I Can Find You! Boundary-Guided Separated Attention Network for Camouflaged Object Detection

Hongwei Zhu[1,2*], Peng Li[1,2*], Haoran Xie[3], Xuefeng Yan[1,2], Dong Liang[1,2], Dapeng Chen[4], Mingqiang Wei[1,2†], Jing Qin[5]

[1]Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2]MIIT Key Laboratory of Pattern Analysis and Machine Intelligence, Nanjing, China
[3]Lingnan University, Hong Kong SAR, China
[4]AI Application Research Center, Huawei Technologies, Shenzhen, China
[5]Hong Kong Polytechnic University, Hong Kong SAR, China

Abstract

Can you find me? By simulating how humans to discover the so-called 'perfectly'-camouflaged object, we present a novel boundary-guided separated attention network (call BSA-Net). Beyond the existing camouflaged object detection (COD) wisdom, BSA-Net utilizes two-stream separated attention modules to highlight the separator (or say the camouflaged object's boundary) between an image's background and foreground: the reverse attention stream helps erase the camouflaged object's interior to focus on the background, while the normal attention stream recovers the interior and thus pay more attention to the foreground; and both streams are followed by a boundary guider module and combined to strengthen the understanding of the boundary. The core design of such separated attention is motivated by the COD procedure of humans: find the subtle difference between the foreground and background to delineate the boundary of a camouflaged object, then the boundary can help further enhance the COD accuracy. We validate on three benchmark datasets that our BSA-Net is very beneficial to detect camouflaged objects with the blurred boundaries and similar colors/patterns with their backgrounds. Extensive results exhibit very clear COD improvements on our BSA-Net over sixteen SOTAs.

Introduction

Camouflaged object detection (COD) from single images aims to find an object that has colors or patterns very similar to but with intrinsic different attributes from its background. A successful effort of COD will facilitate many practically meaningful applications of polyp segmentation (Fan et al. 2020b), lung infection segmentation (Wu et al. 2021), recreational art (Chu et al. 2010), and even the rescue mission in extreme weather and anti-military camouflage.

Traditionally, one may adapt salient object detection (SOD) techniques (Zhao et al. 2019; Qin et al. 2019; Wei et al. 2020; Wu, Su, and Huang 2019b; Feng, Lu, and Ding 2019) to the task of COD; or one may develop various hand-crafted features to deal with COD. Unfortunately, the camouflaged objects often conceal themselves in the background as much as possible instead of highlighting themselves as salient objects do. Those SOD techniques or even the elaborate hand-crafted COD features are not sensitive enough to capture subtle differences between any camouflaged object and its background, leading to poor COD results.

Recently, many promising COD methods based on deep learning have been proposed. For example, inspired by the hunting process of a hunter, SINet (Fan et al. 2020a) pioneers to collect a large-scale COD dataset called COD10K, and utilizes a search module and an identification module to locate and identify camouflaged objects. MirrorNet (Yan et al. 2021) assumes that changing the perspective of the same scene can enhance the difference between an object and its background. Rank-Net (Lv et al. 2021) proposes a model to locate, segment and rank camouflaged objects simultaneously, in which the rank module can rank the capability of COD. However, these efforts seldom consider to simulate how humans to detect camouflaged objects, which cannot well vanquish the camouflaged objects with their ambiguous boundaries to the backgrounds.

You may recall how to detect potential camouflaged objects in an image: you first search throughout the image to find the possible region that contains the camouflaged object, then you focus on both the foreground and background in order to find the inconspicuous difference between them. When you gradually discover the difference between them, the boundary of the camouflaged object is highlighted. Finally, you utilize the boundary as an enhancement guidance to improve the COD capability. Inspired by this observation, we propose an effective boundary-guided separated attention network for the COD task (termed as BSA-Net).

Beyond the existing COD wisdom, BSA-Net is a coarse-to-fine learning model, which exploits three main modules to help improve the COD results: the Residual Multi-scale Feature Extractor (RMFE) to capture rich context information, the Separated Attention (SEA) module to handle the sensitivity-invariance dilemma, and the Boundary Guider (BG) to accurately highlight the boundary of a camouflaged object. Specifically, SEA contains two streams: the normal attention stream and the reverse attention stream focus on the foreground and background of the input image respectively, and then integrate the two streams for collaboration. Since the boundary of any camouflaged object (i.e., the separator of foreground and background of an image) is difficult

[*]Joint first authors

[†]Corresponding author: mingqiang.wei@gmail.com

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

to detect, we utilize the BG module to enhance the boundary detection capability after each stream.

Our contributions are three-fold:

- By simulating how humans to detect camouflaged objects, we propose a novel COD network, which leverages the proposed separated attention modules to improve the performance of cutting-edge COD models.
- We design a refinement paradigm in which a simple yet effective boundary guider is proposed to embed the boundary information into the coarse feature map, forming our boundary-guided separated attention network.
- BSA-Net is validated on three popular COD datasets and achieves the most outstanding performance among sixteen SOD and COD methods, especially when the camouflaged objects have very blurred boundaries and similar colors/patterns with their backgrounds.

Related Work

Salient Object Detection (SOD)

SOD aims to localize the most "eye-catching" object(s) in an image, and segment it at the pixel level to generate a binary map. Since saliency and camouflage are slightly opposite to each other, we can regard salient objects as negative samples in the COD tasks. The SOD methods can be potentially exploited to train on the camouflaged object datasets for COD, which will be compared with our proposed BSA-Net.

