Camouflaged Object Segmentation with Distraction Mining

Haiyang Mei[1] Ge-Peng Ji[2,4] Ziqi Wei[3,⋆] Xin Yang[1,⋆] Xiaopeng Wei[1] Deng-Ping Fan[4]
[1] Dalian University of Technology [2] Wuhan University [3] Tsinghua University [4] IIAI
https://mhaiyang.github.io/CVPR2021_PFNet/index

Abstract

Camouflaged object segmentation (COS) aims to identify objects that are "perfectly" assimilate into their surroundings, which has a wide range of valuable applications. The key challenge of COS is that there exist high intrinsic similarities between the candidate objects and noise background. In this paper, we strive to embrace challenges towards effective and efficient COS. To this end, we develop a bio-inspired framework, termed ***P****ositioning and* ***F****ocus* ***Net****work (****PFNet****), which mimics the process of predation in nature. Specifically, our PFNet contains two key modules, i.e., the positioning module (PM) and the focus module (FM). The PM is designed to mimic the detection process in predation for positioning the potential target objects from a global perspective and the FM is then used to perform the identification process in predation for progressively refining the coarse prediction via focusing on the ambiguous regions. Notably, in the FM, we develop a novel* ***distraction mining strategy*** *for the distraction discovery and removal, to benefit the performance of estimation. Extensive experiments demonstrate that our PFNet runs in real-time (72 FPS) and significantly outperforms 18 cutting-edge models on three challenging datasets under four standard metrics.*

1. Introduction

Camouflage is the concealment of animals or objects by any combination of material, coloration, or illumination, for making the target objects hard to see (crypsis) or disguising them as something else (mimesis) [47]. Benefiting from the capability of finding out the camouflaged objects that are "seamlessly" embedded in their surroundings, camouflaged object segmentation (COS) has a wide range of valuable applications in different fields, ranging from medical diagnosis (*e.g.*, polyp segmentation [13] and lung infection segmentation [14]), industry (*e.g.*, inspection of unqualified products on the automatic production line), agriculture (*e.g.*, locust detection to prevent invasion), security and surveillance (*e.g.*, search-and-rescue mission and the detection of pedestrians or obstacles in bad weather for automatic driving), scientific research (*e.g.*, rare species discovery), to art (*e.g.*, photo-realistic blending and recreational art).

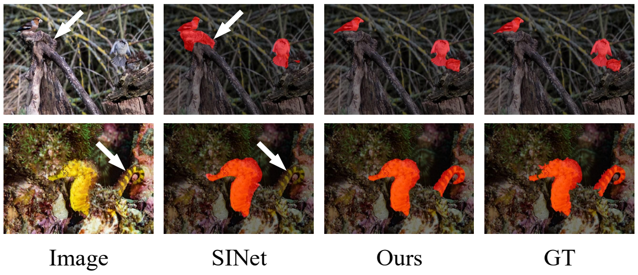


Figure 1. Visual examples of camouflaged object segmentation. While the state-of-the-art method SINet [12] confused by the background region which shares similar appearance with the camouflaged objects (pointed to by a arrow in the top row) or the camouflaged region that cluttered in the background (pointed to by a arrow in the bottom row), our method can eliminate these distractions and generate accurate segmentation results.

However, COS is a fundamentally challenging task due to the fact that the camouflage strategy works by deceiving the visual perceptual system of the observer [47] and thus a significant amount of visual perception knowledge [50] is required to eliminate the ambiguities caused by the high intrinsic similarities between the target object and the background. Research into camouflaged object segmentation has a long and rich history in many fields such as biology and art [47]. Early methods are dedicated to distinguishing the foreground and background based on handcrafted low-level features such as texture [45], 3D convexity [39] and motion [28]. These features, however, have limited capability to distinguish between the camouflaged and non-camouflaged objects, so the approaches based on them often fail in complex scenes. Despite the recently proposed deep learning-based approaches [26, 12, 58] have achieved performance improvement to some extent, there is still a large room for exploring the effective way of accurate COS.

In nature, prey animals make use of mechanisms such

[⋆] Xin Yang (xinyang@dlut.edu.cn) and Ziqi Wei are the corresponding authors.

as camouflage to misdirect the visual sensory mechanisms of predators for reducing the risk of being detected [47]. Under the pressure of natural selection, predatory animals have evolved a variety of adaptations such as sharp senses and intelligent brains for the successful predation which can be divided into three stages, *i.e.*, detection, identification, and capture [15]. This motivates our bio-inspired solution to segment camouflaged objects by mimicking the first two stages of predation.

In this paper, we propose a positioning and focus network (PFNet) which greatly improves the existing camouflaged object segmentation performance. Our PFNet contains two key modules, *i.e.*, the positioning module (PM) and the focus module (FM). The PM is designed to mimic the detection process in predation for positioning the potential target objects from a global perspective and the FM is then used to perform the identification process in predation for refining the initial segmentation results by focusing on the ambiguous regions. Specifically, the PM consists of a channel attention block and a spatial attention block and both of them are implemented in a non-local way to capture long-range semantic dependencies in terms of channel and spatial position for inferring the initial location of the target objects from a global perspective. The FM first perform multi-scale context exploration based on the foreground-attentive (background-attentive) features for discovering the false-positive (false-negative) distractions and then remove these distractions to get the purer representations about the target objects. Such distraction mining strategy is implemented in an implicit way and is applied on different levels of features to progressively refine the segmentation results, enabling our PFNet to possess the strong capability of accurately segmenting the camouflaged objects (see Figure 1 as an example). To sum up, our contributions are as follows:

- We introduce the concept of distraction to the COS problem and develop a novel distraction mining strategy for distraction discovery and removal, to benefit the accurate segmentation of the camouflaged object.

- We propose a new COS framework, named positioning and focus network (PFNet), which first positioning the potential target objects by exploring long-range semantic dependencies and then focuses on distraction discovery and removal to progressively refine the segmentation results.

- We achieve state-of-the-art camouflaged object segmentation performance on three benchmark datasets. Experimental results demonstrate the effectiveness of our method.

