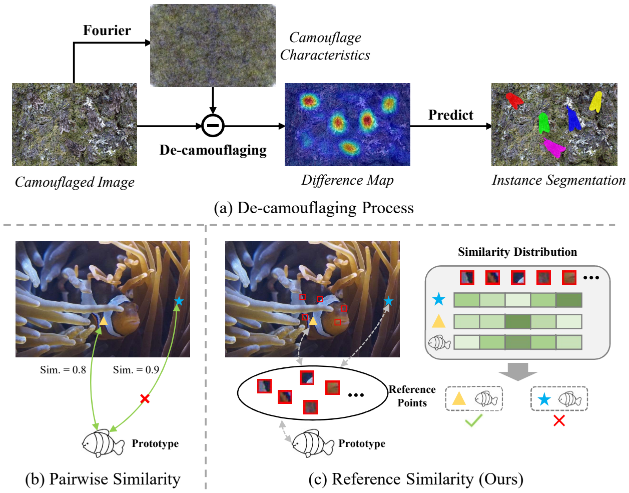
Naisong Luo[1]*, Yuwen Pan[1]*, Rui Sun[1], Tianzhu Zhang[1,2,3]†, Zhiwei Xiong[1,2], Feng Wu[1,2]
[1]University of Science and Technology of China
[2]Institute of Artificial Intelligence, Hefei Comprehensive National Science Center
[3]Deep Space Exploration Lab
{lns6, panyw, issunrui}@mail.ustc.edu.cn, {tzzhang, zwxiong, fengwu}@ustc.edu.cn

Abstract

Camouflaged Instance Segmentation (CIS) aims at predicting the instance-level masks of camouflaged objects, which are usually the animals in the wild adapting their appearance to match the surroundings. Previous instance segmentation methods perform poorly on this task as they are easily disturbed by the deceptive camouflage. To address these challenges, we propose a novel De-camouflaging Network (DCNet) including a pixel-level camouflage decoupling module and an instance-level camouflage suppression module. The proposed DCNet enjoys several merits. First, the pixel-level camouflage decoupling module can extract camouflage characteristics based on the Fourier transformation. Then a difference attention mechanism is proposed to eliminate the camouflage characteristics while reserving target object characteristics in the pixel feature. Second, the instance-level camouflage suppression module can aggregate rich instance information from pixels by use of instance prototypes. To mitigate the effect of background noise during segmentation, we introduce some reliable reference points to build a more robust similarity measurement. With the aid of these two modules, our DCNet can effectively model de-camouflaging and achieve accurate segmentation for camouflaged instances. Extensive experimental results on two benchmarks demonstrate that our DCNet performs favorably against state-of-the-art CIS methods, e.g., with more than 5% performance gains on COD10K and NC4K datasets in average precision.

1. Introduction

In the field of biology, camouflage is defined as a strategy that animals use to adapt their body's appearance (*e.g.*, color and pattern) to match their surroundings in order to achieve concealing and avoid being hunted by predators [37]. Camouflaged Instance Segmentation (CIS) [22] aims at identifying the location and predicting instance-level masks of camouflaged objects, which has attracted more and more attention due to its widespread applications in medical image analysis [11, 44], search-and-rescue work [9] and recreational art [7], *etc.*



(a) De-camouflaging Process

(b) Pairwise Similarity

(c) Reference Similarity (Ours)

Figure 1. Illustration of our motivation. (a) We propose to extract camouflage characteristics to model explicit de-camouflaging for CIS. (b) Pairwise similarity between prototype and pixel is always erroneous. (c) The prototype-pixel correlation is based on the prototype-reference and pixel-reference similarity distribution, which is more accurate.

Despite the tremendous progress in instance segmentation [14, 40, 41], there are few efforts [22, 33] working on Camouflaged Instance Segmentation, as CIS is a more challenging task where most camouflaged instances lack obvious contrast with the background, making general instance segmentation methods work poorly on this task. Recent CIS approaches [22, 33] are generally based on traditional instance segmentation models, either by naively fusing various general instance segmentation models [22] to get better representations, or by directly using global in-

*Equal contribution
†Corresponding author

teraction [33]. However, these approaches fail to explore the core of CIS: de-camouflaging, *i.e*., eliminating camouflage characteristics of the target object. Without explicitly de-camouflaging, the previous methods are easily disturbed by similar background. De-camouflaging is a challenging problem as no image-specific camouflage information is provided as supervision signals. Intuitively, humans have the ability to quickly recognise objects in a highly camouflaged scene. They first repeatedly discriminate the real target characteristics from the camouflage characteristics at the pixel level, and then aggregate the pixel information to discern the whole target instance from the background. This human visual mechanism motivates us to explore de-camouflaging strategy from the pixel level to the instance level in a progressive manner.

In order to model de-camouflage at both the pixel and instance levels, a series of issues need to be considered. (1) The pixel level de-camouflage. In essence, each pixel feature of a camouflaged image contains the camouflage characteristics and the target object characteristics. Note that our goal is to remove camouflage characteristics and maintain target object information, thus a question naturally arises: *How to decouple these two pieces of information at the pixel level?* (2) The instance level segmentation. Based on the de-camouflaged pixels, we can naturally aggregate the pixel information and infer the mask of the target instance. Currently, the transformer-based models [5, 6] have achieved leading performance for instance segmentation, where instance-specific prototypes are learned by constantly interacting with pixel features for final segmentation. However, directly applying the transformer to CIS is not trivial, as the prototypes would frequently absorb deceptive background information that has high similarity with the objects during the interaction, thus failing to discover desired targets accurately. As proved in Figure 1(b), the prototype *fish* is more similar to background pixels than target pixels in the camouflaged image. Therefore, we inevitably face another question: *How to focus on camouflaged instances to achieve de-camouflaging?*

