Frequency-aware Camouflaged Object Detection

JIAYING LIN*, City University of Hong Kong, Hong Kong
XIN TAN*, East China Normal University, China and City University of Hong Kong, Hong Kong
KE XU†, City University of Hong Kong, Hong Kong
LIZHUANG MA, Shanghai Jiao Tong University, China
RYNSON W.H. LAU†, City University of Hong Kong, Hong Kong

Camouflaged object detection (COD) is important as it has various potential applications. Unlike salient object detection (SOD), which tries to identify visually salient objects, COD tries to detect objects that are visually very similar to the surrounding background. We observe that recent COD methods try to fuse features from different levels using some context aggregation strategies originally developed for SOD. Such an approach, however, may not be appropriate for COD as these existing context aggregation strategies are good at detecting distinctive objects while weakening the features from less discriminative objects. To address this problem, we propose in this paper to exploit frequency learning to suppress the confusing high-frequency texture information, to help separate camouflaged objects from their surrounding background, and a frequency-based method, called FBNet, for camouflaged object detection. Specifically, we design a frequency-aware context aggregation (FACA) module to suppress high-frequency information and aggregate multi-scale features from a frequency perspective, an adaptive frequency attention (AFA) module to enhance the features of the learned important frequency components, and a gradient-weighted loss function to guide the proposed method to pay more attention to contour details. Experimental results show that our model outperforms relevant state-of-the-art methods.

CCS Concepts: • **Computing methodologies → Computer vision tasks**.

Additional Key Words and Phrases: Camouflaged object detection, frequency learning

1 INTRODUCTION

Camouflaged object detection (COD) aims to detect objects with similar patterns to their surroundings. It has many potential applications in various fields, *e.g.*, medical image segmentation of polyps and COVID-19 infected regions. Works in art and information forensics [12, 52] attempt to hide information in natural images, and COD techniques may help reveal this camouflaged information.

The COD problem was first addressed using deep-learning in [20], which adopted a two-stream network for classification and COD segmentation. Recently, Fan *et al.* [10] proposed a SINet model for COD and achieved a significant performance improvement over existing methods. It uses a modified Receptive Field Block (RFB) [28] to search for camouflaged objects. RFB was first proposed for object detection [28], to enhance the deep features of discriminative regions, and then extended for salient object detection (SOD) [45]. The objective of SOD is

*Both authors contributed equally to this research.
†Both authors are the joint corresponding authors.

Authors' addresses: Jiaying Lin, jiayinlin5-c@my.cityu.edu.hk, City University of Hong Kong, Hong Kong, Hong Kong; Xin Tan, tanxin2017@sjtu.edu.cn, East China Normal University, Shanghai, China and City University of Hong Kong, Hong Kong; Ke Xu, City University of Hong Kong, Hong Kong, kkangwing@gmail.com; Lizhuang Ma, ma-lz@cs.sjtu.edu.cn, Shanghai Jiao Tong University, Shanghai, China; Rynson W.H. Lau, Rynson.Lau@cityu.edu.hk, City University of Hong Kong, Hong Kong.

1551-6857/2022/6-ART $15.00
https://doi.org/10.1145/3545609

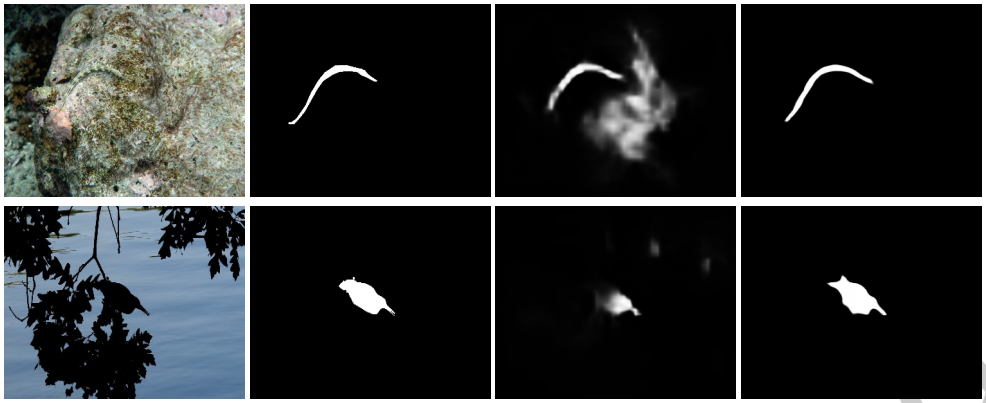

Fig. 1. Two common scenarios where the state-of-the-art COD method [10] fails. Top row: as SINet can be easily affected by high-frequency information due to the use of a context aggregation strategy developed for SOD, it mistakenly recognizes the background (with a similar pattern to the object) as part of the object. Bottom row: SINet fails to detect the complete object as it is not able to gather sufficient low-frequency shape information. In contrast, our method can detect these objects accurately by exploiting frequency learning in context aggregation.

to identify discriminative objects in the input image, and RFB can obtain discriminative features through (i) concatenation of multi-scale features and (ii) addition (*i.e.*, short connection) between input and output features. However, we observe that detecting camouflaged objects is very different from detecting generic or salient objects. Since camouflaged objects would deliberately conceal themselves by changing their own appearances to match with their surroundings, simply adopting these existing context aggregation strategies for COD may lead to a lower performance. As shown in the first row of Figure 1, although SINet can recognize part of the camouflaged object, it over-predicts the object region and mis-recognizes part of the background to belong to the object, as confused by the high-frequency texture shared by the camouflaged object and its surrounding. In the second row of Figure 1, SINet fails to detect the camouflaged object correctly, as it is unable to learn and aggregate sufficient low-frequency shape information of the camouflaged object during context aggregation. In this work, we aim to address this context aggregation problem for COD from a frequency-based perspective.