Existing SOD methods are based on either feature fusion or boundary information. The feature-fusion strategy is dedicated to integrating multi-scale features to enhance the performance of SOD models. For example, Wei et al. (Wei, Wang, and Huang 2020) propose a selective fusion scheme to suppress redundant features and adopt a multi-layer feedforward mechanism to supplement the output features of the previous layers and eliminate the differences between them. Zhu et al. (Zhu et al. 2019) propose an attentional dense atrous spatial pyramid pooling (AD-ASPP) module to expand the receptive field which refines the feature maps of each layer. Pang et al. (Pang et al. 2020) integrate feature maps of adjacent levels and fuse multi-scale information. The boundary-aware strategy utilizes boundary information to optimize the saliency detection influence. Qin et al. (Qin et al. 2019) leverage a supervised encoder-decoder structure and a residual refinement module for coarse-to-fine detection. Wu et al. (Wu, Su, and Huang 2019b) employ the Cross Refinement Unit (CRU) to refine the saliency map and the boundary map at the same time. Zhao et al. (Zhao et al. 2019) obtain salient boundary features and couple salient objects at various resolutions together.

Camouflaged Object Detection (COD)

Due to the fact that the patterns and colors of camouflaged objects are similar to the background, it is difficult to separate them from the background. Traditional COD methods mostly utilize hand-crafted features, such as color, convex intensity, boundary, texture, and brightness (Bhajantri and Nagabhushan 2006; Xue et al. 2015; Huerta et al. 2007; Tankus and Yeshurun 2001; Kavitha, Rao, and Govardhan 2011) to distinguish camouflaged objects from their backgrounds. However, these methods are only suitable for simple uneven backgrounds. Since the foreground and background are highly similar, these methods are often deceived by camouflaged objects, leading to poor detection results.

Recent works resort to the huge capacity of deep neural networks to recognize the more complex properties of camouflage objects, and have acquired the outstanding performance (Sun et al. 2021; Liu, Zhang, and Barnes 2021; Zhang et al. 2021; Li et al. 2021; Mao et al. 2021). Le et al. (Le et al. 2019) propose an end-to-end network that employs a strategy of first classifying the camouflaged image and then segmenting the camouflaged objects. Fan et al. (Fan et al. 2020a) design a search module and an identification module based on the motivation of searching for prey first and then identifying prey during hunting. Ren et al. (Ren et al. 2021) propose a texture-aware module to amplify the subtle texture difference between the camouflaged object and the background. Yan et al. (Yan et al. 2021) take both the original and flipped images as input, and deliver them to the dual-stream mirror network to achieve the intention of changing the viewpoint in the same scene to identify the camouflaged object. Mei et al. (Mei et al. 2021) propose PFNet which is able to position the potential camouflaged objects and then wipe out the false positive and false negative areas which can distract the segmentation results. Zhai et al. (Zhai et al. 2021) propose a Mutual Graph Learning model for COD, which decouples the image into two feature maps for locating the object and capturing the boundary. Different from the existing methods, inspired by the way of humans to recognize camouflaged objects, our method pays more attention to the synergy of the foreground and background information to highlight the boundaries of objects, thereby improving the accuracy of the model.

Methodology

By simulating how humans to find camouflaged objects, we raise an intriguing question: Does the synergy of leveraging the background by erasing the foreground, and focusing on the foreground by ignoring the background enhance the capability of cutting-edge COD networks? To answer it, we present the boundary-guided separated attention network for COD. BSA-Net captures multi-scale feature information to locate where possibly the object is, then employs separated attention to excavate deep information in the foreground and background of the image, and coalesce the additional boundary information in order to strengthen the detection performance at the junction of the foreground and the background. Finally, it uses shuffle attention (Yang 2021) and feature fusion mechanisms to refine features.

Network Architecture

The architecture of BSA-Net is shown in Figure 1. For the input $I \in \mathbb{R}^{W\times H\times 3}$, where W and H denote the width and height of an image, we employ Res2Net as our backbone to extract the multi-level features $F_i, i \in \{1, 2, 3, 4, 5\}$. Generally, BSA-Net is a coarse-to-fine model. First, we input F_2, F_3, F_4, F_5 into the residual multi-scale feature extractors to capture features of different receptive fields, then

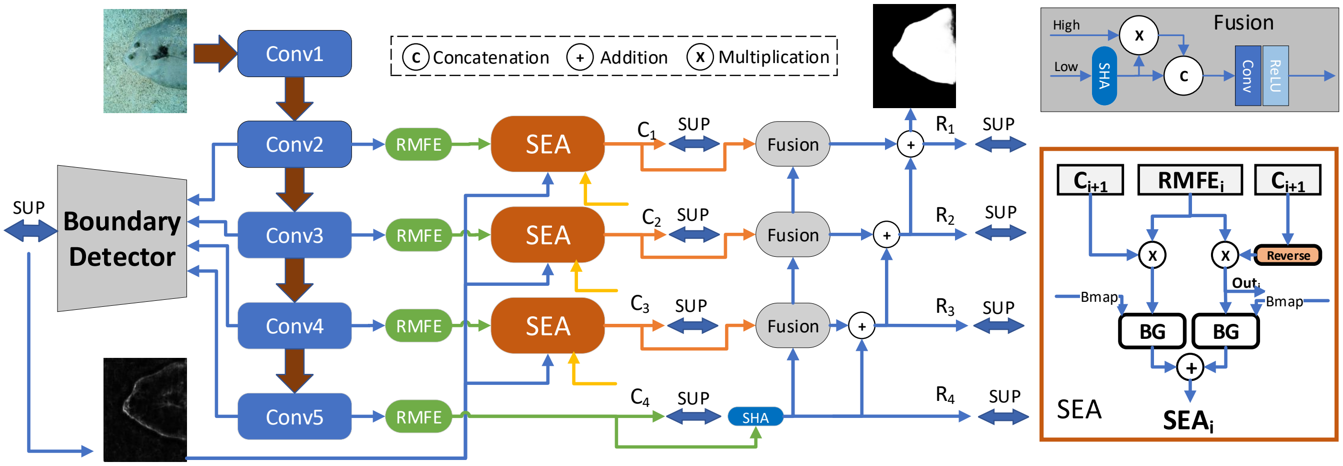

Figure 1: Overview of BSA-Net. BSA-Net simulates the procedure of how humans to detect camouflaged objects. We adopt Res2Net as the backbone encoder. After capturing rich context information by the Residual Multi-scale Feature Extractor (RMFE), we design the Separated Attention (SEA) module to distinguish the subtle difference of foreground and background. The Boundary Guider (BG) module is included in the SEA module to strengthen the model's ability to understand the boundary. Finally, we employ the Shuffle Attention (SHA) block and a feature fusion module to refine our COD result.