Motivated by the above discussions, we propose an end-to-end **De-camouflaging Network (DCNet)** by jointly modeling pixel-level camouflage decoupling and instance-level camouflage suppression for CIS. In the **Pixel-level Camouflage Decoupling module** (PCD), we focus on decoupling camouflage characteristics and target information fused in the pixel feature. First, we extract camouflage characteristics with the assistance of frequency domain information. The Fourier spectrum amplitude contains low-level statistics [34, 42] (*e.g*., color and texture of the environment, see Figure 5) that accords with the camouflage characteristics. Based on the obtained description of camouflage characteristics, we propose a novel difference attention mechanism to acquire de-camouflaged pixel features. In this mechanism, we calculate the discrepancy between features of the original image and camouflage characteristics (see Figure 1(a)), thereby decoupling the camouflage characteristics and valuable target information while filtering out the background interference. In the **Instance-level Camouflage Suppression module** (ICS), we aggregate the de-camouflaged pixels to achieve final segmentation and meanwhile mitigate the effect of background noise in prototype-pixel interactions. Specifically, we introduce a set of instance prototypes to capture each camouflaged instance through long-range context-aware interactions. To constrain the interactions to favor target pixels over background pixels, we design a novel reference attention mechanism, where we select de-camouflaged pixels with high contribution to prototypes as reference points (see Figure 1(c)). Then we calculate the similarity of prototype-reference and pixel-reference, respectively, thereby obtaining similarity distributions serving as soft multilabel to measure correlations between prototypes and pixels. Highly similar pixels and prototypes must have consistent similarity distributions. In this way, the soft multilabel-based correlation is more accurate than the normal pairwise correlation, as it benefits from consensus among reliable reference points with a global receptive field. As a result, the correlation noise brought by deceptive background can be suppressed and more effective prototypes that contain rich instance information can be obtained.

The contributions of our method could be summarized as follows: (1) We propose a novel De-camouflaging Network (DCNet) by jointly modeling pixel-level camouflage decoupling and instance-level camouflage suppression for CIS. (2) We propose two effective designs in DCNet, *i.e*., difference attention mechanism and reference attention mechanism, which can highlight target information and suppress background interference. (3) Extensive experimental results on two benchmark datasets demonstrate the effectiveness of the proposed method, *e.g*., with more than 5% performance gains on COD10K and NC4K datasets in average precision.

2. Related Work

Since camouflaged instance segmentation is a new emerging task with very few papers, we mainly focus on introducing several lines of research in camouflaged object detection and instance segmentation.

2.1. Camouflaged Object Detection

Camouflaged Object Detection (COD), aiming at localizing camouflaged objects from its background, has a long history in biology and art [10, 23]. Early works [18, 31, 36] in this field focus on distinguishing objects in their camouflaging environment by utilizing handcrafted features, such as texture, boundary, and intensity features. With the significant advances in deep learning (DL), DL-based meth-

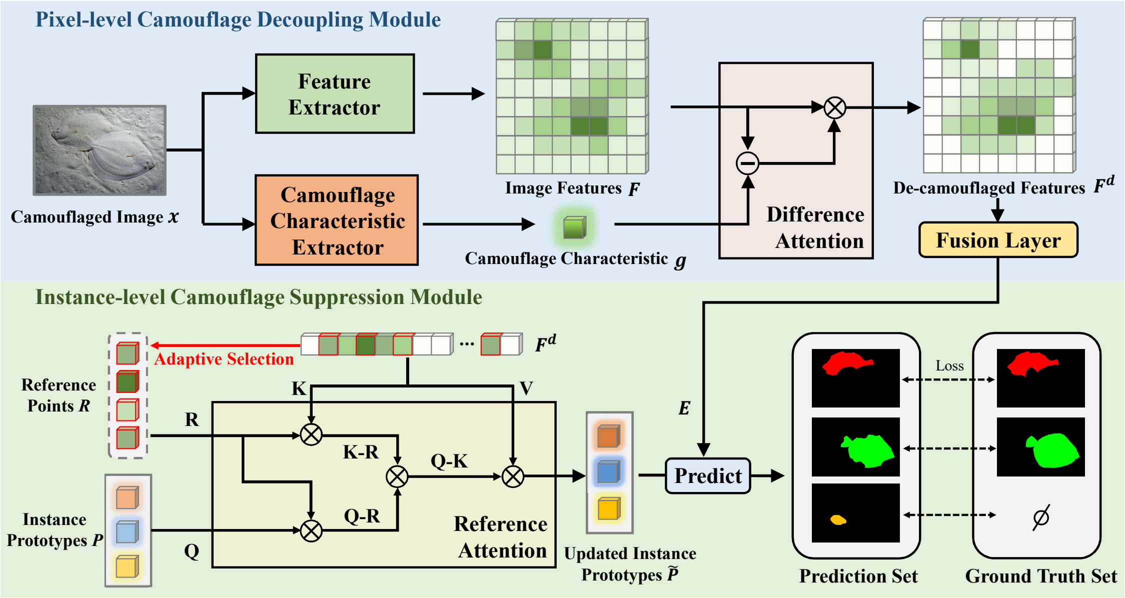


Figure 2. **Illustration of the proposed DCNet.** Our framework mainly consists of two components: a pixel-level camouflage decoupling module and an instance-level camouflage suppression module. (1) The pixel-level camouflage decoupling module aims at extracting camouflaged characteristics and eliminating camouflaged characteristics while reserving target object characteristics to obtain de-camouflaged pixel features. (2) The instance-level camouflage suppression module is responsible for aggregating instance prototypes to achieve instance segmentation meanwhile suppressing the background noise in prototype-pixel interactions.

ods [16, 19, 29, 32, 43, 45, 46] have significantly improved the COD performance through end-to-end learning. Mei *et al.* [29] developed a distraction mining strategy to benefit the accurate segmentation of the camouflaged object. Pang *et al.* [32] proposed a mixed-scale triplet network, ZoomNet, to capture objects in complex scenes at different "zoom" scales. Zhong *et al.* [45] was the first to claim the COD task should go beyond the RGB domain and introduced frequency clues to better detect camouflaged objects. Despite the large progress in camouflaged object detection, camouflaged instance segmentation is a rarely studied task. In this paper, inspired by [34, 42, 45], we introduce frequency domain transform to model the camouflage characteristics and then achieve de-camouflaging for camouflaged instance segmentation. To the best of our knowledge, we are the first to explore explicit de-camouflaging for camouflaged instance segmentation.