As pointed out by [6], information extracted from less discriminative objects may be weakened during context aggregation, *e.g.*, using Atrous Spatial Pyramid Pooling (ASPP) [3] or RFB [28]. Geirhos *et al.* [13] also observed that CNNs would bias towards high-frequency information like textures in images. Such a bias can affect the performance of existing context aggregation strategies when they are applied to COD, as the patterns of camouflaged objects are similar to those of their surroundings. It may also make it more difficult for the network to detect accurate boundaries of camouflaged objects. These observation inspire us to consider a better design for COD with suppressing the confusing high-frequency texture feature and enhancing the low-frequency shape information from images with camouflaged objects.

To tackle the above problems, we propose in this paper a novel frequency-based method, named *FBNet*, to both reduce the bias caused by high-frequency texture information and gather low-frequency shape information from camouflaged images. Our FBNet is based on two novel modules. First, we propose a Frequency-aware Context Aggregation (FACA) module to suppress confusing high-frequency texture information and aggregate multi-scale features via frequency modeling, which can benefit the learning of the differences between camouflaged objects and their surrounding backgrounds from a frequency perspective. Second, we propose an Adaptive Frequency Attention

(AFA) mechanism to enhance the features of the learned important frequency components. Besides, we note that camouflaged objects usually have complex boundaries, we further design a gradient-weighted loss function to help our model focus on the contours of camouflaged objects.

Our main contribution can be summarized as follows:

- We propose a frequency-based approach to address the limitation of existing context aggregation strategies when used in COD. Based on this frequency-based approach, we propose FBNet to exploit frequency learning for COD.
- We propose two novel blocks for FBNet: a frequency-aware context aggregation (FACA) module to disentangle frequency modeling in context aggregation, and an adaptive frequency attention (AFA) mechanism to learn to enhance the features of the learned important frequency components. We further propose a gradient-weighted loss function to guide FBNet to focus on the contours of camouflaged objects.
- Our extensive experimental evaluations demonstrate the superiority of the proposed model over state-of-the-art methods from relevant fields.

2 RELATED WORK

In this section, we first briefly survey related works on context aggregation strategies. We then summarize recent works on camouflaged object detection (COD) and salient object detection (SOD).

Context Aggregation. This aims to aggregate contextual features for different tasks, including object detection and semantic segmentation. PSPNet [54] uses the Pyramid Pooling Module (PPM) to extract multi-scale contextual representations for semantic segmentation. Chen *et al.* [3] proposed Atrous Spatial Pyramid Pooling (ASPP) to concatenate spatial features extracted by convolutional kernels with different dilation rates. Liu *et al.* [28] proposed the Receptive Field Block (RFB) to adopt suitable kernel sizes of different dilation rates to improve ASPP [3].

Due to the success of the context aggregation strategies applied to general detection tasks, recent SOD methods also adopt a similar approach in their network design. For example, Zhao *et al.* [56] proposed a Context-aware Pyramid Feature Extraction module (CPFE) based on ASPP. Wu *et al.* [46] explored how to aggregate segmentation features and edge features for SOD. The state-of-the-art method for COD [10] also proposed a modified RFB [28] module for context aggregation.

However, as CNNs would bias towards extracting high-frequency information (*e.g.*, textures) [13] while context aggregation tends to weaken the features from less discriminative objects [6], existing context aggregation strategies may not be suitable for COD. In this work, we propose a new context aggregation strategy that focuses more on extracting low-frequency shape information for COD.

Camouflaged Object Detection (COD). It aims to detect objects with similar patterns (*e.g.*, textures and colors) to their natural habitats. Early methods [36, 48, 51] were mostly based on hand-crafted, low-level features, and they mainly focused on detecting camouflaged regions in the given images. COD has lots of potential applications, including mirror detection [14, 24, 41] and glass surface detection [23, 35].

Recent CNN-based methods [10, 20] have achieved significant improvement on COD. Le *et al.* [20] proposed the first CNN-based model, ANet-SRM, and a benchmark dataset (consisting of 2500 images with pixel-level annotations) for COD. Fan *et al.* [10] proposed the largest dataset COD10K with 10K images and a state-of-the-art model, SINet, for COD. Mei *et al.*[34] proposed a bio-inspired network, PFNet, for COD. However, these latest COD methods [10, 20] are based on some context aggregation strategies, which may not be suitable for COD. In addition, these COD methods have complex frameworks. While ANet-SRM [20] uses a classification-segmentation pipeline, SINet [10] uses a search-identification pipeline.

To address these limitations, we propose in this work to exploit frequency learning to suppress the confusing high-frequency texture information and a new context aggregation strategy to extract low-frequency shape information, enabling efficient and accurate camouflaged object detection.

Salient Object Detection (SOD). It aims to detect objects that are most salient to humans. While early methods relied heavily on hand-craft features and saliency priors [49, 60], recent methods are mostly based on CNNs to enhance feature extraction [43, 53]. Zhang *et al.* [53] proposed a progressive attention based network for adaptive multi-scale context integration. Wang *et al.* [43] presented a new pyramid pooling module as well as a multi-stage refinement mechanism to capture detailed spatial information. Zhao *et al.* [56] proposed a context-aware pyramid feature extraction module to extract pyramid context features. BASNet [38] and SCRN [46] leveraged boundaries of salient objects as explicit guidance to encourage a finer segmentation. Pang *et al.* [37] proposed a multi-scale network to integrate features in an interactive way. GateNet [58] explored a gating method for SOD to adaptively control the information flow in a balanced way.

However, camouflaged objects are very different from salient objects. While salient objects have very different appearances from their surroundings, camouflaged objects tend to have very similar appearances in order to conceal themselves. Hence, we argue that SOD methods are not suitable for COD.