we utilize separated attention blocks containing the normal attention stream and reverse attention stream to focus on both the foreground and background. The coarse maps $C_i, i \in \{1, 2, 3, 4\}$ are obtained from the reverse attention stream of separated attention blocks, which are supervised by the ground truth. Furthermore, we design an effective boundary detection network to obtain the boundary map BM, which is exploited in the boundary guider module in separated attention blocks. After that, the shuffle attention module is utilized to make the model pay attention to the informative channels. Finally, we obtain 4 refined maps of the first round predictions marked as $R_i, i \in \{1, 2, 3, 4\}$. We choose R_1 as the final output in the inference stage.

Residual Multi-scale Feature Extractor

Since the ResNet-based backbone networks use convolution operations serially, they cannot extract abundant context information. Moreover, only using 3×3 convolutions is difficult to obtain multiple-scale features in one stage, which is adverse to image understanding and segmentation. Inspired by the Inception module and Res2Net (Gao et al. 2021) block, we develop a Residual Multi-scale Feature Extractor (RMFE) to solve these problems.

RMFE adopts the 3×3 convolution in parallel while employing residual blocks to enlarge the receptive fields successively. To be more specific, for an input feature F_i, we utilize 4 branches to capture different characterizations. Each branch is equipped with a 1×1 convolution to reduce the number of channels, a 1×3 and 3×1 asymmetric convolution for reducing the computational load. The output of each branch is added to the input of the next branch. The general formulation of the operation is defined as

$$Bout_k^i = \begin{cases} Conv_r(F_i) & k = 1 \\ Conv_r(F_i \oplus Bout_{k-1}) & k = 2, 3, 4 \end{cases} \quad (1)$$

where F_i denotes the ith feature map produced by the backbone network, k is the branch number, $Bout_k^i$ denotes the output of the kth branch, $\oplus$ is element-wise addition, $Conv_r()$ denotes the stacked convolutional layer mentioned above. After that, we concatenate the outputs of 4 branches followed by a 1×1 convolution to adjust the channel to 64 and add it to the input feature F_i. Finally, we obtain the output feature $RMFE_i, i \in \{2, 3, 4, 5\}$ embedded with multi-scale information which is computed as

$$RMFE_i = Conv(F_i) \oplus Conv(Cat_{k=1}^4(Bout_k^i)), \quad (2)$$

where $Conv()$ denotes 1×1 convolution, $Cat_{k=1}^4$ denotes the concatenation of all 4 branches. The overall structure of RMFE is shown in Figure 2.

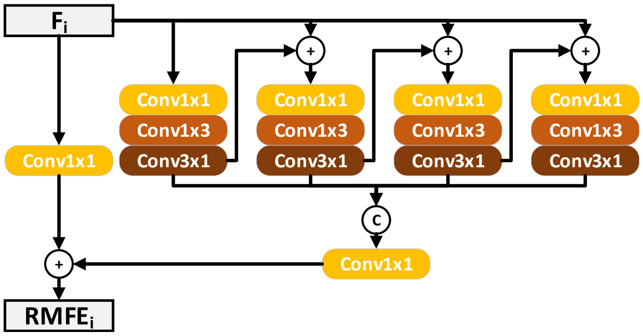

Figure 2: Structure diagram of RMFE. It utilizes stacked residual blocks to enlarge the receptive field layer by layer.

Separated Attention

When delineating salient/camouflaged objects, the information at the boundary between the foreground and background is an important cue. The fusion of information from the interior of the object and the background allows our eyes to perceive the boundary information in an effective manner. Inspired by (Chen et al. 2018), we adopt the mechanism called separated attention, which is the combination of the reverse attention and normal attention to focus on the background and foreground respectively. The module contains two streams. In the first stream, we erase the internal details of the objects to focus on the background. Meanwhile, the internal information of the object is recovered in the second stream to focus on the foreground. Through the synergy of foreground and background information, the separator between them is highlighted, which is the boundary of the object. In detail, each stream of the separated attention is multiplied by the corresponding attention map. The foreground attention map in the ith layer is the result of upsampling on the coarse map of the $(i+1)th$ layer marked as C_{i+1}, written as $W_{fai} = \sigma(C_{i+1})$, where σ denotes the sigmoid function, and the background attention map is the foreground attention map in the ith layer subtracted from 1, which is defined as $W_{bai} = 1 - \sigma(C_{i+1})$. Please note that we expand the channel of all attention maps to 64 before element-wise multiplication. The attention part can be written as