2.2. Instance Segmentation

Instance segmentation is a challenging task as it requires both pixel-level and instance-level mask prediction. Existing instance segmentation approaches can be broadly separated into two paradigms: two-stage methods [14, 17, 21] and one-stage methods [1, 40, 41]. Two-stage methods generally utilize object detectors to generate proposal regions and then segment the mask of each detected object. Based on Faster R-CNN [35], Mask R-CNN [14] generates region-of-interests (ROIs) in the first stage and utilizes an extra mask head to acquire the instance segmentation mask. The following works continue to improve the performance for this task in the two-stage manner. Recently, one-stage methods have emerged and driven a new trend in this field. YOLACT [1] combines the prototypes with the mask coefficients to produce instance masks without the detection step. SOLOv2 [41], evolved from SOLO [40], directly decouples the original mask prediction into kernel learning and feature learning to generate final instance segmentation results. Currently, the transformer-based methods [5, 6] introduce instance-specific prototypes to constantly interact with pixel features with the help of some attention mechanisms for final instance segmentation, achieving leading performance. However, the general instance segmentation methods can not directly apply to CIS, as the camouflaged objects present high similarity with the background. Therefore, we base on the one-stage transformer-based paradigm to design an end-to-end De-camouflaging Network, where we introduce reliable reference points to build accurate similarity measurement that is robust to deceptive backgrounds.

3. Our Method

In this section, we first present the overall architecture of the proposed De-camouflage Network (DCNet), and then describe each module in detail.

3.1. Overall Architecture

Our DCNet mainly consists of two modules. (1) The pixel-level camouflage decoupling module aims at extracting camouflage characteristics based on the amplitude information of Fourier spectrum, and eliminating camouflage characteristics via a difference attention mechanism. (2) The instance-level camouflage suppression module is responsible for learning instance prototypes from interactions with pixel features, through a reference attention mechanism to suppress the erroneous correlations caused by deceptive background.

3.2. Pixel-level Camouflage Decoupling

In order to decouple camouflage characteristics and target information fused in the pixel feature, we first model the representation of camouflage characteristics by exploring Fourier frequency domain. Then we design a difference attention mechanism to remove camouflage characteristics while maintaining target information.

Camouflage Characteristics Extraction. Generally, the camouflage characteristics are mainly comprised of color and texture information [10, 22]. On the other hand, in the frequency domain, it is known that the amplitude component of Fourier spectrum preserves low-level statistics information. Therefore, we can utilize the amplitude of Fourier spectrum to represent background information in the camouflaged image. Specifically, given a camouflaged image $x \in \mathbb{R}^{H \times W \times 3}$, its Fourier transformation $\mathcal{F}(x)$ is formulated as:

$$\mathcal{F}(x)_{u,v} = \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} x_{i,j} e^{-J 2\pi \left(\frac{i}{H} u + \frac{j}{W} v\right)}, \tag{1}$$

where J refers to the imaginary unit and each channel of the image is computed independently. The amplitude and phase components are then respectively expressed as:

$$\begin{aligned} \mathcal{A}(x)_{u,v} &= \left[R^2(x)_{u,v} + I^2(x)_{u,v}\right]^{1/2} \\ \mathcal{P}(x)_{u,v} &= \arctan\left[\frac{I(x)_{u,v}}{R(x)_{u,v}}\right], \end{aligned} \tag{2}$$

where $R(x)$ and $I(x)$ represent the real and imaginary part of $\mathcal{F}(x)$, respectively. To represent the camouflaged background, we fix the phase to a constant b (*e.g.*, average value) and reconstruct the image with the amplitude information as

$$\tilde{x} = \mathcal{F}^{-1}[\mathcal{A}(x)_{u,v} e^{-Jb}], \tag{3}$$

where $\mathcal{F}^{-1}$ indicates the inverse Fourier transformation. Then, we feed the reconstructed image $\tilde{x}$ into a light-weight CNN (*e.g.*, ResNet-18 [15]) with global average pooling layer to obtain the global camouflage characteristics $\boldsymbol{g}$ of the original image.

Difference Attention Mechanism. Given image features $\mathbf{F} \in \mathbb{R}^{h \times w \times c}$ extracted from a backbone network (*e.g.*, ResNet-50 [15]), we remove the camouflage characteristics $\boldsymbol{g}$ while retaining valuable target information by a novel difference attention mechanism. In specific, we first utilize two 1×1 convolution layers to map $\mathbf{F}$ and $\boldsymbol{g}$ to the same dimension C, resulting in $\hat{\mathbf{F}} \in \mathbb{R}^{h \times w \times C}$ and $\hat{\boldsymbol{g}} \in \mathbb{R}^{C}$, respectively. The discrepancy between image features and camouflage characteristics indicates the saliency information of the target objects, helping to highlight the target information. Thus, we integrate the difference on all channels for each pixel to obtain a *difference map* $\mathbf{D} \in \mathbb{R}^{h \times w}$:

$$\mathbf{D}_{i,j} = \sum_{k=1}^{C} (\hat{\mathbf{F}}_{i,j,k} - \hat{\boldsymbol{g}}_k)^2, \tag{4}$$

where i, j, and k are the index of height, width, and channel, respectively. By multiplying the original image features with the *difference map* $\mathbf{D}$, we can get the de-camouflaged pixel features $\mathbf{F}^d$ as

$$\mathbf{F}^d = \mathbf{F} \circ \mathbf{D}, \tag{5}$$

where $\circ$ refers to the Hadamard product and $\mathbf{D}$ is broadcasted to the same dimension as $\mathbf{F}$. In addition, we follow [33] to use a deformable self-attention layer [47] to further incorporate context information from other pixels.