3 OUR METHOD

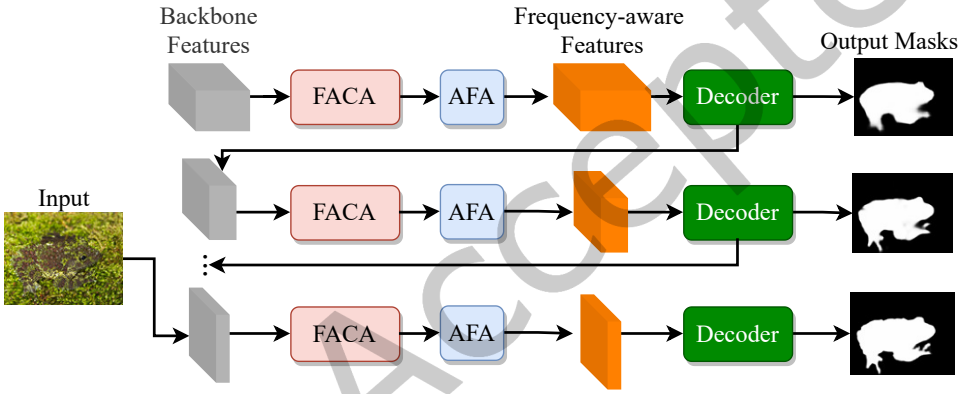

Fig. 2. The pipeline of our proposed FBNet. It leverages the frequency-aware context aggregation (FACA) module to disentangle frequency modeling in context aggregation, and the adaptive frequency attention (AFA) mechanism to enhance the features of the learned important frequency components, for the COD task.

In this paper, we propose a novel frequency-based method for camouflaged object detection. Our method aims to both reduce the bias caused by high-frequency texture information and adaptively enhance the feature representation from learned important frequency components for camouflaged object detection.

Figure 2 illustrates the pipeline of our proposed FBNet. We first feed the input image to a backbone network [17] to extract multi-scale backbone features. Specifically, we use the outputs of four stages of the network, *i.e.*, *res2c*, *res3b3*, *res4b22* and *res5c*, as our backbone features. The deepest features from the final stage (*i.e.*, *res5c*) are first fed into the proposed FACA module to capture the frequency-aware features while suppressing the confusing high-frequency texture information. The output features are then fed into an adaptive frequency attention (AFA) module to enhance the features of the learned important frequency components in an adaptive way. The enhanced features are further fed into a decoder to generate a coarse binary mask of the potential camouflaged objects. This coarse mask serves as an attention map to the previous stage (*i.e.*, *res4b22*) to guide the refinement of the camouflaged object mask. We use the predicted mask of the first stage as the final output. In this way, the output mask is progressively refined by integrating with the earlier backbone features.

3.1 Frequency-aware Context Aggregation (FACA)

To suppress the confusing high-frequency texture information in camouflaged images, we propose a frequency-aware context aggregation (FACA) module to aggregate multi-scale features while preserving the information of less discriminative regions from the input image.

Figure 3 shows the structure of our proposed FACA. We first define the input features to the FACA module as f_{in}, and each set of intermediate features as $f_k = \text{Conv}_{s_k,d_k}(f_{in})$, where s_k is the kernel size and d_k is the dilation rate of the "k^{th}" convolution kernel. In the original ASPP [3], all convolution kernels Conv_k corresponding to f_k have the same kernel size but different dilation rates. As discussed in [28], such a design probably leads to confusion between the object and its context. RFB [28] improves ASPP by adjusting the kernel size s_k for Conv_k and adding shortcut connections between the input features f_{in} and output features f_o, after aggregating all intermediate features f_k. However, it still treats all f_k equally, and directly concatenates all f_k without any enhancement or filtering. As a result, the information of a less discriminative object would be weakened during the aggregation.

Unlike ASPP and RFB, our FACA computes the intermediate features f_k by feeding the input features f_{in} into a learnable low-pass layer LP_k, followed by the convolution kernel Conv_{s_k,d_k}. All permuted pairs of f_k are aggregated by a subtraction operator and then concatenated channel-wise. The low-pass layer can help filter the confusing high-frequency texture information in the image features and preserve the important information from f_{in} in f_k across different levels. In our FACA, inspired by [61], we design a convolution kernel as low-pass layer LP_k for each spatial location of the input feature to filter high-frequency information and learn to preserve important information in a feature level. Following [61], we apply a softmax layer on the weights of the low-pass layer to constrict them as positive values, and we initialize the weights of LP_k by Kaiming Normal initialization [16] to ensure LP_k be Gaussian-like.

Specifically, given a set of input feature $\mathbf{X}$, we construct a low-pass filter $z_{h,w}$ by a 3×3 convolution kernel for each spatial location (h, w) on $\mathbf{X}$. The predicted $z_{h,w}$ is then applied to $\mathbf{X}$ to produce the high-frequency suppressed feature $\mathbf{F}$ as:

$$\mathbf{F} = \sum_{h=0}^{H-1} \sum_{w=0}^{W-1} \sum_{p,q \in \Omega} z_{h,w}^{p,q} \cdot X_{h+p,w+q}, \tag{1}$$

where $\mathbf{F}$ denotes output features at location (h, w) and Ω denotes the region surrounding (h, w) on which we apply the predicted $z_{h,w}$.

Finally, existing context aggregation strategies, such as ASPP and RFB, typically use an addition operator to capture the features from discriminative regions. In contrast, we use the subtraction operator here in order to suppress the features representing high-frequency camouflaged textures. Formally, we can compute the output features f_o as:

$$\begin{aligned} f_k &= \text{Conv}_{s_k,d_k}(LP_k(f_{in})), \\ f_o &= \bigotimes_{i=1}^{n} \bigotimes_{j=i+1}^{n} (f_i - f_j), \end{aligned} \tag{2}$$

where $n = 4$ in Figure 3. $\bigotimes$ represents concatenation followed by a 1×1 convolution. s_k is the kernel size. d_k is the dilation rate of the corresponding k^{th} convolution kernel. For each FACA module in each stage, we set n, s_k and d_k to 4, $\{3, 3, 3, 3\}$ and $\{2, 4, 6, 8\}$, respectively.