$$Ba_i = Out_i = Conv_s(RMFE_i \otimes expand(W_{bai})), \quad (3)$$

$$Fa_i = Conv_s(RMFE_i \otimes expand(W_{fai})), \quad (4)$$

where $RMFE_i$ denotes the ith layer feature map produced by the RMFE module, W_{bai} and W_{fai} are attention maps, $expand()$ is to expand the channel of attention maps to the same as $RMFE_i$, $Conv_s$ is a 1×1 convolution, $\otimes$ indicates the multiplication operation. Out_i is the coarse output map of the ith layer, which is supervised by the GT map, Ba_i and Fa_i are the feature maps after the attention operation. To discover the contribution of each stream in SEA, we adopt Multi-scale Channel Attention Module (MS-CAM) (Dai et al. 2021), a dual-branch block to get the weight of feature maps in global and local scale. In detail, the weight matrix W can be written as $W(X) = G(\sigma(G(X))) + G(\sigma(L(X)))$. $G(X)$ leads with a global average pooling layer to discover the global information, while $L(X)$ utilizes point-wise convolution to extract local feature. We use Ba_i and Fa_i as the input to MS-CAM. The employment of MS-CAM is beneficial to represent different scales of features in a more general way. After each attention module, we add a Boundary Guider module to enhance the model's ability to understand the boundary, so that the boundaries can be more prominent after the two streams are merged. The introduction of Boundary Guider will be covered in the next section. In the end, we integrate the information of these two streams together by a simple addition operation. The proposed SEA module can be written as

$$SEAF_i = BG_i(W(Ba_i + Fa_i) \otimes Ba_i, Bmap), \quad (5)$$

$$SEAB_i = BG_i((1 - W(Ba_i + Fa_i)) \otimes fa_i, Bmap), \quad (6)$$

$$SEA_i = SEAF_i \oplus SEAB_i \; i = 2, 3, 4, \quad (7)$$

where BG_i denotes the Boundary Guider module, $SEAF_i$ and $SEAB_i$ are the output results of the foreground and background streams respectively, SEA_i is the output of SEA module, $Bmap$ is the predicted boundary map. The structure of the separated attention map is shown in Figure 3.

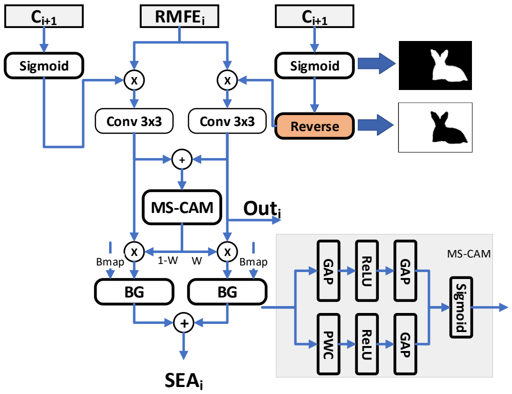

Figure 3: The output of the deep layer is inverted as the weight of the shallow feature map. The normal attention stream and the reverse attention stream focus on the foreground and background respectively. By coordinating the foreground and background information, the boundary of camouflaged object is highlighted. The boundary information is utilized to find camouflaged object.

Boundary Guider

For the SOD task, it is known that predicting the pixels which are close to the boundary of the object(s) is complicated. There are two main reasons for this situation. One is that the distribution of pixels around the boundary is abnormal, and the other is that SOD is a high-resolution task, which requires pixel-level classification. Since many convolution and pooling layers are used to extract features, it requires many upsampling operations like interpolation to restore the resolution, which causes the loss of spatial information to some extent. Such the problem is more obvious in the COD task, since camouflaged objects are concealed and merged in the backgrounds, making the boundaries more blurred. Hence, we try to integrate the boundary information into the feature space to enhance the sensitivity of the model to the boundary. In the beginning, we try to use the ground-truth boundary map (the gradient map of a binary ground truth) as the guidance map, and find that the results are surprisingly good. This is because, with the prior knowledge of an object's boundary, we can easily recognize the object which has similar patterns to the background. Therefore, the boundary information plays an important role in the COD task. However, the ground-truth map is not available in the inference stage. We have to train a network to obtain the boundary map in a supervised way.

In detail, we design a simple network for boundary detection, which concatenates four layer features from the backbone and utilizes a convolution to obtain the boundary map supervised by the boundary map of GT. As shown in Figure 4, the boundary prediction result BM and the feature maps produced in the SEA module marked as Attention Stream Map (ASM) are delivered to a conditional batch normalization (BG) module. In general batch normalization, the parameters of the affine operation (γ and β) are unable to learn enough information without prior knowledge. To address this problem, we use a boundary map to learn these affine parameters. We consider the boundary prediction as our condition and such a module embeds the spatial information into the feature map, which allows the original feature map to learn better boundary features. The formulation of the operation is defined as

$$BGM_i = CB(ASM_i) \otimes \gamma(BM) \oplus \beta(BM), \quad (8)$$

where CB denotes a 3×3 convolution and batch normalization, ASM_i represents the feature map generated in the SEA module mentioned above, γ and β are the affine parameters, each of which contains a 3×3 convolutional layer to encode information about the boundary map and enlarge the channel to 64, which is the same as coarse feature maps.

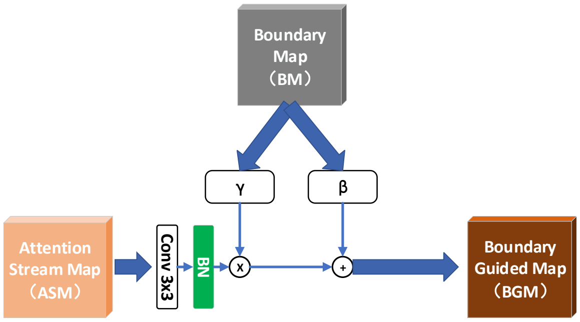

Figure 4: Illustration of the BG module, which uses adaptive space normalization to embed the boundary information into the feature map effectively.