2. Related Work

Generic Object Detection (GOD) seeks to locate object instances from several predefined generic categories in natural images [31], which is one of the most fundamental and challenging problems in computer vision and forms the basis for solving complex or high-level vision tasks such as segmentation [23], scene understanding [29], and object tracking [61]. The generic objects in a scene can be either conspicuous or camouflaged, and the camouflaged ones can be seen as hard cases. Therefore, directly applying GOD methods (*e.g.*, [30, 17, 22]) to segment camouflaged objects may not get the desired results.

Salient Object Detection (SOD) aims to identify and segment the most attention-grabbing object(s) in an input image. Hundreds of image-based SOD methods have been proposed in the past decades [9]. Early methods are mainly based on the handcrafted low-level features as well as heuristic priors (*e.g.*, color [1] and contrast [6]). Recently, deep convolutional neural networks (CNNs) have set new state-of-the-art on salient object detection. Multi-level feature aggregation is explored for robust detection [27, 19, 63, 68]. Recurrent and iterative learning strategies are also employed to refine the saliency map progressively [64, 52]. Due to the effectiveness for feature enhancement, attention mechanisms [51, 54] are also applied to saliency detection [32, 4]. In addition, edge/boundary cues are leveraged to refine the saliency map [43, 67, 48]. However, applying the above SOD approaches for camouflaged object segmentation may not appropriate as the term "salient" is essentially the opposite of "camouflaged", *i.e.*, standout versus immersion.

Specific Region Segmentation (SRS) we defined here refers to segmenting the specific region such as shadow [20, 25, 72, 70], mirror [59, 36], glass [38, 57] and water [16] region in the scene. Such regions are special and has a critical impact on the vision systems. For the water, shadow and mirror region, there typically exists intensity or content discontinuities between the foreground and background. Instead, both the intensity and content are similar between the camouflaged objects and the background, leading to a great challenge of COS. Besides, the camouflaged objects are typically with more complex structures, compared with the glass region, and thus increasing the difficulty of accurate segmentation.

Camouflaged Object Segmentation (COS) has a long and rich research history in many fields such as biology and art [47], and hugely influenced by two remarkable studies [49, 7]. Early works related to camouflage are dedicated to distinguishing foreground and background based on the handcrafted low-level features such as texture [45], 3D convexity [39] and motion [28]. These methods work for a few simple cases but often fail in complex scenes. Recently, Le *et al.* [26] propose an end-to-end network for camouflaged

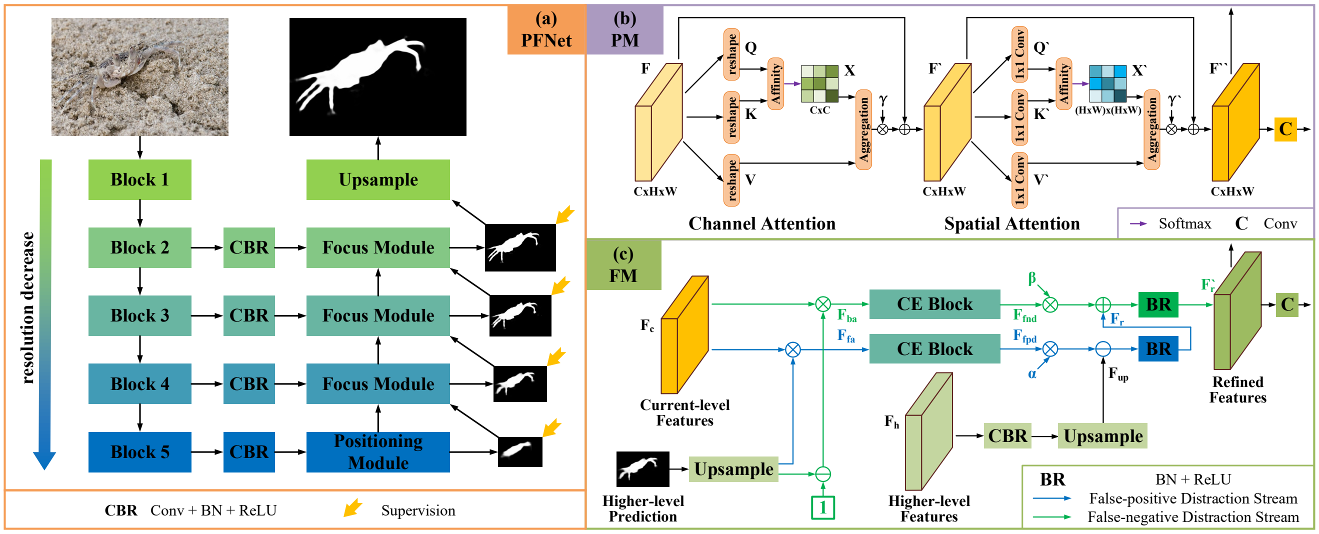


Figure 2. (a) Overview of our positioning and focus network (PFNet) and its two main building blocks: (b) a positioning module (PM) and (c) a focus module (FM).

object segmentation through integrating classification information into pixel-level segmentation. Yan *et al.* [58] further introduce the adversarial attack to boost up the segmentation accuracy. Fan *et al.* [12] develop a simple yet effective framework, termed as *SINet*, and construct the current largest COS dataset *COD10K* to facilitate the advance of the COS in the deep learning era.

Contextual Feature Learning plays an important role in achieving high performance for many computer vision tasks. Many works are devoted to exploiting contexts to enhance the ability of feature representation. Specifically, multi-scale contexts are developed in [3, 65, 37] and multi-level contexts are extracted in [60, 62]. Large-field contextual features are captured in [42, 38], direction-aware contexts are explored in [20], and contextual contrasted features are leveraged in [8, 59]. However, exploring contextual features indiscriminately may not contribute much to COS as the contexts would often be dominated by features of conspicuous objects. Our method differs from the above works by focusing on exploring contexts from the foreground/background-attentive features for contextual reasoning and distraction discovery. And we validate the effectiveness of our method by the experiments.

3. Methodology

It has been pointed in the biological study [15] that the process of predation can be broken down into three stages, *i.e.*, detection, identification and capture. Inspired by the first two stages of predation, we design a positioning and focus network (PFNet) which consists of two key modules, *i.e.*, the positioning module (PM) and the focus module (FM). The PM is designed to mimic the detection process in predation for positioning the potential target objects from a global perspective and the FM is then used to perform the identification process in predation for refining the initial segmentation results by focusing on the ambiguous regions.