Fusion Layer. To acquire fine-grained target information for more accurate segmentation, we use multi-scale features from different stages of the backbone. Specifically, we first obtain the features from different stages of the backbone network, and then obtain the corresponding de-camouflaged features through the difference attention mechanism. Finally, we fuse the processed features with the FPN [24] network to generate the high-resolution pixel-level feature $\mathbf{E}$, which is used for subsequent instance segmentation.

3.3. Instance-level Camouflage Suppression

In order to capture different camouflaged instances, we learn a set of instance prototypes $\mathbf{P} = \{\boldsymbol{p}^i\}_{i=1}^{N}$, and each $\boldsymbol{p}^i \in \mathbb{R}^{1 \times L}$ is responsible for identifying whether each pixel belongs to a camouflaged instance. To eliminate the background interference for prototype learning, we design reference attention to aggregate target information.

Reference Attention Mechanism. Given the de-camouflaged pixel feature $\mathbf{F}^d$ derived from the pixel-level camouflage decoupling module, we first obtain the queries $\mathbf{Q} = [\boldsymbol{q}_1; \boldsymbol{q}_2; ...; \boldsymbol{q}_N]$ from instance prototypes $\mathbf{P}$, keys $\mathbf{K} = [\boldsymbol{k}_1; \boldsymbol{k}_2; ...; \boldsymbol{k}_{hw}]$ and values $\mathbf{V} = [\boldsymbol{v}_1; \boldsymbol{v}_2; ...; \boldsymbol{v}_{hw}]$ from pixel features $\mathbf{F}^d = [\boldsymbol{f}_1; \boldsymbol{f}_2; ...; \boldsymbol{f}_{hw}]$ as

$$\boldsymbol{q}_i = \boldsymbol{p}_i \mathbf{W}^q, \boldsymbol{k}_j = \boldsymbol{f}_j \mathbf{W}^k, \boldsymbol{v}_j = \boldsymbol{f}_j \mathbf{W}^v, \tag{6}$$

where $i \in 1, 2, ..., N, j \in 1, 2, ..., hw$, and $\mathbf{W}^q \in \mathbb{R}^{L\times L}, \mathbf{W}^k, \mathbf{W}^v \in \mathbb{R}^{c\times L}$ are linear projections. With the prototype queries, we can compute the correlation between each query-key pair independently as

$$s_{i,j} = \frac{d\left(\boldsymbol{q}_i, \boldsymbol{k}_j\right)}{\sqrt{L}}, \tag{7}$$

where $d(\cdot,\cdot)$ denotes the distance metric which is the dot product similarity in [39], and $\sqrt{L}$ is a scaling factor. As the camouflaged object presents high similarity with the surrounding backgrounds, the prototype-pixel correlation represented by $s_{i,j}$ is susceptible to background pixel interference, resulting in erroneous segmentation. Intuitively, humans are able to identify camouflaged instances by repeatedly comparing the target with the surrounding reference areas. Motivated by this, we first use an adaptive selection process to select some reliable pixels from de-camouflaged pixels $\mathbf{F}^d$ as references. In specific, to measure the reliability of each pixel, we define and calculate the total contribution of each pixel with regard to all the prototypes, which is formulated as:

$$u_j = \sum_{i=1}^{N} s_{i,j}, j \in 1, 2, ..., hw. \tag{8}$$

Then we select top-K pixels with the largest correlations with the prototypes as reference points $\mathbf{R} \in \mathbb{R}^{K\times L}$. In this way, we can pick out reliable reference points that are adaptive to the instance prototypes. Based on reference points, we first calculate the prototype-reference and pixel-reference similarity as

$$s_i^q = \phi_1(\boldsymbol{q}_i\mathbf{R}^\top), s_j^k = \phi_2(\boldsymbol{k}_j\mathbf{R}^\top), \tag{9}$$

where ϕ_1 and ϕ_2 are feed-forward networks (FFN) containing two fully connected layers. s_i^q and s_j^k can serve as soft multilabel to measure the correlation between the i-th prototype and the j-th pixel:

$$s_{i,j}^{qk} = d(\boldsymbol{q}_i, \boldsymbol{k}_j; \mathbf{R}) = s_i^q(s_j^k)^\top. \tag{10}$$

Compared to $s_{i,j}$, $s_{i,j}^{qk}$ leverages the consensus among reliable reference points, thus suppressing the background noises and leading to fewer erroneous correlations. Given the correlation $s_{i,j}^{qk}$, we can extract and purify target information from pixel features to update instance prototypes as

$$\tilde{\boldsymbol{p}}_i = \sum_{j=1}^{hw} a_{i,j}\boldsymbol{v}_j, a_{i,j} = \frac{\exp(s_{i,j}^{qk})}{\sum_{j=1}^{hw}\exp(s_{i,j}^{qk})}. \tag{11}$$

Note that we have omitted the scaling factor and the multi-head mechanism for notation simplicity.