3.2 Adaptive Frequency Attention (AFA)

The benefit of attention mechanisms has been shown in different vision tasks. One of the most widely used attention mechanisms is the SE-block [18], which aims to explore attention in a channel-wise manner. However, as proven by [7], the SE-block only models the lowest frequency components by applying global average pooling (GAP) on

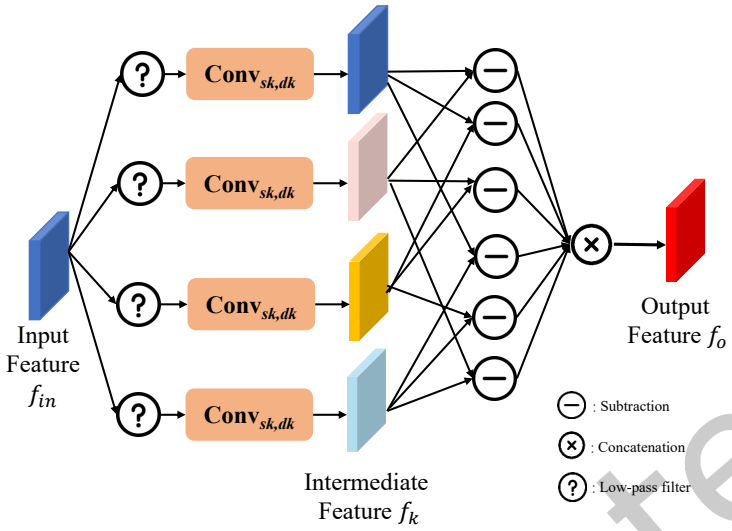

Fig. 3. The structure of our proposed FACA. It aims to aggregate multi-scale features and preserve the information of less discriminative regions from the input feature.

the input features. As such, it may not be suitable for COD, due to the lack of frequency modeling. Recently, a frequency channel attention mechanism (FCA) [39] was proposed to extend the SE-block by replacing GAP with multiple frequency components. However, FCA requires a heavy pre-computation on the selection of frequency components, and the selection of frequency components is empirical and fixed before training. Thus, it is difficult to generalize FCA to different datasets, especially to the camouflaged object datasets, which have a large distribution gap with the general object datasets (*e.g.*, ImageNet [22]). To address these limitations of the attention mechanisms when used on COD, we propose an adaptive frequency attention (AFA) mechanism here. Unlike FCA, our proposed AFA selects frequency components in an adaptive way.

Formally, given the input features $\mathbb{F} \in \mathbb{R}^{C\times H\times W}$, our AFA first computes the discrete cosine transform (DCT) weight $\Gamma^{i,j}_{h,w}$ for each spatial pixel (h, w) in the input features $\mathbb{F}$, where (i, j) is the index of the frequency components in the DCT matrix. We also design a learnable parameter α to learn the importance of each frequency component:

$$\begin{aligned} \Gamma^{i,j}_{h,w} &= \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right), \\ \mathbb{M} &= \sum_{i=0}^{K-1}\sum_{j=0}^{K-1}\sum_{h=0}^{H-1}\sum_{w=0}^{W-1}\alpha_{iK-i+j}\mathbb{F}_{h,w}\Gamma^{i,j}_{h,w}, \end{aligned} \tag{3}$$

where $K\times K$ is the size of the DCT square matrix. Typically, we adopt a 8×8 DCT matrix to represent the frequency space. Finally, the output features $\widetilde{\mathbb{F}}$ are obtained by following a similar structure as the SE-block:

$$\widetilde{\mathbb{F}} = \sigma(\mathbf{W}_2(\delta(\mathbf{W}_1\mathbb{M}))), \tag{4}$$

where σ, δ, $\mathbf{W}_1 \in \mathbb{R}^{\frac{C}{r}\times C}$ and $\mathbf{W}_2 \in \mathbb{R}^{C\times\frac{C}{r}}$ denote the sigmoid function, the ReLU function and two learnable weights, respectively. Following [18], we set r to 16 for a good tradeoff between accuracy and complexity.

3.3 Gradient-Weighted Loss

Most CNN-based existing methods [10, 20] for COD adopt cross entropy (CE) as their training loss function. However, as CE treats all pixels equally, the boundaries of camouflaged objects can be easily missed or blurred by the methods trained with CE, especially when the boundaries are complex.

To tackle this problem, we propose a weighted loss that puts more attention on the boundary details. We denote the predicted mask as P and the ground truth mask as G. To obtain the spatial importance of each location in G, we apply a Laplacian filter ∇ on G and compute the difference between the normalized ∇G and G. The weight of spatial importance w can be computed as (with σ representing the sigmoid function):

$$w = |\sigma(\nabla G) - G|\,, \tag{5}$$

$$\ell_{bce} = -w\left[G \cdot \log \sigma(P) + (1 - G) \cdot \log(1 - \sigma(P))\right]. \tag{6}$$

Our final loss function is then:

$$\ell = \sum_{i=1}^{N} \ell_{bce}^{i} + \ell_{IoU}^{i}, \tag{7}$$

where ℓ_{bce}^{i} is the weighted loss and ℓ_{IoU}^{i} is the IoU loss [38] for the i^{th} predicted camouflaged object mask.

4 EXPERIMENTS

4.1 Datasets and Evaluation Metrics

Following SINet [10], we train our model on the combined dataset with 4,640 images and evaluate our method on three datasets: CHAMELEON [1] with 76 images, CAMO-Test [20] with 250 images, and COD10K-Test [10] with 2,066 images.

We employ four popular metrics to evaluate the performance of our model quantitatively: S-measure (S_α) [8], mean E-measure (E_ϕ) [9], weighted F-measure (F_β^ω) [33], and Mean Absolute Error (MAE). MAE is the average pixel-wise error between the predicted mask and ground truth as:

$$MAE = \frac{1}{HW} \sum_{i=1}^{H} \sum_{j=1}^{W} |P(i, j) - G(i, j)|, \tag{8}$$

where P is the predicted mask. G is the ground truth. H and W are the width and height of the input image.

4.2 Implementation Details

We use ResNet-50 [17] pre-trained on ImageNet as our backbone, and the remaining layers are initialized randomly with the default setting in PyTorch. All training and test images are uniformly resized to 352×352. We use stochastic gradient descent (SGD) as the optimizer with a momentum of 0.9 and a weight decay of 5×10^{-4}. We adopt the "Poly" decay strategy [29], where the current learning rate is the base learning rate multiplied by $(1 - \frac{current_{iter}}{max_{iter}})^{Power}$. The base learning rate is 0.001 and *Power* is 0.9. The batch size is 32. We run 40 epochs in the training.