3.1. Overview

The overview of our proposed network is shown in Figure 2 (a). Given a single RGB image, we first feed it into a ResNet-50 [18] backbone to extract multi-level features which are further fed into four convolution layers for channel reduction. Then, a positioning module (PM) is applied on the highest-level features to locate the potential target objects. Finally, multiple focus modules (FMs) are leveraged to progressively discover and remove both false-positive and false-negative distractions, for the accurate identification of the camouflaged object.

3.2. Positioning Module

Figure 2 (b) illustrates the detailed structure of the well-designed positioning module (PM). Given the input highest-level features, the PM aims to harvest semantic-enhanced high-level features and further generate the initial segmentation map. It consists of a channel attention block and a spatial attention block. Both of them are implemented in a non-local way, to capture long-range dependencies in terms of channel and spatial position, for enhancing the semantic representation of the highest-level features from a global perspective.

Specifically, given the input features $F \in \mathbb{R}^{C \times H \times W}$, where C, H, and W represent the channel number, height, and width, respectively, we first reshape F to get the query

Q, key K, and value V, respectively, where $\{Q, K, V\} \in \mathbb{R}^{C\times N}$ and $N = H \times W$ is the number of pixels. Then we perform a matrix multiplication between Q and the transpose of K, and apply a softmax layer to calculate the channel attention map $X \in \mathbb{R}^{C\times C}$:

$$x_{ij} = \frac{exp(Q_{i:} \cdot K_{j:})}{\sum_{j=1}^{C} exp(Q_{i:} \cdot K_{j:})}, \qquad (1)$$

where $Q_{i:}$ denotes the i-th row of matrix Q and x_{ij} measures the j^{th} channel's impact on the i^{th} channel. After that, we perform a matrix multiplication between X and V and reshape the aggregated attentive features to $\mathbb{R}^{C\times H\times W}$. Finally, to enhance the fault-tolerant ability, we multiply the result by a learnable scale parameter γ and perform an identify mapping operation to obtain the final output $F' \in \mathbb{R}^{C\times H\times W}$:

$$F'_{i:} = \gamma \sum_{j=1}^{C} (x_{ij} V_{j:}) + F_{i:}, \qquad (2)$$

where γ gradually learns a weight from an initial value of 1. The final feature F' models the long-range semantic dependencies between the channels of feature maps and thus is more discriminative than the input feature F.

Then, we feed the output features of channel attention block into the spatial attention block as the input. We first employ three 1×1 convolution layers on the input features F' and reshape the convolution results to generate three new feature maps Q', K', and V', respectively, where $\{Q', K'\} \in \mathbb{R}^{C_1\times N}$ and $C_1 = C/8$, and $V' \in \mathbb{R}^{C\times N}$. After that we perform a matrix multiplication between the transpose of Q' and K', and use the softmax normalization to generate the spatial attention map $X' \in \mathbb{R}^{N\times N}$:

$$x'_{ij} = \frac{exp(Q'_{:i} \cdot K'_{:j})}{\sum_{j=1}^{N} exp(Q'_{:i} \cdot K'_{:j})}, \qquad (3)$$

where $Q'_{:i}$ denotes the i-th column of matrix Q' and x'_{ij} measures the j^{th} position's impact on the i^{th} position. Meanwhile, we conduct a matrix multiplication between V' and the transpose of X' and reshape the result to $\mathbb{R}^{C\times H\times W}$. Similar to the channel attention block, we multiply the result by a learnable scale parameter γ' and add a skip-connection to obtain the final output $F'' \in \mathbb{R}^{C\times H\times W}$:

$$F''_{:i} = \gamma' \sum_{j=1}^{N} (V'_{:j} x'_{ji}) + F'_{:i}, \qquad (4)$$

where γ' is also initialized as 1. Based on F', F'' further gains the semantic correlations between all positions and thus enhancing the semantic representation of the feature.

Finally, we can get the initial location map of the targets by applying a 7×7 convolution with the padding of 3 on F''. The F'' and the initial location map would be refined progressively by the following focus modules (FMs).

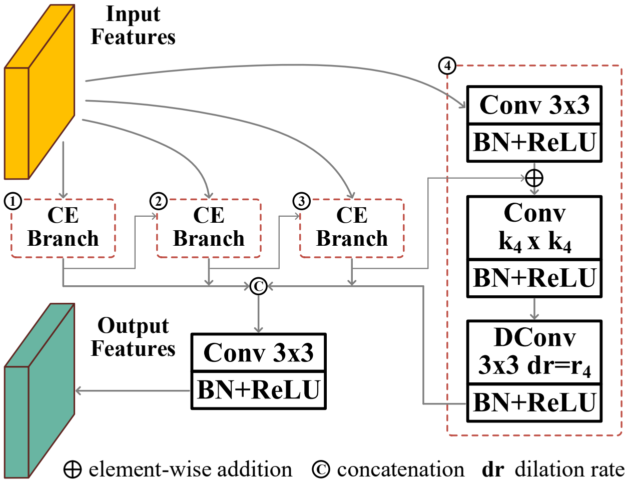


Figure 3. The architecture of our context exploration (CE) block.

3.3. Focus Module

As the camouflaged objects typically share a similar appearance with the background, both false-positive and false-negative predictions would naturally occur in the initial segmentation. The focus module (FM) is designed to first discover and then remove these false predictions. It takes the current-level features derived from the backbone and the higher-level prediction and features as the input, and outputs the refined features and a more accurate prediction.

Distraction discovery. We note that humans could distinguish the distractions well after a careful analysis. Our observation is that humans would do context reasoning, *i.e.*, comparing the patterns (*e.g.*, texture and semantic) of the ambiguous regions with that of the confident regions, to make the final decision. This inspires us to conduct contextual exploration for all the predicted foreground (or background) regions, for the purpose of discovering the false-positive distractions (or the false-negative distractions) that are heterogeneous with the confident foreground (or background) prediction regions. As shown in Figure 2 (c), we first upsample the higher-level prediction and normalize it with a sigmoid layer. Then we use this normalized map and its reverse version to multiply the current-level features F_c, to generate the foreground-attentive features F_{fa} and the background-attentive features F_{ba}, respectively. Finally, we feed these two types of features into two parallel context exploration (CE) blocks to perform contextual reasoning for discovering the false-positive distractions F_{fpd} and the false-negative distractions F_{fnd}, respectively.