Loss Function

In pixel-wise binary classification tasks like SOD and COD, the binary cross-entropy loss is widely used in many scenarios. But it has an obvious shortcoming when the number of foreground pixels is far less than that of background pixels, the model is heavily biased towards the background, leading to poor performances. Inspired by (Dong et al. 2021), we assign a weight factor for each pixel, which can be written as $w = \sigma|P_n - G_n|$. The weighted BCE loss is defined as

$$\mathcal{L}_{Wbce} = -\sum_{n=1}^{N} w[G_n ln(P_n) + (1-G_n) ln(1-P_n)], \quad (9)$$

where P_n and G_n are the values at pixel n in the predicted map and ground-truth label, which gives larger weights to pixels that are difficult to predict. In addition, we use the IOU loss for map-level supervision. We combine the two loss functions as the total loss for supervision formulated as

$$\mathcal{L}_t = \mathcal{L}_{Wbce} + \mathcal{L}_{IOU}. \quad (10)$$

Our model includes 9 supervised outputs, including 4 coarse maps $(C1, C2, C3, C4)$, 4 refined maps $(R1, R2, R3, R4)$ and 1 boundary map B. Thus, the final total loss function can be represented as

$$\mathcal{L} = \sum_{i=1}^{4} [\mathcal{L}_t(C_i, G) + \mathcal{L}_t(R_i, G)] + \mathcal{L}_{bce}(B, BG), \quad (11)$$

where BG denotes the boundary ground-truth label and i is the index of predictions.

Experiment

Settings

Datasets We evaluate our BSA-Net on three benchmark datasets: CAMO (Le et al. 2019), CHAMELEON (Skurowski et al. 2018) and COD10K (Fan et al. 2020a). In CAMO, the camouflaged object image set consists of 1250 images (1000 images in the training set and 250 images in the test set), and the non-camouflaged object images are collected from MS-COCO (1000 images in the training set and 250 images in the test set). CHAMELEON is a small dataset containing only 76 images. COD10K is currently the largest dataset containing 6000 training images and 4000 test images. Here, both the training set and the test set are the same as (Fan et al. 2020a).

Implementation details We implement BSA-Net on PyTorch (Paszke et al. 2019). Res2Net (Gao et al. 2021), pretrained on ImageNet, is utilized to initialize the backbone (i.e., block1 to block5). We use kaiming-normal to initialize all the convolutional layers and linear layers. We utilize the Adam optimizer to train our model. The learning rate and weight decay are set to 8e-5 and 0.1, respectively. During training, the batch size is set to 36 and the maximum epoch is set to 35. The input image is simply resized to 384 × 384 and then fed into the network to obtain the predicted binary map. The bilinear interpolation operation for image resizing. All the experiments are running with an Nvidia GeForce RTX 3090 GPU (with 24GB memory). The code is available at https://github.com/WolfberryCoke/BSA-Net.

Evaluation metrics We use four evaluation metrics that are widely-used in image segmentation, including E-measure (Fan et al. 2018), S-measure (Fan et al. 2017), weighted F-measure (Margolin, Zelnik-Manor, and Tal 2014) and Mean Absolute Error (Perazzi et al. 2012) denoted as E_ϕ, S_α, F_β^ω and MAE, respectively. **E-measure** is an enhanced-alignment measure that combines local pixels with image-level average values so that both local and global information can be considered simultaneously. **S-measure** is a structure-based metric employed to calculate the structural similarity of objects and regions. **Weighted F-measure** is an improved version of F-measure, utilizing weighted precision and recall. **Mean absolute error** is a pixel-level indicator defined to calculate the pixel-wise difference between the predicted map and the ground-truth map.

Comparison with SOTA Methods

Since COD is relatively new, there are not many methods about this topic, we also compare our method with some typical salient object detection (SOD) methods. Totally, we compare our BSA-Net with 16 state-of-the-art methods, including UNet++ (Zhou et al. 2018), PiCANet (Liu, Han, and Yang 2018), MSRCNN (Huang et al. 2019), BASNet (Qin et al. 2019), PFANet (Zhao and Wu 2019), CPD (Wu, Su, and Huang 2019a), HTC (Chen et al. 2019), EGNet (Zhao et al. 2019), ANet-SRM (Le et al. 2019), SINet (Fan et al. 2020a), PraNet (Fan et al. 2020b), MCIF-Net (Dong et al. 2021), TANet (Ren et al. 2021), R-MGL (Zhai et al. 2021), PFNet (Mei et al. 2021) and Rank-Net (Lv et al. 2021).