4.3 Comparison with the State-of-the-Arts

Quantitative Evaluation. We compare our proposed method with the state-of-the-art methods from relevant fields, including SINet [10] and PFNet [34] for camouflaged object detection; MINet [37], GCPA [4], F3Net [44], SCRN [46], DeepCRF [47], BANet [40], EGNet [55], CPD [45], PFANet [57], BASNet [38], PoolNet [26] and PiCANet [27] for saliency object detection; FPN [25], PSPNet [54] and UNet++ [59] for semantic segmentation; MaskRCNN [15], MSRCNN [19] and HTC [2] for instance segmentation. For fair comparison, all methods use

ResNet50 [17] as the backbone network for feature extraction, and are trained in the same way as SINet [10] on their combined training set.

Table 1 shows the quantitative comparison on the four metrics. We can see that our proposed method outperforms all the other methods on most metrics and all three datasets, except for two results are the second best. It is interesting to note that although SINet is designed for COD, some recent SOD methods such as F3Net [44] and SCRN [46] outperform it. This shows that context aggregation strategies used in these two methods may be more effective in COD than the one used in SINet.

Table 1. Quantitative results on three COD datasets. The best three results are shown in red (best), green (second), and blue (third). The first group of methods are semantic and instance segmentation methods. The second group of methods are SOD methods. The last group of methods are COD methods. For fair comparison, all methods are trained in the same way as SINet [10] on their combined training set.

Baseline Models	CHAMELEON [1]				CAMO-Test [20]				COD10K-Test [10]			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
2017 FPN [25]	0.794	0.783	0.590	0.075	0.684	0.677	0.483	0.131	0.697	0.691	0.411	0.075
2017 MaskRCNN [15]	0.643	0.778	0.518	0.099	0.574	0.715	0.430	0.151	0.613	0.748	0.402	0.080
2017 PSPNet [54]	0.773	0.758	0.555	0.085	0.663	0.659	0.455	0.139	0.678	0.680	0.377	0.080
2018 UNet++ [59]	0.695	0.762	0.501	0.094	0.599	0.653	0.392	0.149	0.623	0.672	0.350	0.086
2019 MSRCNN [19]	0.637	0.686	0.443	0.091	0.617	0.669	0.454	0.133	0.641	0.706	0.419	0.073
2019 HTC [2]	0.517	0.489	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.520	0.221	0.088
2018 PiCANet [27]	0.769	0.749	0.536	0.085	0.609	0.584	0.356	0.156	0.649	0.643	0.322	0.090
2019 PoolNet [26]	0.776	0.779	0.555	0.081	0.702	0.698	0.494	0.129	0.705	0.713	0.416	0.074
2019 BASNet [38]	0.687	0.721	0.474	0.118	0.618	0.661	0.413	0.159	0.634	0.678	0.365	0.105
2019 PFANet [57]	0.679	0.648	0.378	0.144	0.659	0.622	0.391	0.172	0.636	0.618	0.286	0.128
2019 CPD [45]	0.853	0.866	0.706	0.052	0.726	0.729	0.550	0.115	0.747	0.770	0.508	0.059
2019 EGNet [55]	0.872	0.895	0.749	0.040	0.754	0.773	0.625	0.098	0.780	0.813	0.581	0.048
2019 BANet [40]	0.647	0.739	0.447	0.115	0.632	0.711	0.471	0.149	0.663	0.734	0.434	0.090
2019 DeepCRF [47]	0.560	0.535	0.270	0.142	0.607	0.601	0.379	0.167	0.606	0.591	0.296	0.108
2019 SCRN [46]	0.878	0.894	0.751	0.041	0.777	0.795	0.644	0.090	0.791	0.820	0.588	0.046
2020 F3Net [44]	0.854	0.901	0.747	0.046	0.779	0.825	0.666	0.092	0.784	0.851	0.611	0.047
2020 GCPA [5]	0.851	0.842	0.687	0.051	0.744	0.727	0.563	0.109	0.763	0.748	0.520	0.055
2020 MINet [37]	0.847	0.895	0.749	0.042	0.731	0.759	0.604	0.096	0.767	0.816	0.598	0.042
2021 Swin-T [30]	0.873	0.916	0.795	0.035	0.744	0.800	0.634	0.090	0.761	0.821	0.586	0.045
2021 Swin-B [30]	0.906	0.938	0.854	0.021	0.860	0.912	0.811	0.048	0.851	0.906	0.748	0.026
2020 SINet [10]	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
2021 RankNet [32]	0.846	0.913	0.767	0.045	0.712	0.791	0.583	0.104	0.767	0.861	0.611	0.045
2021 JCOD [21]	0.870	0.924	NA	0.039	0.792	0.839	NA	0.082	0.800	0.872	NA	0.041
2021 PFNet [34]	0.882	0.942	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
Ours	0.888	0.939	0.828	0.032	0.783	0.839	0.702	0.081	0.809	0.889	0.684	0.035

Qualitative Evaluation. Figure 4 shows visual comparison on some challenging images from the test sets of the three COD datasets. We can see that the proposed method outperforms current state-of-the-art methods in these

Table 2. Ablation Study on using different context aggregation strategies. All ablation models are trained on the combined training set in [10] and tested on COD10K-test.

Ablation	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
ASPP [3]	0.666	0.708	0.449	0.092
RFB [28]	0.754	0.813	0.547	0.055
FACA (ours)	0.809	0.889	0.684	0.035

Table 3. Ablation Study on using different attention mechanisms. All ablation models are trained on the combined training set in [10] and tested on COD10K-test.

Ablation	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
SE-Block [18]	0.790	0.879	0.652	0.041
FCA [39]	0.777	0.864	0.632	0.044
AFA (Ours)	0.809	0.889	0.684	0.035

challenging cases. In particular, the first seven row shows a camouflaged object with patterns very similar to its surrounding. Our method produces the sharpest and most accurate boundary. The eighth and ninth rows show two partially occluded camouflaged objects. Both objects have two parts. The object in the second row requires long-range context information to detect, as it covers a large extent, while the one in the third row has a larger body but a smaller tail. Our method can correctly detect both parts in both cases. The tenth to sixth rows show images with multiple camouflaged objects, where the sizes of the camouflaged objects include medium, small and mixed, respectively. While other methods tend to miss some objects in some cases but overpredict them in other cases, our method can detect all objects in different scales correctly. The final row is an extremely challenging case, as the target object is difficult to see even for humans. Only our method can detect it successfully.