As shown in Figure 3, the CE block consists of four context exploration branches and each branch includes a 3×3 convolution for channel reduction, a $k_i \times k_i$ convolution for local feature extraction, and a 3×3 dilated convolution with a dilation rate of r_i for context perceiving. We set

$k_i, i \in \{1,2,3,4\}$ to 1, 3, 5, 7, and set $r_i, i \in \{1,2,3,4\}$ to 1, 2, 4, 8, respectively. Each convolution is followed by a batch normalization (BN) layer and a ReLU nonlinearity operation. The output of the $i^{th}, i \in \{1,2,3\}$ CE branch will be fed into $(i+1)^{th}$ branch to be further processed in a larger receptive field. The outputs of all four branches are then concatenated and fused via a 3×3 convolution. By such design, the CE block gains the capability of perceiving rich contexts over a wide range of scales and thus could be used for context reasoning and distraction discovery.

Distraction Removal. After distraction discovery, we can perform distraction removal in the following way:

$$\begin{aligned} F_{up} &= U(CBR(F_h)), \\ F_r &= BR(F_{up} - \alpha F_{fpd}), \\ F_r' &= BR(F_r + \beta F_{fnd}), \end{aligned} \tag{5}$$

where F_h and F_r' denote the input higher-level features and the output refined features, respectively; CBR is the combination of convolution, batch normalization (BN) and ReLU; U is the bilinear upsampling; and α and β are the learnable scale parameters which are initialized as 1. Here we use the element-wise subtraction operation to suppress the ambiguous backgrounds (*i.e.*, false-positive distractions) and the element-wise addition operation to augment the missing foregrounds (*i.e.*, false-negative distractions).

Finally, a more accurate prediction map can be obtained by applying a convolution layer on the refined feature F_r'. We use the ground truth map to supervise the generated map, to force the F_r' into a purer representation than F_h, *i.e.*, the distraction removed features. This would further guide the CE block to discover the specific form of distractions and make the whole focus module (FM) works on distraction discovery and removal in an implicit way. Note that we do not adopt the specific distraction map to explicitly supervise the F_{fpd} and F_{fnd}, based on the following two considerations: (i) annotating false positives and false negatives are both expensive and subjective, making it difficult to obtain sufficient and representative distractions; and (ii) using a fixed distraction supervision for all focus modules (FMs) is suboptimal as the input higher-level features for each FM is different and the distractions we hope to discover and remove should vary dynamically with the gradually refined input higher-level features.

Discussion. Distraction cues have been explored in many vision tasks such as salient object detection [4, 56], semantic segmentation [21] and visual tracking [73]. Existing works leverage either the false-positive distraction [21, 56, 73] or the false-negative distraction [4] to obtain more accurate results. Unlike the above methods, we explore both two types of distractions and propose a well-designed focus module (FM) to first discover and then remove these distractions. Although the distraction-aware shadow (DS) module in [70] also exploits both two distractions, our proposed focus module (FM) is inherently different from the DS module in the following three aspects. First, DS module extracts features to predict two types of distractions based on the same input features while our focus module (FM) finds false-positive distraction from the foreground-attentive features and discovers false-negative distraction from the background-attentive features. Second, the feature extractor in the DS module contains two 3×3 convolutions while our context exploration (CE) block consists of four branches, which could capture multi-scale contexts for better distraction discovery. Third, the supervision for the DS module is acquired based on the differences between the predictions from existing shadow detection models (*i.e.*, [20, 72, 25]) and the ground truths. Such an explicit supervision strategy would be constrained by the specific methods and thus may have limited generality. By contrast, we design an implicit distraction mining strategy via imposing ground truth supervision on the distraction-removed features to force each CE block exploring the specific form of distractions. To the best of our knowledge, we are the first to mine distractions for camouflaged object segmentation and we believe that the proposed strategy of distraction mining could provide insights to other vision tasks.

3.4. Loss Function

There are four output predictions in the PFNet, *i.e.*, one from the positioning module (PM) and three from the focus modules (FMs). For the PM, we impose binary cross-entropy (BCE) loss ℓ_{bce} and IoU loss ℓ_{iou} [44] on its output, *i.e.*, $\mathcal{L}_{pm} = \ell_{bce} + \ell_{iou}$, to guide the PM to explore the initial location of the target object. For the FM, we hope it could focus more on the distraction region. Such region is typically located at the object's boundaries, elongated areas or holes. Thus we combine the weighted BCE loss ℓ_{wbce} [53] and weighted IoU loss ℓ_{wiou} [53], *i.e.*, $\mathcal{L}_{fm} = \ell_{wbce} + \ell_{wiou}$, to force the FM pay more attention to the possible distraction region. Finally, the overall loss function is:

$$\mathcal{L}_{overall} = \mathcal{L}_{pm} + \sum_{i=2}^{4} 2^{(4-i)} \mathcal{L}_{fm}^{i}, \tag{6}$$

where $\mathcal{L}_{fm}^{i}$ denotes the loss for the prediction of the focus module at i-th level of the PFNet.

4. Experiments

4.1. Experimental Setup

Datasets. We evaluate our method on three benchmark datasets: CHAMELEON [46], CAMO [26], and COD10K [12]. CHAMELEON [46] has 76 images collected from the Internet via the Google search engine using "camouflaged animal" as a keyword and corresponding manually