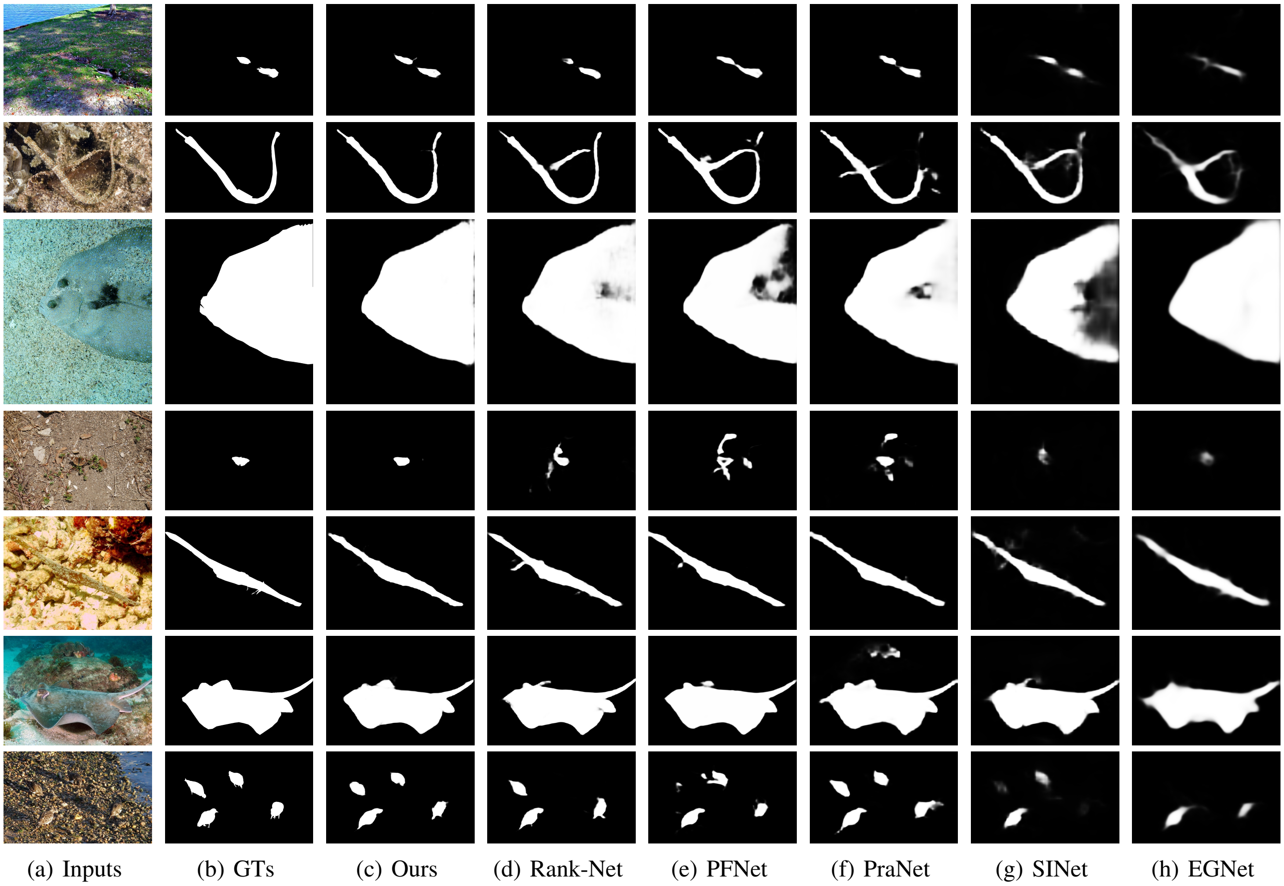

Figure 5: Visual comparisons with SOTAs. BSA-Net produces the COD results with clear boundaries while SOTAs mostly not.

Quantitative comparisons We exploit all four metrics to compare with SOTAs. Table 1 summarizes the results of all methods on three benchmark datasets, where the best ones are highlighted in bold. First, as observed, these SOD methods are not sensitive enough to capture subtle differences between the camouflaged objects and their backgrounds, since the camouflaged objects in the datasets often conceal themselves 'perfectly' in the background instead of highlighting themselves as the salient objects do. Second, our BSA-Net outperforms all the COD methods in terms of the four metrics. For example, compared to SINet on the COD10K dataset, our method increases S_α by **0.047**, E_ϕ by **0.085**, F_β^ω by **0.148** and MAE by **0.017**.

Visual comparisons Figure 5 shows visual comparisons of different methods. As illustrated, our method achieves better results compared to these SOTA methods, which are consistent to the quantitative comparisons. We observe that our proposed BSA-Net can accurately capture the camouflaged objects, and can reduce omissions compared to the other methods. Simultaneously, as shown in the Row 2 and Row 6, our method is more accurate in detecting the boundary of the object than the other methods. That is, it can obtain more detailed boundaries, which also verifies the effectiveness of our utilization of boundary information. It should be noted that our method also has good performance in detecting small camouflaged objects, as shown in the Row 1, Row 4 and Row 7. In a word, the binary map obtained by our method is clearer and more accurate.

Ablation Study

We conduct ablation studies to demonstrate the effectiveness of our components, including the SEA and BG modules. Model a is composed of the backbone network Res2Net and Residual Multi-scale Feature Extractor (RMFE) module; Model b adds the separated attention (SEA) module based on Model a; Model c adds boundary guider (BG) on the Model a, and Model d is our final model. We evaluate the 4 models on three benchmark datasets. Quantitative experimental results are shown in Table 2.

Effectiveness of Separated Attention By comparing Model a with Model b, we observe that Model b outperforms Model a in terms of all the evaluation metrics. It means by adding the separated attention modules, our model can perform better. The apparent improvement in the evaluation metrics shows that SEA can highlight the boundaries of objects by focusing on the foreground and background information separately, thereby improving the accuracy of COD.

Methods	CAMO				CHAMELEON				COD10K			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
2018 UNet++	0.599	0.653	0.392	0.149	0.695	0.762	0.501	0.094	0.623	0.672	0.350	0.086
2018 PiCANet	0.609	0.584	0.356	0.156	0.769	0.749	0.536	0.085	0.649	0.643	0.322	0.090
2019 MSRCNN	0.617	0.669	0.454	0.133	0.637	0.686	0.443	0.091	0.641	0.706	0.419	0.073
2019 BASNet	0.618	0.661	0.413	0.159	0.687	0.721	0.474	0.118	0.634	0.678	0.365	0.105
2019 PFANet	0.659	0.622	0.391	0.172	0.679	0.648	0.378	0.144	0.636	0.618	0.286	0.128
2019 CPD	0.726	0.729	0.550	0.115	0.853	0.866	0.706	0.052	0.747	0.770	0.508	0.059
2019 HTC	0.476	0.442	0.174	0.172	0.517	0.489	0.204	0.129	0.548	0.520	0.221	0.088
2019 EGNet	0.732	0.768	0.583	0.104	0.848	0.870	0.702	0.050	0.737	0.779	0.509	0.056
2019 ANet-SRM	0.682	0.685	0.484	0.126	‡	‡	‡	‡	‡	‡	‡	‡
2020 SINet	0.751	0.771	0.606	0.100	0.869	0.891	0.740	0.044	0.771	0.806	0.551	0.051
2020 PraNet	0.769	0.824	0.663	0.094	0.860	0.907	0.763	0.044	0.789	0.861	0.629	0.045
2021 MCIF-Net	0.784	0.845	0.677	0.084	‡	‡	‡	‡	0.787	0.872	0.636	0.042
2021 TANet	0.793	0.834	0.690	0.083	0.888	0.911	0.786	0.036	0.803	0.848	0.629	0.041
2021 R-MGL	0.775	0.847	0.673	0.088	0.893	0.923	0.813	0.030	0.814	0.865	0.666	0.035
2021 PFNet	0.782	0.841	0.695	0.085	0.882	0.931	0.810	0.033	0.800	0.877	0.660	0.040
2021 Rank-Net	0.787	0.838	0.696	0.080	0.890	0.935	0.822	0.030	0.804	0.880	0.673	0.037
Ours	**0.796**	**0.851**	**0.717**	**0.079**	**0.895**	**0.946**	**0.841**	**0.027**	**0.818**	**0.891**	**0.699**	**0.034**