Computation Time. We compare the inferring time of our method and the state-of-the-art COD method (SINet) on a 352×352 image. While SINet takes 0.10s, our method takes only 0.03s. Hence, our method is roughly 3.3 times faster than SINet, due to the simpler structure of our model.

4.4 Ablation Study

Context Aggregation Strategies. To evaluate how different context aggregation strategies perform under the same setting, we have experimented three strategies on our FBNet framework: ASPP [3], RFB [28] and our FACA. Table 2 shows the experimental results. We can see that our FACA performs better than ASPP and RFB. This is because FACA can aggregate contextual features with the suppression of redundant high-frequency information by its low-pass filters and subtraction aggregation operator. Figure 5 shows a visual example to compare the three context aggregation strategies in resolving an ambiguous scene.

Attention Mechanisms. Table 3 compare our proposed AFA with two other attention mechanisms: SE-Block [18] and FCA [39]. It may be interesting to see that although FCA has shown to outperform the SE-Block in many other research problems, such as image recognition and object detection, it degenerates on COD and performs worse than the SE-Block, which only models the lowest frequency components in its squeezing operation. The main reason is that FCA lacks adaptive frequency learning on COD and is more susceptible to the high-frequency textures in camouflaged images. We can see that our proposed AFA outperforms other attention mechanisms, due to the proper and adaptive frequency modeling.

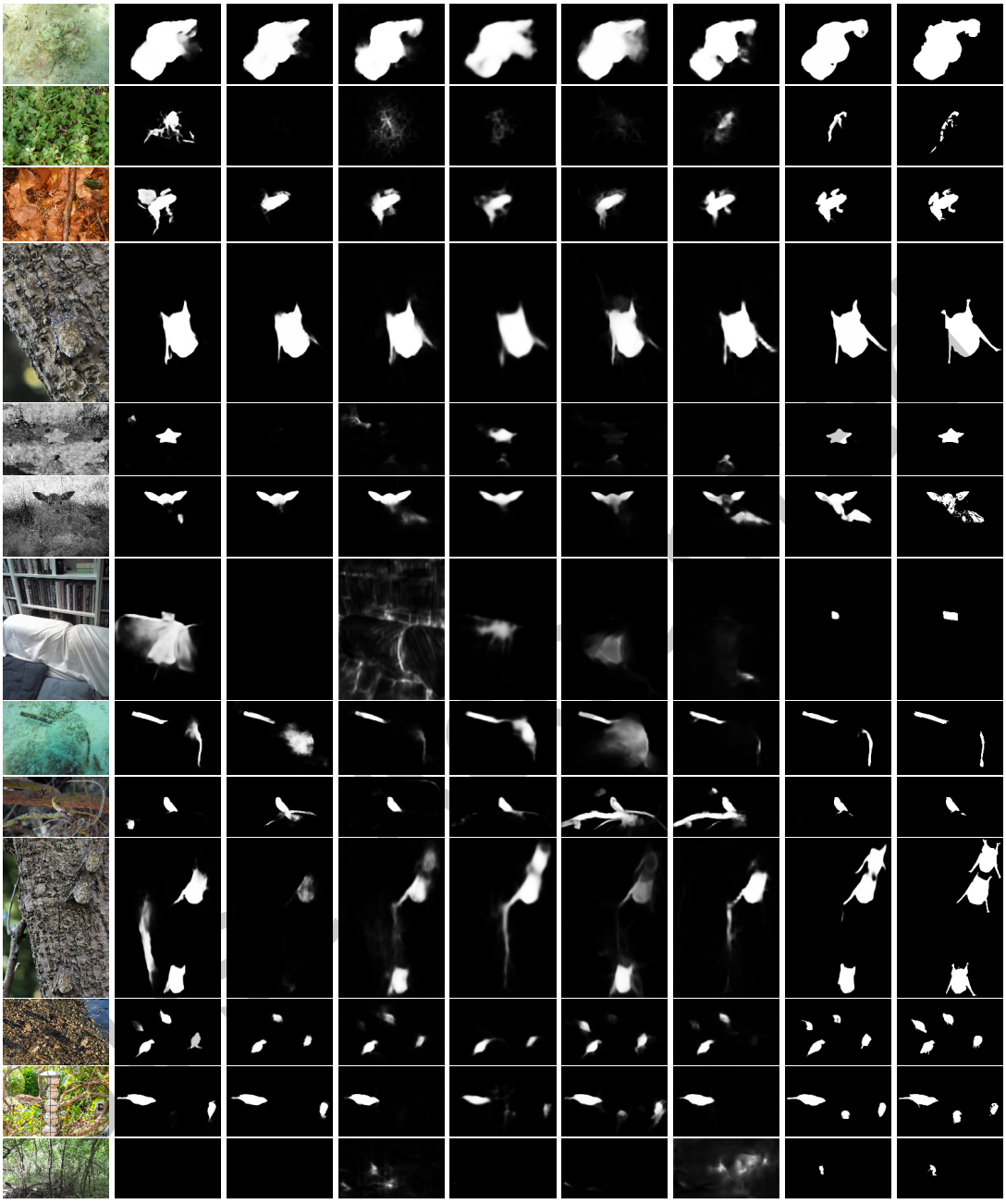

Fig. 4. Visual comparison with state-of-the-art methods. For fair comparison, all methods are trained in the same way as SINet [10] on their combined training set.

Fig. 5. A visual example of the ablation study.

Operator for Context Aggregation. To directly compare how the operators affect the performances of different context aggregation strategies, we perform a experiments based on our FACA-Net in Table 4. We experiment with four operators: addition, concatenation, subtraction and multiplication. We can find that using subtraction as the contextual aggregation FACA outperforms among all ablation models.

Table 4. Ablation Study. All ablation models are trained on the combined training set in [10] and tested on COD10K-test. We denote that Add, Cat, Sub, Mul is addition, concatenation, subtraction, multiplication operator.