Methods	Pub.'Year	CHAMELEON (76 images)				CAMO-Test (250 images)				COD10K-Test (2,026 images)			
		$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
FPN$^\circ$ [30]	CVPR'17	0.794	0.835	0.590	0.075	0.684	0.791	0.483	0.131	0.697	0.711	0.411	0.075
PSPNet$^\bullet$ [66]	CVPR'17	0.773	0.814	0.555	0.085	0.663	0.778	0.455	0.139	0.678	0.688	0.377	0.080
Mask RCNN* [17]	ICCV'17	0.643	0.780	0.518	0.099	0.574	0.716	0.430	0.151	0.613	0.750	0.402	0.080
UNet++§ [71]	DLMIA'17	0.695	0.808	0.501	0.094	0.599	0.740	0.392	0.149	0.623	0.718	0.350	0.086
DSC$^\triangle$ [20]	CVPR'18	0.850	0.888	0.714	0.050	0.736	0.830	0.592	0.105	0.758	0.788	0.542	0.052
PiCANet† [33]	CVPR'18	0.769	0.836	0.536	0.085	0.609	0.753	0.356	0.156	0.649	0.678	0.322	0.090
BDRAR$^\triangle$ [72]	ECCV'18	0.779	0.881	0.663	0.064	0.759	0.825	0.664	0.093	0.753	0.836	0.591	0.051
HTC* [2]	CVPR'19	0.517	0.490	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.521	0.221	0.088
MSRCNN* [22]	CVPR'19	0.637	0.688	0.443	0.091	0.617	0.670	0.454	0.133	0.641	0.708	0.419	0.073
BASNet† [44]	CVPR'19	0.687	0.742	0.474	0.118	0.618	0.719	0.413	0.159	0.634	0.676	0.365	0.105
CPD† [55]	CVPR'19	0.853	0.878	0.706	0.052	0.726	0.802	0.550	0.115	0.747	0.763	0.508	0.059
PFANet† [69]	CVPR'19	0.679	0.732	0.378	0.144	0.659	0.735	0.391	0.172	0.636	0.619	0.286	0.128
EGNet† [67]	ICCV'19	0.848	0.879	0.702	0.050	0.732	0.827	0.583	0.104	0.737	0.777	0.509	0.056
F3Net† [53]	AAAI'20	0.854	0.899	0.749	0.045	0.779	0.840	0.666	0.091	0.786	0.832	0.617	0.046
GCPANet† [5]	AAAI'20	0.876	0.891	0.748	0.041	0.778	0.842	0.646	0.092	0.791	0.799	0.592	0.045
PraNet§ [13]	MICCAI'20	0.860	0.898	0.763	0.044	0.769	0.833	0.663	0.094	0.789	0.839	0.629	0.045
MINet-R† [40]	CVPR'20	0.844	0.919	0.746	0.040	0.749	0.835	0.635	0.090	0.759	0.832	0.580	0.045
SINet* [12]	CVPR'20	0.869	0.899	0.740	0.044	0.751	0.834	0.606	0.100	0.771	0.797	0.551	0.051
PFNet*	Ours	**0.882**	**0.942**	**0.810**	**0.033**	**0.782**	**0.852**	**0.695**	**0.085**	**0.800**	**0.868**	**0.660**	**0.040**

Table 1. Comparison of our proposed method and other 18 state-of-the-art methods in the relevant fields on three benchmark datasets in terms of the structure-measure S_α (larger is better), the adaptive E-measure E_ϕ^{ad} (larger is better), the weighted F-measure F_β^w (larger is better), and the mean absolute error M (smaller is better). All the prediction maps are evaluated with the same code. The best results are marked in **bold**. ○: object detection method. ●: semantic segmentation method. ⋆: instance segmentation methods. △: shadow detection methods. §: medical image segmentation methods. †: SOD methods. *: COS methods. Our method outperforms other counterparts with a large margin under all four standard evaluation metrics on all three benchmark datasets.

annotated object-level ground-truths. CAMO [26] contains 1,250 camouflaged images covering different categories, which are divided into 1,000 training images and 250 testing images. COD10K [12] is currently the largest benchmark dataset, which includes 5,066 camouflaged images (3,040 for training and 2,026 for testing) downloaded from multiple photography websites, covering 5 super-classes and 69 sub-classes. We follow previous work [12] to use the training set of CAMO [26] and COD10K [12] as the training set (4,040 images) and others as testing sets.

Evaluation Metrics. We use four widely used and standard metrics to evaluate our method: structure-measure (S_α) [10], adaptive E-measure (E_ϕ^{ad}) [11], weighted F-measure (F_β^w) [35], and mean absolute error (M). Structure-measure (S_α) focuses on evaluating the structural information of the prediction maps, which is defined as: $S_\alpha = \alpha S_o + (1-\alpha)S_r$, where S_o and S_r denote the object-aware and region-aware structural similarity, respectively; and α is set to be 0.5 as suggested in [10]. E-measure (E_ϕ) simultaneously evaluates the pixel-level matching and image-level statistics, which is shown to be related to human visual perception [11]. Thus, we include this metric to assess the overall and localized accuracy of the camouflaged object segmentation results. F-measure (F_β) is a comprehensive measure on both the precision and recall of the prediction map. Recent studies [10, 11] have suggested that the weighted F-measure (F_β^w) [35] can provide more reliable evaluation results than the traditional F_β. Thus, we also consider this metric in the comparison. The mean absolute error (M) metric is widely used in foreground-background segmentation tasks, which calculates the element-wise difference between the prediction map and the ground truth mask.

Implementation Details. We implement our model with the PyTorch toolbox [41]. An eight-core PC with an Intel Core i7-9700K 3.6 GHz CPU (with 64GB RAM) and an NVIDIA GeForce RTX 2080Ti GPU (with 11GB memory) is used for both training and testing. For training, input images are resized to a resolution of 416×416 and are augmented by randomly horizontal flipping and color jittering. The parameters of the encoder network are initialized with the ResNet-50 model [18] pre-trained on ImageNet while the remaining layers of our PFNet are initialized randomly. We use the stochastic gradient descent (SGD) optimizer with the momentum of 0.9 and the weight decay of 5×10^{-4} for loss optimization. We set the batch size to 16 and adjust the learning rate by the poly strategy [34] with the basic learning rate of 0.001 and the power of 0.9. It takes only about 76 minutes for the network to converge for 45 epochs. For testing, the image is first resized to 416×416 for network inference and then the output map is resized back to the original size of the input image. Both the resizing processes use bilinear interpolation. We do not use any post-processing such as the fully connected conditional random field (CRF) [24] to further enhance the final output. The inference for a 416×416 image takes only 0.014 seconds (about 72 FPS).