Given the high-resolution pixel-level feature embedding $\mathbf{E}$ and the learned instance prototypes $\tilde{\boldsymbol{p}}_i$, we can obtain N mask predictions and according scores by

$$m_i = \text{sigmoid}(\mathbf{E}\tilde{\boldsymbol{p}}_i^\top), y_i = h(\tilde{\boldsymbol{p}}_i \mid \sigma), \tag{12}$$

where $h(\cdot \mid \sigma)$ is a classifier parameterized by σ to predict the confidence score. Following [6], we match a ground truth label (including “no object”) for each predicted segment by the Hungarian matching algorithm. For training, we use a cross-entropy loss to constrain the instance score y_i, and a linear combination of a focal loss [25] and a dice loss [30] to constrain the mask prediction m_i.

4. Experiment

4.1. Experiment Setup

Datasets. We evaluate our method on two datasets: COD10K [10] and NC4K [28]. COD10K contains 5086 camouflaged images with high-quality instance-level annotations, which are divided into 3040 training images and 2026 testing images. NC4K contains 4121 test camouflaged images to evaluate the generalization ability of CIS models. Following the standard benchmark [33], we use the instance-level annotations in COD10K to train our DCNet and evaluate it on the test set of COD10K and NC4K.

Evaluation Metrics. We adopt AP$_{50}$, AP$_{75}$ and AP scores as evaluation metrics to quantify the effectiveness of our method. A true positive is counted if the intersection over union (IoU) between the ground truth and the segmentation is at least 50% or 75%, and the AP score is an overall metric that combines different IoU thresholds.

Implementation Details. For fair comparisons, the feature extractor is implemented by ResNet-50 [15] which is pre-trained on ImageNet [8]. During our training, our model is trained with batch size of 2, using the Adam optimizer [27] with an initial learning rate of 0.0001 for 90,000 iterations. We set the channel dimension $C = L = 256$. In the instance-level camouflage suppression module, we set the number of prototypes N as 10, and the number of reference points K as 64, which turns out to generate the best segmentation during our experiments as shown in Figure 4.

4.2. Comparison with State-of-the-arts

Quantitative results. Considering that CIS is a newly emerging task, we also compare some general instance segmentation methods besides the previous CIS methods and all the methods adopt ResNet-50 [15] backbone for fair comparisons. As demonstrated in Table 1, the proposed DCNet consistently outperforms the state-of-the-art methods by a large margin on both datasets. **(1) On the COD10K** [10], our method outperforms the previous best method OSFormer [33] by **4.3%** in AP, indicating that our DCNet is able to acquire more accurate segmentation of camouflaged targets. **(2) On the NC4K** [28], Our model yields an accuracy of 52.8% in AP, making an obvious performance improvement by **7.0%** in AP compared to the second best performing method Mask2Former [5]. Since the model is trained on the COD10K training set, the high

Table 1. Comparisons of existing CIS methods and general instance approaches on COD10K [10] and NC4K [28] testing set, where all the methods use ResNet-50 as feature extractor. The best results are shown in **bold**.

Methods	COD10K-Test			NC4K-Test			Params(M)	GFLOPs
	AP	AP_{50}	AP_{75}	AP	AP_{50}	AP_{75}		
Mask R-CNN [14]	25.0	55.5	20.4	27.7	58.6	22.7	43.9	186.3
MS R-CNN [17]	30.1	57.2	28.7	31.0	58.7	29.4	60.0	198.5
Cascade R-CNN [2]	25.3	56.1	21.3	29.5	60.8	24.8	71.7	334.1
HTC [4]	28.1	56.3	25.1	29.8	59.0	26.6	76.9	331.7
BlendMask [3]	28.2	56.4	25.2	27.7	56.7	24.2	35.8	233.8
Mask Transfiner [20]	28.7	56.3	26.4	29.4	56.7	27.2	44.3	185.1
YOLACT [1]	24.3	53.3	19.7	32.1	65.3	27.9	-	-
CondInst [38]	30.6	63.6	26.1	33.4	67.4	29.4	34.1	200.1
QueryInst [12]	28.5	60.1	23.1	33.0	66.7	29.4	-	-
SOTR [13]	27.9	58.7	24.1	29.3	61.0	25.6	63.1	476.7
SOLOv2 [41]	32.5	63.2	29.9	34.4	65.9	31.9	46.2	318.7
MaskFormer [6]	38.2	65.1	37.9	44.6	71.9	45.8	45.0	174.2
Mask2Former [5]	39.4	67.7	38.5	45.8	73.6	47.5	43.9	241.0
OSFormer [33]	41.0	**71.1**	40.8	42.5	72.5	42.3	46.6	324.7
DCNet (ours)	**45.3**	70.7	**47.5**	**52.8**	**77.1**	**56.5**	53.4	207.0

Table 2. Comparison of the different components in pixel-level camouflage decoupling module in terms of AP scores.

	COD10K	NC4K
PCD (ours)	**45.3**	**52.8**
−camouflage attribute extractor	$43.1_{(-2.2)}$	$48.1_{(-4.7)}$
−difference attention mechanism	$44.1_{(-1.2)}$	$51.3_{(-1.5)}$
−both 2 components above	$40.2_{(-5.1)}$	$46.4_{(-6.4)}$

Table 3. Comparisons of different attention mechanisms in the instance-level camouflage suppression module in terms of AP scores.

Attention Mechanism	COD10K	NC4K
cross-attention [39]	42.4	49.5
masked attention [5]	44.7	51.7
reference attention (ours)	**45.3**	**52.8**

performance on NC4K indicates that our method also has better generalization ability. Notice that our method does not achieve the highest performance on AP_{50} metric for the NC4K because different from the previous CIS method, we do not use additional Non-maximum suppression (NMS) post-processing to remove redundant predictions. For AP_{50}, our more accurate segmentation results cannot increase the number of true positive (TP) under the 50% IoU threshold, while redundant predictions will increase the number of false positive (FP), thus our method has no obvious advantages. For AP_{75}, our high precision segmentation will increase TP with a large margin. The overall metric AP reflects the superiority of our method over other methods. There is a similar trend on NC4K. In addition, the lower GFLOPs of our DCNet compared to state-of-the-art methods indicates that our model achieves high segmentation ability while maintaining high computation efficiency.