Ablation	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
FACA Add	0.788	0.871	0.641	0.041
FACA Cat	0.782	0.874	0.650	0.042
FACA Mul	0.785	0.867	0.632	0.044
FACA Sub (ours)	0.809	0.889	0.684	0.035

Visualization of ERFs of Different Context Aggregation Strategies. Figure 6 visualizes the ERFs [31] of different back-propagated positions in the input image. Comparing with other methods, the ERFs of our method focus on the target object especially on its low-frequency shape information. In the second row of Figure 6, although the back-propagated position moves out of the object, the ERFs of our proposed FACA are still compact and less interfered by the background texture due to the suppression of high-frequency texture information in FACA.

Loss Functions. Table 5 shows the experimental results of different loss functions. We select binary cross entropy and IoU loss [38] as baseline loss functions. Our proposed loss function outperforms the others. Figure 7 shows a visual example to compare the results of the models trained by three different loss functions.

5 DISCUSSION

Adding an explicit edge detection module to various detection tasks has been shown to be very effective [46, 55], and is becoming very popular in recent works. In our model, a straightforward way to learn the edge and shape information is to apply an edge detection module as an explicit guidance during network training. As such, we add an edge guidance module derived from EGNet [55], which has been shown to help improve the SOD performance, in our FBNet to study if adding this edge guidance may help improve the COD performance. We then compare our model with two strong baselines EGNet [55] and SCRN [46] (from Table 1), which use explicit edge guidance

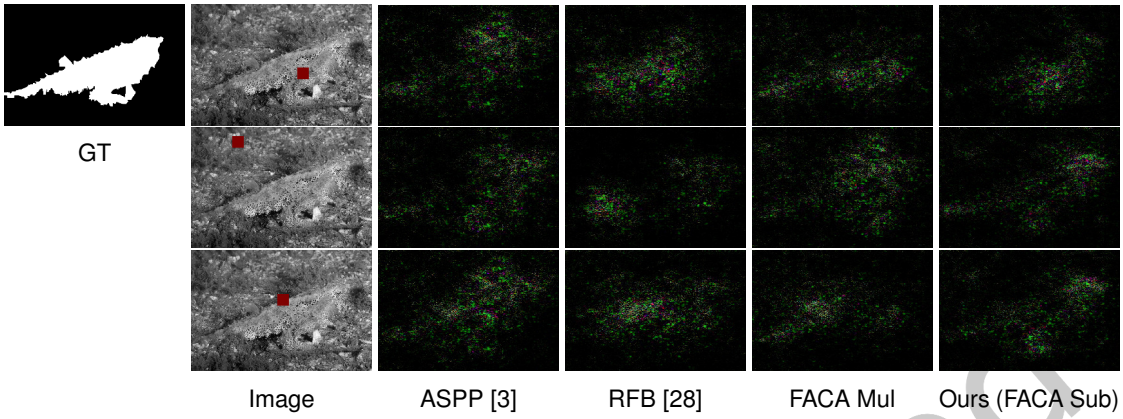

Fig. 6. Visualization of the effective receptive fields (ERFs) of specific back-propagated positions, *i.e.*, inside, outside and on the edge of the target object (red dots). While other strategies have messy ERFs, the ERFs of our method are more compact and focus on the object.

Table 5. Ablation Study on using different loss functions. All ablation models are trained on the combined training set in [10] and tested on COD10K-test.

Ablation	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
BCE	0.775	0.784	0.551	0.051
IOU [38]	0.777	0.857	0.654	0.038
Ours	0.809	0.889	0.684	0.035

Fig. 7. A visual example of the ablation study on using different loss functions. The method trained by our proposed loss function predicts more precise contours and shapes of the camouflaged objects.

during their network training. Table 6 shows the results of EGNet, SCRN, and our model with and without the edge guidance module. It is interesting to see that the edge guidance module does not really help improve the COD performances in all three cases. We believe that this is because camouflaged objects deliberately conceal their boundaries, so that they cannot be easily detected by their predictors. As such, explicitly supervising the networks with edge guidance does not benefit too much for COD.

6 CONCLUSION

In this paper, we have investigated different context aggregation strategies for camouflaged object detection (COD), and proposed a new frequency-aware context aggregation (FACA) module and adaptive frequency attention (AFA)

Table 6. Evaluation of the effectiveness of edge guidance on COD. All baselines are trained on the training dataset provided by [10] and tested on COD10K-test. EG stands for explicit edge guidance.

Ablation models	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^\omega \uparrow$	MAE↓
EGNet (with EG) [55]	0.780	0.813	0.581	0.048
EGNet (without EG)	0.779	0.814	0.580	0.047
SCRN (with EG) [46]	0.791	0.820	0.588	0.046
SCRN (without EG)	0.790	0.819	0.584	0.046
Ours (with EG)	0.807	0.884	0.687	0.037
Ours (without EG)	0.809	0.889	0.684	0.035

mechanism to aggregate multi-scale features broadly but selectively from important frequency components. Based on FACA and AFA, we present the FBNet model for COD. We have also proposed a novel gradient-weighted loss function to learn the complex contours of camouflaged objects. We have conducted extensive experiments to evaluate the performance of the proposed model against state-of-the-art methods from relevant fields. Our results demonstrate the superiority of the proposed model for COD.

Our method does have limitations. As shown in Figure 8, if a camouflaged object has very complex contours due to occlusion, our method may not be able to predict it well. As a future work, we would like to consider incorporating techniques such as object de-occlusion to address the camouflaged object detection problem. In addition, COD is potentially useful for night-time semantic segmentation [42] since many objects are camouflaged in under-exposed regions, even with semi-supervised [11] and noisy-label [50] settings.

Fig. 8. A failure case from COD10K (with a black cat hidden behind the leaves). Our method may fail to detect the camouflaged object with extremely complex contours caused by occlusion.

ACKNOWLEDGMENTS

This work is partially supported by a GRF (RGC Ref: 11205620) from the Research Grants Council of Hong Kong. Xin Tan is supported by the Postgraduate Studentship (Mainland Schemes) from City University of Hong Kong.