Compared Methods. To demonstrate the effectiveness of our PFNet, we compare it against 18 state-of-the-art base-

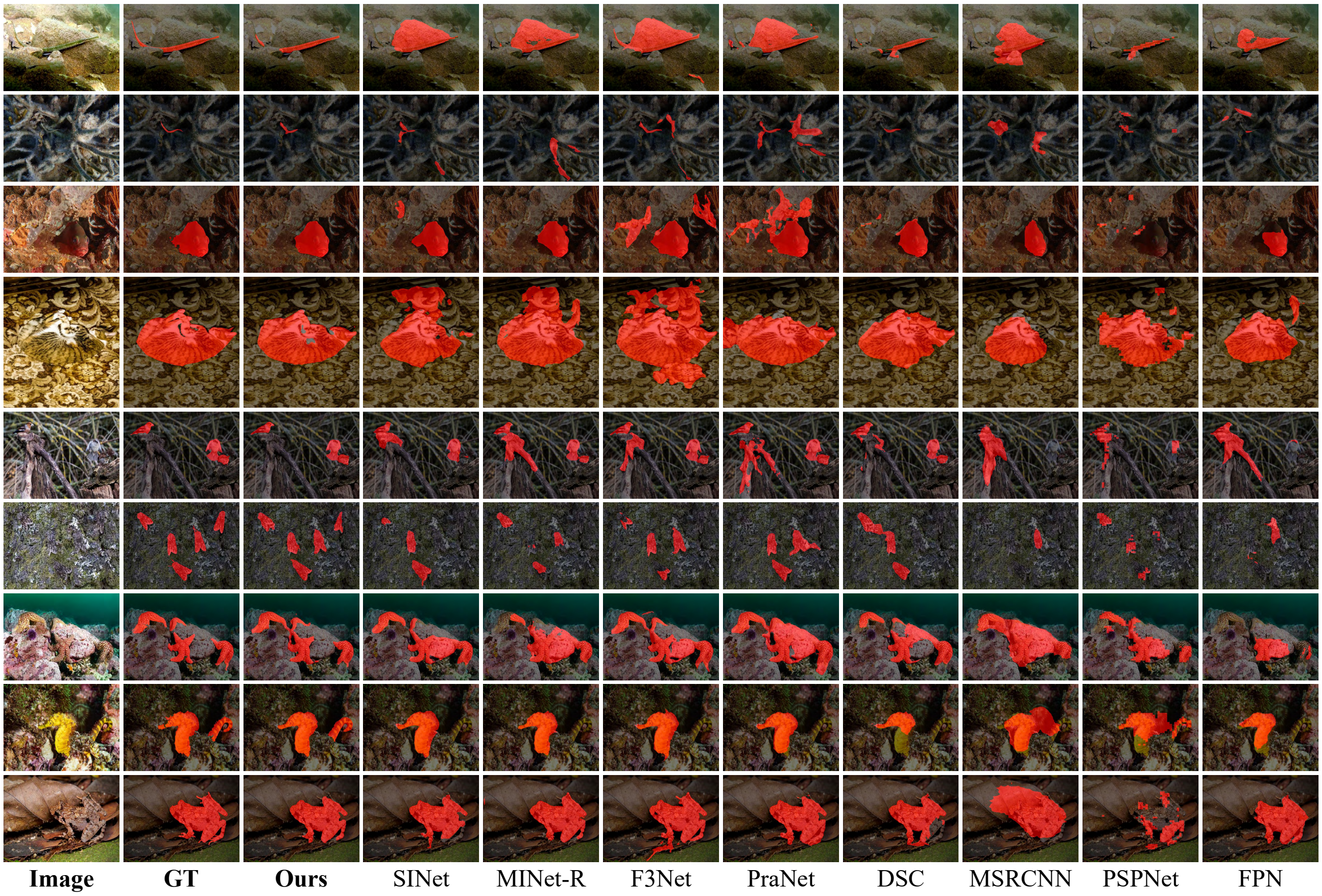


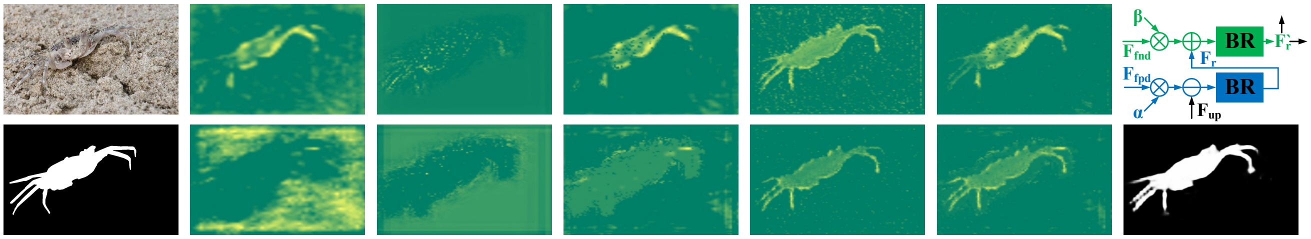
Figure 4. Visual comparison of the proposed model with state-of-the-art methods. Obviously, our approach is capable of segmenting various camouflaged objects concealed in different environments more accurately.

lines: object detection method FPN [30]; semantic segmentation method PSPNet [66]; instance segmentation methods Mask RCNN [17], HTC [2], and MSRCNN [22]; shadow detection methods DSC [20] and BDRAR [72]; medical image segmentation methods UNet++ [71] and PraNet [13]; salient object detection methods PiCANet [33], BASNet [44], CPD [55], PFANet [69], EGNet [67], F3Net [53], GCPANet [5], and MINet-R [40]; and camouflaged object segmentation method SINet [12]. For fair comparison, all the prediction maps of the above methods are either provided by the public website or produced by running the models retrained with open source codes. Besides, all the prediction maps are evaluated with the same code.

4.2. Comparison with the State-of-the-arts

Table 1 reports the quantitative results of PFNet against other 18 state-of-the-art methods on three benchmark datasets. We can see that our method outperforms all the other methods with a large margin on all four standard metrics. For example, compared with the state-of-the-art COS method SINet [12], our method improves F_β^w by 7.0%, 8.9%, and 10.9% on the CHAMELEON [46], CAMO [26], and COD10K [12] dataset, respectively. Note that our method is also faster than SINet, *i.e.*, 72 versus 51 FPS.