Qualitative Results. As shown in Figure 3, our proposed method is capable of de-camouflaging by using the difference attention mechanism at the pixel level, and also able to identify the accurate locations of target camouflaged objects at the instance level. Compared to previous methods, our DCNet performs better at boundaries of camouflage instances (see the first 4 columns), suppresses distracting background regions (see 5^{th} to 6^{th} columns), and distinguishes multiple instances well (see the last 2 columns).

4.3. Ablation Study

We conduct comprehensive ablation studies on COD10K and NC4K to verify the effectiveness of our modules.

Effectiveness of the Pixel-level Camouflage Decoupling Module. The pixel-level camouflage decoupling module mainly consists of a camouflage attribute extractor and a difference attention mechanism. We validate the importance of each component by removing them one at a time. As shown in Table 2, we replace the proposed camouflage attribute extractor with a learnable vector, and the performance drops by an average of 3.5% across two datasets. It indicates that the Fourier spectrum amplitude is effective in modeling the camouflage characteristics, verifying that it is hard to learn de-camouflage without image-specific camouflage information supervision. Then we replace the difference attention mechanism with general subtraction operations. The reduced performance indicates that our pro-

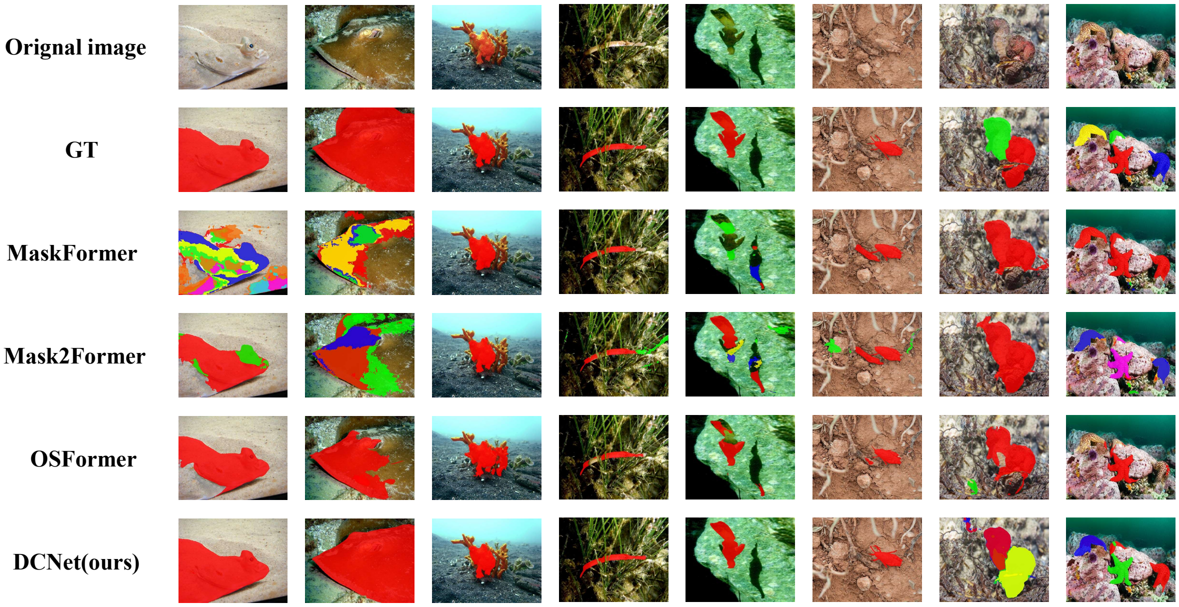


Figure 3. Qualitative results of DCNet compared with state-of-the-art methods. Different colored masks represent different instances.

Table 4. Comparison of using different scales of features in terms of AP scores.

	COD10K	NC4K
single scale (1/32)	45.2	52.5
single scale (1/16)	44.8	51.6
single scale (1/8)	45.1	52.2
multiple scales	**45.3**	**52.8**

Table 5. Comparisons with OSFormer [33] of different backbones on COD10K [10] and NC4K [28] in terms of AP scores.

Method	Backbone	COD10K	NC4K
OSFormer	R50	41.0	42.5
DCNet (ours)		45.3	52.8
OSFormer	R101	42.0	44.4
DCNet (ours)		46.8	53.5
OSFormer	Swin-T	47.7	50.2
DCNet (ours)		50.3	56.3
OSFormer	Swin-S	52.1	56.7
DCNet (ours)		52.3	58.4

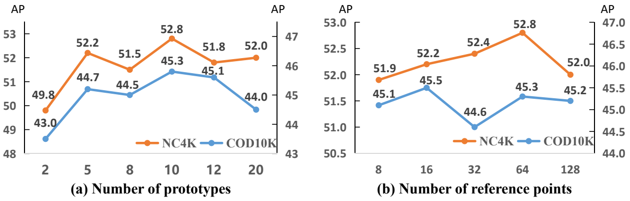


Figure 4. Comparisons of performance with different numbers of prototypes and reference points on COD10K [10] and NC4K [28].

posed mechanism can well model the discrepancy between image features and camouflage characteristics. Removing both of the components leads to huge performance degradation, showing that the combination of them is essential in improving the camouflaged object segmentation.

Effectiveness of the Instance-level Camouflage Suppression Module. As shown in Table 3, the reference attention mechanism in our instance-level camouflage suppression module has the highest AP score on both datasets. Compared to vanilla cross-attention [39], our proposed reference points can effectively suppress the noise brought by background pixels in the prototype-pixel interaction, thus obtaining significant improvement for accurate segmentation. Meanwhile, our method is superior to existing cross-attention variants, such as the masked attention [5] in the segmentation field.