Table 1: Comparisons with SOTAs for COD on three benchmark datasets. The best results are highlighted in bold.

model	CAMO-Test				CHAMELEON				COD10K-Test			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
a. Baseline	0.783	0.840	0.691	0.083	0.879	0.933	0.807	0.032	0.805	0.880	0.672	0.037
b. Baseline+SEA	0.796	**0.854**	0.716	0.080	0.882	0.934	0.815	0.030	0.811	0.881	0.685	0.037
c. Baseline+BG	0.786	0.848	0.704	0.081	0.875	0.921	0.808	0.031	0.812	0.883	0.686	0.036
d. Ours	**0.796**	0.851	**0.717**	**0.079**	**0.895**	**0.946**	**0.841**	**0.027**	**0.818**	**0.891**	**0.699**	**0.034**

Table 2: Ablation study

Effectiveness of Boundary Guider In order to validate the effectiveness of the boundary map guider, we compare the results of Model b and Model d. After removing the boundary director, the performance of our model decreases. Since the separator between the foreground and the background, that is, the boundary of the camouflaged object contains fewer pixels, we need to exploit the BG module to embed additional boundary information into the feature to strengthen the model's understanding of boundary. With the help of boundary guider, the predicted result can maintain a clear boundary structure of the object.

Conclusion

We propose a novel boundary-guided separated attention network for camouflaged object detection, called BSA-Net. Our BSA-Net well responds to the intriguing question if the synergy of leveraging the background by erasing the foreground, and focusing on the foreground by ignoring the background enhance the capability of cutting-edge COD networks. BSA-Net is inspired by the way of how humans to find camouflaged objects in an image: humans pay attention to the foreground and background of the image to find the difference between them, and when the difference between them is distinguished, the outline of the object can be depicted. Therefore, the boundary information will enhance the capability of camouflaged object detection. Based on the above observation, we first utilize the Residual Multi-scale Feature Extraction module to extract multi-scale features. Then, we use the two-stream Separated Attention modules: one stream focuses on the foreground and ignores the background, and the other stream focuses on the background while erasing the foreground. After each stream, we exploit the Boundary Guider module to embed the boundary information into the features. Finally, the two stream are merged to highlight the boundary of the camouflaged object and strengthen the model's ability to detect the boundary of the object. We conduct experiments on three COD datasets. The experimental results show that our method achieves very competitive performance compared with the sixteen SOTA methods. In the future, we will consider to generate camouflaged objects from natural images and use depth images.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China (No. 62172218, No. 62032011), in part by the Joint Fund of National Natural Science Foundation of China and Civil Aviation Administration of China (No. U2033202), in part by the Free Exploration of Basic Research Project, Local Science and Technology Development Fund Guided by the Central Government of China (No. 2021Szvup060), and in part by the Project of Strategic Importance in The Hong Kong Polytechnic University (No. ZE2Q).

References

Bhajantri, N. U.; and Nagabhushan, P. 2006. Camouflage defect identification: a novel approach. In *9th International*

Conference on Information Technology (ICIT'06), 145–148. IEEE.

Chen, K.; Pang, J.; Wang, J.; Xiong, Y.; Li, X.; Sun, S.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. 2019. Hybrid task cascade for instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4974–4983.

Chen, S.; Tan, X.; Wang, B.; and Hu, X. 2018. Reverse attention for salient object detection. In *Proceedings of the European Conference on Computer Vision (ECCV)*, 234–250.

Chu, H.; Hsu, W.; Mitra, N. J.; Cohen-Or, D.; Wong, T.; and Lee, T. 2010. Camouflage images. *ACM Trans. Graph.*, 29(4): 51:1–51:8.

Dai, Y.; Gieseke, F.; Oehmcke, S.; Wu, Y.; and Barnard, K. 2021. Attentional Feature Fusion. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021, Waikoloa, HI, USA, January 3-8, 2021*, 3559–3568. IEEE.

Dong, B.; Zhuge, M.; Wang, Y.; Bi, H.; and Chen, G. 2021. Towards Accurate Camouflaged Object Detection with Mixture Convolution and Interactive Fusion. *CoRR*, abs/2101.05687.

Fan, D.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M.; and Borji, A. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In Lang, J., ed., *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, 698–704. ijcai.org.

Fan, D.-P.; Cheng, M.-M.; Liu, Y.; Li, T.; and Borji, A. 2017. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE international conference on computer vision*, 4548–4557.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2777–2787.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 263–273. Springer.

Feng, M.; Lu, H.; and Ding, E. 2019. Attentive feedback network for boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1623–1632.

Gao, S.; Cheng, M.; Zhao, K.; Zhang, X.; Yang, M.; and Torr, P. H. S. 2021. Res2Net: A New Multi-Scale Backbone Architecture. *IEEE Trans. Pattern Anal. Mach. Intell.*, 43(2): 652–662.