Besides, Figure 4 shows the qualitative comparison of our method with the others. It can be seen that our method is capable of accurately segmenting small camouflaged objects (*e.g.*, the first two rows), large camouflaged objects (*e.g.*, 3-*rd* and 4-*th* rows), and multiple camouflaged objects (*e.g.*, 5-*th* and 6-*th* rows). This is mainly because that the positioning module (PM) can provide the initial location of the camouflaged objects with different scales for the following distraction mining, via exploring long-range semantic dependencies. While the state-of-the-arts are typically confused by the background which shares similar appearance with the camouflaged objects (*e.g.*, 7-*th* row) or the foreground region that cluttered in the background (*e.g.*, 8-*th* row), our method can successfully infer the true camouflaged region. This is mainly contributed by the proposed distraction mining strategy which could help suppress the false-positive distraction region and augment the false-negative distraction region. Furthermore, benefited by


β
$\mathbf{F}_{\mathrm{fnd}}$
BR
$\mathbf{F}_{\mathrm{r}}$
$\mathbf{F}_{\mathrm{r}}^{`}$
$\mathbf{F}_{\mathrm{fpd}}$
BR
α
$\mathbf{F}_{\mathrm{up}}$

(a) Image & GT (b) F_{up} (c) $\mathrm{F}_{\mathrm{fpd}}$ (d) F_{r} (e) $\mathrm{F}_{\mathrm{fnd}}$ (f) $\mathrm{F}_{\mathrm{r}}^{`}$ (g) Prediction

Figure 5. Visualizing feature maps in the last FM. Best viewed in color and zoomed-in.

	Networks	COD10K-Test (2,026 images)			
		$S_\alpha\uparrow$	$E_\phi^{ad}\uparrow$	$F_\beta^w\uparrow$	$M\downarrow$
(a)	B	0.779	0.803	0.591	0.051
(b)	B + CA	0.788	0.819	0.618	0.046
(c)	B + SA	0.791	0.826	0.624	0.046
(d)	B + PM	0.792	0.835	0.631	0.045
(e)	B + FPD	0.790	0.844	0.632	0.043
(f)	B + FND	0.790	0.837	0.628	0.043
(g)	B + FM *w/o* A	0.796	0.843	0.639	0.042
(h)	B + FM	0.797	0.860	0.649	0.041
(i)	B + PM + FPD	0.796	0.854	0.645	0.042
(j)	B + PM + FND	0.796	0.847	0.644	0.043
(k)	B + PM + FM *w/o* A	0.796	0.851	0.647	0.042
(l)	PFNet	**0.800**	**0.868**	**0.660**	**0.040**

Table 2. Ablation analyses. "B" denotes our network with the channel attention block ("CA") and spatial attention block ("SA") removed from positioning module ("PM") and the false-positive distraction stream ("FPD") and false-negative distraction stream ("FND") in the focus module ("FM") replaced by a simple skip-connection. "*w/o* A" denotes that the higher-level prediction is not used as the attention map to guide the current-level features in the focus module. As can be observed, each proposed component plays an important role and contributes to the performance.

the multi-scale context exploration in the distraction mining process, our method can capture detailed distraction information and thus has the ability to finely segment the camouflaged objects with complex structures (*e.g.*, the last row).

4.3. Ablation Study

We conduct ablation studies to validate the effectiveness of two key components tailored for accurate camouflaged object segmentation, *i.e.*, positioning module (PM) and focus module (FM), and report the results in Table 2.

The effectiveness of PM. In Table 2, we can see that adding the channel attention block (b) or the spatial attention block (c) on the base model (a) can boost the segmentation performance to some extent and the combination of the two (d) can achieve better results. This confirms that the PM can benefit the accurate camouflaged object segmentation.

The effectiveness of FM. Based on (a), introducing our proposed false-positive distraction mining (e) or false-negative distraction mining (f) would greatly improve the segmentation results. Considering both two types of distractions, *i.e.*, (h), we obtain better results. For example, adding the focus module gains 5.7% and 5.8% performance improvement in terms of E_ϕ^{ad} and F_β^w, respectively. This shows that the FM enables our approach to possess the strong capability of accurately segmenting the camouflaged objects. When removing the guidance from the higher-level prediction, *i.e.*, (g), the performance would decline to some extent. This is because that indiscriminately mining distractions from the input features increases the difficulty of the distraction discovery and thus hinder the effective distraction removal. This validates the rationality of our design to learn distractions from the attentive input features. From the results of (i-l), we can see that the above conclusions still keep true when adding the partial/full focus module on (d). In addition, we visualize the feature maps in the last FM in Figure 5. By mining the false-positive distractions (c), the false-positive predictions in (b) can be greatly suppressed (d). Through mining false-negative distractions (e), the purer representation of the target object can be obtained (f). This clearly demonstrates the effectiveness of the proposed distraction mining strategy which is designed to discover and remove distractions.

5. Conclusion

In this paper, we strive to embrace challenges towards accurate camouflaged object segmentation. We develop a novel distraction mining strategy for distraction discovery and removal. By adopting the distraction mining strategy in our bio-inspired framework, *i.e.*, positioning and focus network (PFNet), we show that our approach achieves state-of-the-art performance on three benchmarks. In the future, we plan to explore the potential of our method for other applications such as polyp segmentation and COVID-19 lung infection segmentation and further enhance its capability for segmenting camouflaged objects in videos.

Acknowledgements. This work was supported in part by the National Natural Science Foundation of China under Grant 91748104, Grant 61972067, Grant U1908214, in part by the Innovation Technology Funding of Dalian (Project No. 2020JJ26GX036).

References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, 2009.

[2] 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 *CVPR*, 2019.

[3] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2017.

[4] Shuhan Chen, Xiuli Tan, Ben Wang, and Xuelong Hu. Reverse attention for salient object detection. In *ECCV*, 2018.

[5] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, 2020.

[6] Ming-Ming Cheng, Niloy J Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. Global contrast based salient region detection. *IEEE TPAMI*, 2014.

[7] Hugh Bamford Cott. Adaptive coloration in animals. *Methuen & Co. Ltd*, 1940.

[8] Henghui Ding, Xudong Jiang, Bing Shuai, Ai Qun Liu, and Gang Wang. Context contrasted feature and gated multiscale aggregation for scene segmentation. In *CVPR*, 2018.

[9] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, 2018.

[10] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[11] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 2021.

[12] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020.

[13] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, 2020.

[14] 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 TMI*, 2020.

[15] Joanna R Hall, Innes C Cuthill, Roland J Baddeley, Adam J Shohet, and Nicholas E Scott-Samuel. Camouflage, detection and identification of moving targets. *Proceedings of The Royal Society B: Biological Sciences*, 2013.

[16] Xiaofeng Han, Chuong Nguyen, Shaodi You, and Jianfeng Lu. Single image water hazard detection using fcn with reflection attention units. In *ECCV*, 2018.