Effectiveness of Hierarchical Features To explore decamouflage at multiple feature levels, we feed 3 scales of features (*i.e.*, stride of 32, 16, and 8) into the difference attention mechanism, respectively. As shown in Table 4, low-resolution features exceed other scales. This may attribute to the retention of object semantic information. The highest performance in this experiment shows that the contribution of each scale is complementary to each other. Using single-scale features alone is not enough to model the target information, thus combining them together can further boost the performance.

Compatibility for Feature Extractor. We equip our DCNet with different feature extractor backbones, *i.e.*, ResNet-

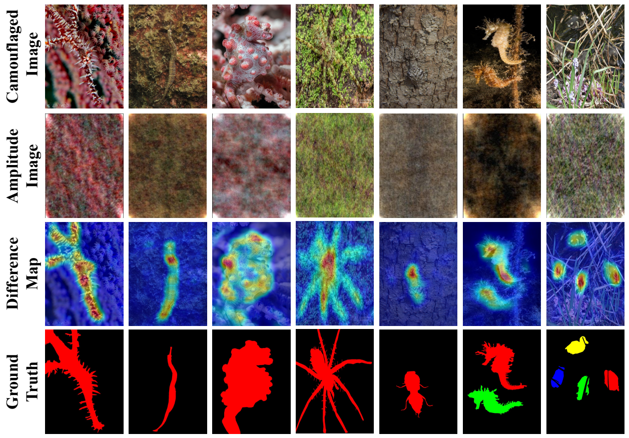


Figure 5. Illustration of the amplitude images and difference maps in our proposed difference attention.

50 [15], ResNet-101 [15], Swin-T [26], and Swin-S [26]. All of them are pre-trained on ImageNet-1K [8]. As shown in Table 5, our method performs the best in all cases on both datasets, proving that our approach has tremendous potential for further improvement.

Analysis of Hyperparameters. In the instance-level camouflage suppression module, prototypes play an important role in capturing instances by aggregating object pixels. As shown in Figure 4(a), we conduct quantitative experiments to analyze how many prototypes are suitable for instance segmentation. It can be seen that too many or too few prototypes will damage the AP metric. In fact, the number of prototypes N should be larger than the count of instances in an image to avoid instance fusion error. While superabundant prototypes will cause many prototypes to match "no object" during training. We choose $N = 10$ by balancing the performance on both datasets. Besides, in the reference attention mechanism, the number of reference points K influences the similarity distributions between prototypes and pixels. As shown in Figure 4(b), we can observe that the performance on NC4K continues to grow until $K = 64$, which means that it is sufficient for building appropriate prototype-pixel correction.

4.4. Visualizations

Visualization of Difference Map. As shown in Figure 5, the 2^{nd} row exhibits the amplitude images that are generated through Fourier transformation, which can be considered as camouflage characteristics. With the help of the difference attention mechanism, our network can remove the confusing background and localize the accurate salient areas (see 3^{rd} row), which turn out to be the correct regions of camouflaged instances.

Visualization of Reference Attention. To explore the reference attention mechanism's ability to suppress the background noise, we make qualitative visualization to compare different attention mechanisms. As shown in Figure 6 (d), with the proposed reference attention, the high response is clustered in the foreground area and the prototype pays little attention to the background area. In contrast, in Figure 6 (b), prototypes with the cross-attention [39] tend to absorb pixels from the non-target region. And in in Figure 6 (c), with the masked attention [5], most background pixels are masked, while some foreground pixels are also suppressed, leading to sub-optimal performance. Thus, the reference mechanism plays an important role in mitigating the effect of background noise in prototype-pixel interactions, which helps to achieve more accurate segmentation.

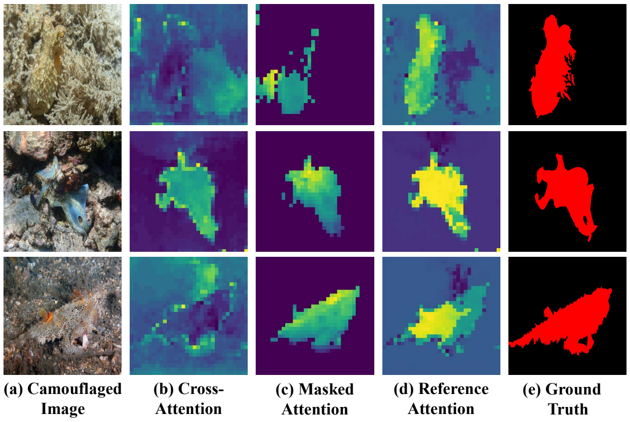


Figure 6. Visualizations of the activation map with different attention mechanisms.

5. Conclusion

In this paper, we propose an end-to-end Decamouflaging Network (DCNet) by jointly modeling pixel-level camouflage decoupling and instance-level camouflage suppression for camouflaged instance segmentation. We design a Pixel-level Camouflage Decoupling module to model and further eliminate the camouflage characteristics based on the frequency domain information. Besides, an Instance-level Camouflage Suppression module is proposed to achieve prototype-based instance segmentation and mitigate the effect of background noise by introducing reference points. Extensive experimental results on two benchmark datasets demonstrate the effectiveness of the proposed method.

6. Acknowledgment

This work was partially supported by the National Nature Science Foundation of China (Grant 62022078, 62021001), National Defense Basic Scientific Research Program of China (Grant JCKY2021601B013).