Huang, Z.; Huang, L.; Gong, Y.; Huang, C.; and Wang, X. 2019. Mask scoring r-cnn. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6409–6418.

Huerta, I.; Rowe, D.; Mozerov, M.; and Gonzàlez, J. 2007. Improving background subtraction based on a casuistry of colour-motion segmentation problems. In *Iberian Conference on Pattern Recognition and Image Analysis*, 475–482. Springer.

Kavitha, C.; Rao, B. P.; and Govardhan, A. 2011. An efficient content based image retrieval using color and texture of image sub blocks. *International Journal of Engineering Science and Technology (IJEST)*, 3(2): 1060–1068.

Le, T.; Nguyen, T. V.; Nie, Z.; Tran, M.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *Comput. Vis. Image Underst.*, 184: 45–56.

Li, A.; Zhang, J.; Lv, Y.; Liu, B.; Zhang, T.; and Dai, Y. 2021. Uncertainty-Aware Joint Salient Object and Camouflaged Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 10071–10081. Computer Vision Foundation / IEEE.

Liu, J.; Zhang, J.; and Barnes, N. 2021. Confidence-Aware Learning for Camouflaged Object Detection. *CoRR*, abs/2106.11641.

Liu, N.; Han, J.; and Yang, M.-H. 2018. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 3089–3098.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D. 2021. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 11591–11601. Computer Vision Foundation / IEEE.

Mao, Y.; Zhang, J.; Wan, Z.; Dai, Y.; Li, A.; Lv, Y.; Tian, X.; Fan, D.; and Barnes, N. 2021. Transformer Transforms Salient Object Detection and Camouflaged Object Detection. *CoRR*, abs/2104.10127.

Margolin, R.; Zelnik-Manor, L.; and Tal, A. 2014. How to evaluate foreground maps? In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 248–255.

Mei, H.; Ji, G.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D. 2021. Camouflaged Object Segmentation With Distraction Mining. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 8772–8781. Computer Vision Foundation / IEEE.

Pang, Y.; Zhao, X.; Zhang, L.; and Lu, H. 2020. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9413–9422.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8024–8035.

Perazzi, F.; Krähenbühl, P.; Pritch, Y.; and Hornung, A. 2012. Saliency filters: Contrast based filtering for salient region detection. In *2012 IEEE conference on computer vision and pattern recognition*, 733–740. IEEE.

Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; and Jagersand, M. 2019. Basnet: Boundary-aware salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7479–7489.

Ren, J.; Hu, X.; Zhu, L.; Xu, X.; Xu, Y.; Wang, W.; Deng, Z.; and Heng, P. 2021. Deep Texture-Aware Features for Camouflaged Object Detection. *CoRR*, abs/2102.02996.

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

Sun, Y.; Chen, G.; Zhou, T.; Zhang, Y.; and Liu, N. 2021. Context-aware Cross-level Fusion Network for Camouflaged Object Detection. In Zhou, Z., ed., *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, 1025–1031. ijcai.org.

Tankus, A.; and Yeshurun, Y. 2001. Convexity-Based Visual Camouflage Breaking. *Comput. Vis. Image Underst.*, 82(3): 208–237.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^{3}Net: Fusion, Feedback and Focus for Salient Object Detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, 12321–12328.

Wei, J.; Wang, S.; Wu, Z.; Su, C.; Huang, Q.; and Tian, Q. 2020. Label Decoupling Framework for Salient Object Detection. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13022–13031. IEEE.

Wu, Y.; Gao, S.; Mei, J.; Xu, J.; Fan, D.; Zhang, R.; and Cheng, M. 2021. JCS: An Explainable COVID-19 Diagnosis System by Joint Classification and Segmentation. *IEEE Trans. Image Process.*, 30: 3113–3126.

Wu, Z.; Su, L.; and Huang, Q. 2019a. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3907–3916.

Wu, Z.; Su, L.; and Huang, Q. 2019b. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7264–7273.

Xue, F.; Guoying, C.; Hong, R.; and Gu, J. 2015. Camouflage texture evaluation using a saliency map. *Multim. Syst.*, 21(2): 169–175.

Yan, J.; Le, T.; Nguyen, K.; Tran, M.; Do, T.; and Nguyen, T. V. 2021. MirrorNet: Bio-Inspired Camouflaged Object Segmentation. *IEEE Access*, 9: 43290–43300.

Yang, Q.-L. Z. Y.-B. 2021. SA-Net: Shuffle Attention for Deep Convolutional Neural Networks. *arXiv preprint arXiv:2102.00240*.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D. 2021. Mutual Graph Learning for Camouflaged Object Detection. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, 12997–13007. Computer Vision Foundation / IEEE.

Zhang, J.; Lv, Y.; Xiang, M.; Li, A.; Dai, Y.; and Zhong, Y. 2021. Depth-Guided Camouflaged Object Detection. *CoRR*, abs/2106.13217.

Zhao, J.-X.; Liu, J.-J.; Fan, D.-P.; Cao, Y.; Yang, J.; and Cheng, M.-M. 2019. EGNet: Edge guidance network for salient object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8779–8788.

Zhao, T.; and Wu, X. 2019. Pyramid feature attention network for saliency detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3085–3094.

Zhou, Z.; Siddiquee, M. M. R.; Tajbakhsh, N.; and Liang, J. 2018. Unet++: A nested u-net architecture for medical image segmentation. In *Deep learning in medical image analysis and multimodal learning for clinical decision support*, 3–11. Springer.

Zhu, L.; Chen, J.; Hu, X.; Fu, C.-W.; Xu, X.; Qin, J.; and Heng, P.-A. 2019. Aggregating attentional dilated features for salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(10): 3358–3371.