[17] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017.

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016.

[19] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip Torr. Deeply supervised salient object detection with short connections. *IEEE TPAMI*, 2019.

[20] Xiaowei Hu, Lei Zhu, Chi-Wing Fu, Jing Qin, and Pheng-Ann Heng. Direction-aware spatial context features for shadow detection. In *CVPR*, 2018.

[21] Qin Huang, Chunyang Xia, Chi-Hao Wu, Siyang Li, Ye Wang, Yuhang Song, and C.-C. Jay Kuo. Semantic segmentation with reverse attention. In *BMVC*, 2017.

[22] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019.

[23] Alexander Kirillov, Kaiming He, Ross Girshick, Carsten Rother, and Piotr Dollar. Panoptic segmentation. In *CVPR*, 2019.

[24] Philipp Krähenbühl and Vladlen Koltun. Efficient inference in fully connected crfs with gaussian edge potentials. In *NeurIPS*, 2011.

[25] Hieu Le, Tomas F. Yago Vicente, Vu Nguyen, Minh Hoai, and Dimitris Samaras. A+d net: Training a shadow detector with adversarial shadow attenuation. In *ECCV*, 2018.

[26] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[27] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *CVPR*, 2016.

[28] Jianqin Yin Yanbin Han Wendi Hou Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *Procedia Engineering*, 2011.

[29] Li-Jia Li, Richard Socher, and Li Fei-Fei. Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In *CVPR*, 2009.

[30] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[31] Li Liu, Wanli Ouyang, Xiaogang Wang, Paul Fieguth, Jie Chen, Xinwang Liu, and Matti Pietikinen. Deep learning for generic object detection: A survey. *IJCV*, 2018.

[32] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[33] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[34] Wei Liu, Andrew Rabinovich, and Alexander C. Berg. Parsenet: Looking wider to see better. *arXiv:1506.04579*, 2015.

[35] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014.

[36] Haiyang Mei, Bo Dong, Wen Dong, Pieter Peers, Xin Yang, Qiang Zhang, and Xiaopeng Wei. Depth-aware mirror segmentation. In *CVPR*, 2021.

[37] Haiyang Mei, Yuanyuan Liu, Ziqi Wei, Li Zhu, Yuxin Wang, Dongsheng Zhou, Qiang Zhang, and Xin Yang. Exploring dense context for salient object detection. *IEEE TCSVT*, 2021.

[38] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. Don't hit me! glass detection in real-world scenes. In *CVPR*, 2020.

[39] 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*, 2011.

[40] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *CVPR*, 2020.

[41] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

[42] Chao Peng, Xiangyu Zhang, Gang Yu, Guiming Luo, and Jian Sun. Large kernel matters — improve semantic segmentation by global convolutional network. In *CVPR*, 2017.

[43] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[44] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[45] P. Sengottuvelan, A. Wahi, and A. Shanmugam. Performance of decamouflaging through exploratory image analysis. In *ETET*, 2008.

[46] P Skurowski, H Abdulameer, J Błaszczyk, T Depta, A Kornacki, and P Kozieł. Animal camouflage analysis: Chameleon database. *Unpublished Manuscript*, 2018.

[47] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B*, 2009.

[48] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *ICCV*, 2019.

[49] Gerald Handerson Thayer and Abbott Handerson Thayer. Concealing-coloration in the animal kingdom : an exposition of the laws of disguise through color and pattern being a summary of abbott h. thayer's discoveries. *New York the Macmillan Co*, 1909.

[50] Tom Troscianko, Christopher P Benton, P. George Lovell, David J Tolhurst, and Zygmunt Pizlo. Camouflage and visual perception. *Philosophical Transactions of the Royal Society B*, 2009.

[51] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017.

[52] Wenguan Wang, Jianbing Shen, Ming-Ming Cheng, and Ling Shao. An iterative and cooperative top-down and bottom-up inference network for salient object detection. In *CVPR*, 2019.

[53] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020.

[54] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. Cbam: Convolutional block attention module. In *ECCV*, 2018.

[55] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[56] Huaxin Xiao, Jiashi Feng, Yunchao Wei, Maojun Zhang, and Shuicheng Yan. Deep salient object detection with dense connections and distraction diagnosis. *IEEE TMM*, 2018.

[57] Enze Xie, Wenjia Wang, Wenhai Wang, Mingyu Ding, Chunhua Shen, and Ping Luo. Segmenting transparent objects in the wild. In *ECCV*, 2020.

[58] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V. Nguyen. Mirrornet: Bio-inspired adversarial attack for camouflaged object segmentation. *arXiv:2007.12881*, 2020.

[59] Xin Yang, Haiyang Mei, Ke Xu, Xiaopeng Wei, Baocai Yin, and Rynson W.H. Lau. Where is my mirror? In *ICCV*, 2019.

[60] Xin Yang, Haiyang Mei, Jiqing Zhang, Ke Xu, Baocai Yin, Qiang Zhang, and Xiaopeng Wei. Drfn: Deep recurrent fusion network for single-image super-resolution with large factors. *IEEE TMM*, 2019.

[61] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Computing Surveys*, 2006.

[62] Jiqing Zhang, Chengjiang Long, Yuxin Wang, Xin Yang, Haiyang Mei, and Baocai Yin. Multi-context and enhanced reconstruction network for single image super resolution. In *ICME*, 2020.

[63] Pingping Zhang, Dong Wang, Huchuan Lu, Hongyu Wang, and Xiang Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *ICCV*, 2017.

[64] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *CVPR*, 2018.

[65] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[66] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[67] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *ICCV*, 2019.

[68] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.

[69] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.

[70] Quanlong Zheng, Xiaotian Qiao, Ying Cao, and Rynson WH Lau. Distraction-aware shadow detection. In *CVPR*, 2019.

[71] Zongwei Zhou, Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. *DLMIA*, 2018.

[72] Lei Zhu, Zijun Deng, Xiaowei Hu, Chi-Wing Fu, Xuemiao Xu, Jing Qin, and Pheng-Ann Heng. Bidirectional feature pyramid network with recurrent attention residual modules for shadow detection. In *ECCV*, 2018.

[73] Zheng Zhu, Qiang Wang, Bo Li, Wei Wu, Junjie Yan, and Weiming Hu. Distractor-aware siamese networks for visual object tracking. In *ECCV*, 2018.