References

[1] Daniel Bolya, Chong Zhou, Fanyi Xiao, and Yong Jae Lee. Yolact: Real-time instance segmentation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9157–9166, 2019. 3, 6

[2] Zhaowei Cai and Nuno Vasconcelos. Cascade r-cnn: high quality object detection and instance segmentation. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1483–1498, 2019. 6

[3] Hao Chen, Kunyang Sun, Zhi Tian, Chunhua Shen, Yongming Huang, and Youliang Yan. Blendmask: Top-down meets bottom-up for instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8573–8581, 2020. 6

[4] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4974–4983, 2019. 6

[5] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1290–1299, 2022. 2, 3, 5, 6, 7, 8

[6] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. *Advances in Neural Information Processing Systems*, 34:17864–17875, 2021. 2, 3, 5, 6

[7] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM Trans. Graph.*, 29(4):51–1, 2010. 1

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 5, 8

[9] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 1

[10] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2777–2787, 2020. 2, 4, 5, 6, 7

[11] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 39(8):2626–2637, 2020. 1

[12] Yuxin Fang, Shusheng Yang, Xinggang Wang, Yu Li, Chen Fang, Ying Shan, Bin Feng, and Wenyu Liu. Instances as queries. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6910–6919, 2021. 6

[13] Ruohao Guo, Dantong Niu, Liao Qu, and Zhenbo Li. Sotr: Segmenting objects with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7157–7166, 2021. 6

[14] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *Proceedings of the IEEE international conference on computer vision*, pages 2961–2969, 2017. 1, 3, 6

[15] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 4, 5, 8

[16] Xiaobin Hu, Deng-Ping Fan, Xuebin Qin, Hang Dai, Wenqi Ren, Ying Tai, Chengjie Wang, and Ling Shao. High-resolution iterative feedback network for camouflaged object detection. *arXiv preprint arXiv:2203.11624*, 2022. 3

[17] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6409–6418, 2019. 3, 6

[18] Iván Huerta, Daniel Rowe, Mikhail Mozerov, and Jordi Gonzàlez. Improving background subtraction based on a casuistry of colour-motion segmentation problems. In *Iberian Conference on Pattern Recognition and Image Analysis*, pages 475–482. Springer, 2007. 2

[19] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *Machine Intelligence Research*, 20(1):92–108, 2023. 3

[20] Lei Ke, Martin Danelljan, Xia Li, Yu-Wing Tai, Chi-Keung Tang, and Fisher Yu. Mask transfiner for high-quality instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4412–4421, 2022. 6

[21] Alexander Kirillov, Yuxin Wu, Kaiming He, and Ross Girshick. Pointrend: Image segmentation as rendering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9799–9808, 2020. 3

[22] Trung-Nghia Le, Yubo Cao, Tan-Cong Nguyen, Minh-Quan Le, Khanh-Duy Nguyen, Thanh-Toan Do, Minh-Triet Tran, and Tam V Nguyen. Camouflaged instance segmentation in-the-wild: Dataset, method, and benchmark suite. *IEEE Transactions on Image Processing*, 31:287–300, 2021. 1, 4

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

[24] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017. 4

[25] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision*, pages 2980–2988, 2017. 5

[26] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10012–10022, 2021. 8

[27] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 5

[28] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11591–11601, 2021. 5, 6, 7

[29] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8772–8781, 2021. 3

[30] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *2016 fourth international conference on 3D vision (3DV)*, pages 565–571. IEEE, 2016. 5

[31] Yuxin Pan, Yiwang Chen, Qiang Fu, Ping Zhang, and Xin Xu. Study on the camouflaged target detection method based on 3d convexity. *Modern Applied Science*, 5(4):152, 2011. 2

[32] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan Lu. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2160–2170, 2022. 3

[33] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *European Conference on Computer Vision*, pages 19–37. Springer, 2022. 1, 2, 4, 5, 6, 7

[34] Leon N Piotrowski and Fergus W Campbell. A demonstration of the visual importance and flexibility of spatial-frequency amplitude and phase. *Perception*, 11(3):337–346, 1982. 2, 3

[35] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3

[36] P Sengottuvelan, Amitabh Wahi, and A Shanmugam. Performance of decamouflaging through exploratory image analysis. In *2008 First International Conference on Emerging Trends in Engineering and Technology*, pages 6–10. IEEE, 2008. 2

[37] Sujit K Singh, Chitra A Dhawale, and Sanjay Misra. Survey of object detection methods in camouflaged image. *IERI Procedia*, 4:351–357, 2013. 1

[38] Zhi Tian, Chunhua Shen, and Hao Chen. Conditional convolutions for instance segmentation. In *European conference on computer vision*, pages 282–298. Springer, 2020. 6

[39] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 5, 6, 7, 8

[40] Xinlong Wang, Tao Kong, Chunhua Shen, Yuning Jiang, and Lei Li. Solo: Segmenting objects by locations. In *European Conference on Computer Vision*, pages 649–665. Springer, 2020. 1, 3

[41] Xinlong Wang, Rufeng Zhang, Tao Kong, Lei Li, and Chunhua Shen. Solov2: Dynamic and fast instance segmentation. *Advances in Neural information processing systems*, 33:17721–17732, 2020. 1, 3, 6

[42] Qinwei Xu, Ruipeng Zhang, Ya Zhang, Yanfeng Wang, and Qi Tian. A fourier-based framework for domain generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14383–14392, 2021. 2, 3

[43] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4146–4155, 2021. 3

[44] Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Automatic polyp segmentation via multi-scale subtraction network. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pages 120–130. Springer, 2021. 1

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

[46] Tao Zhou, Yi Zhou, Chen Gong, Jian Yang, and Yu Zhang. Feature aggregation and propagation network for camouflaged object detection. *IEEE Transactions on Image Processing*, 31:7036–7047, 2022. 3

[47] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *arXiv preprint arXiv:2010.04159*, 2020. 4