REFERENCES

[1] Hassan Abdulameer, Jakub Błaszczyk, Tomasz Depta, Adam Kornacki, and Przemysław Kozieł. 2018. Animal camouflage analysis: Chameleon database.

[2] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. 2019. Hybrid task cascade for instance segmentation. In *CVPR*.

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

[4] Yao Chen, Xu Qianqian, Cong Runmin, and Huang. Qingming. 2020. Global Context-Aware Progressive Aggregation Network for Salient Object Detection. In *AAAI*.

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

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

[7] Max Ehrlich and Larry S Davis. 2019. Deep residual learning in the jpeg transform domain. In *ICCV*. 3484–3493.

[8] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. 2017. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*.

[9] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. 2018. Enhanced-alignment Measure for Binary Foreground Map Evaluation. In *IJCAI*.

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

[11] Zhengyang Feng, Qianyu Zhou, Qiqi Gu, Xin Tan, Guangliang Cheng, Xuequan Lu, Jianping Shi, and Lizhuang Ma. 2022. DMT: Dynamic Mutual Training for Semi-Supervised Learning. *Pattern Recognition* (2022). https://doi.org/10.1016/j.patcog.2022.108777

[12] Shiming Ge, Xin Jin, Qiting Ye, Zhao Luo, and Qiang Li. 2018. Image editing by object-aware optimal boundary searching and mixed-domain composition. *Computational Visual Media* 4, 1 (2018), 71–82.

[13] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A. Wichmann, and Wieland Brendel. 2019. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.. In *ICLR*. https://openreview.net/forum?id=Bygh9j09KX

[14] Huankang Guan, Jiaying Lin, and Rynson W.H. Lau. 2022. Learning Semantic Associations for Mirror Detection. In *CVPR*.

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

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *ICCV*. 1026–1034.

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

[18] Jie Hu, Li Shen, and Gang Sun. 2018. Squeeze-and-excitation networks. In *CVPR*.

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

[20] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. 2019. Anabranch Network for Camouflaged Object Segmentation. *CVIU* 184 (2019), 45–56.

[21] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. 2021. Uncertainty-Aware Joint Salient Object and Camouflaged Object Detection. In *CVPR*.

[22] Yu Li and Michael S Brown. 2014. Single image layer separation using relative smoothness. In *CVPR*.

[23] Jiaying Lin, Zebang He, and Rynson W.H. Lau. 2021. Rich Context Aggregation with Reflection Prior for Glass Surface Detection. In *CVPR*.

[24] Jiaying Lin, Guodong Wang, and Rynson W.H. Lau. 2020. Progressive Mirror Detection. In *CVPR*.

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

[26] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. 2019. A Simple Pooling-Based Design for Real-Time Salient Object Detection. In *CVPR*.

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

[28] Songtao Liu, Di Huang, et al. 2018. Receptive field block net for accurate and fast object detection. In *ECCV*. 385–400.

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

[30] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. 2021. Swin Transformer: Hierarchical Vision Transformer using Shifted Windows. *ICCV* (2021).

[31] Wenjie Luo, Yujia Li, Raquel Urtasun, and Richard Zemel. 2016. Understanding the effective receptive field in deep convolutional neural networks. In *NeurIPS*.

[32] Yunqiu Lyu, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. 2021. Simultaneously Localize, Segment and Rank the Camouflaged Objects. In *CVPR*.
[33] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. 2014. How to evaluate foreground maps?. In *CVPR*.
[34] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. 2021. Camouflaged Object Segmentation with Distraction Mining. In *CVPR*.
[35] Haiyang Mei, Xin Yang, Yang Wang, Yuanyuan Liu, Shengfeng He, Qiang Zhang, Xiaopeng Wei, and Rynson W.H. Lau. 2020. Don't Hit Me! Glass Detection in Real-World Scenes. In *CVPR*.
[36] Yuxin Pan, Yiwang Chen, Qiang Fu, Ping Zhang, and Xin Xu. 2011. Study on the camouflaged target detection method based on 3D convexity. *Modern Applied Science* 5, 4 (2011), 152.
[37] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. 2020. Multi-scale Interactive Network for Salient Object Detection. In *CVPR*.
[38] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. 2019. BASNet: Boundary-Aware Salient Object Detection. In *CVPR*.
[39] Zequn Qin, Pengyi Zhang, Fei Wu, and Xi Li. 2020. FcaNet: Frequency Channel Attention Networks. *arXiv:2012.11879* (2020).
[40] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong. Tian. 2019. Selectivity or Invariance: Boundary-aware Salient Object Detection. In *ICCV*.
[41] Xin Tan, Jiaying Lin, Ke Xu, Chen Pan, Lizhuang Ma, and Rynson W. H. Lau. 2022. Mirror Detection with the Visual Chirality Cue. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022). https://doi.org/10.1109/TPAMI.2022.3181030
[42] Xin Tan, Ke Xu, Ying Cao, Yiheng Zhang, Lizhuang Ma, and Rynson W. H. Lau. 2021. Night-time Scene Parsing with a Large Real Dataset. *IEEE Transactions on Image Processing* 30 (2021), 9085–9098.
[43] Tiantian Wang, Ali Borji, Lihe Zhang, Pingping Zhang, and Huchuan Lu. 2017. A stagewise refinement model for detecting salient objects in images. In *ICCV*.
[44] Jun Wei, Shuhui Wang, and Qingming Huang. 2020. F3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*.
[45] Zhe Wu, Li Su, and Qingming Huang. 2019. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection. In *CVPR*.
[46] Zhe Wu, Li Su, and Qingming Huang. 2019. Stacked Cross Refinement Network for Edge-Aware Salient Object Detection. In *ICCV*.
[47] Yingyue Xu, Dan Xu, Xiaopeng Hong, Wanli Ouyang, Rongrong Ji, Min Xu, and Guoying Zhao. 2019. Structured Modeling of Joint Deep Feature and Prediction Refinement for Salient Object Detection. In *ICCV*.
[48] Feng Xue, Chengxi Yong, Shan Xu, Hao Dong, Yuetong Luo, and Wei Jia. 2016. Camouflage performance analysis and evaluation framework based on features fusion. *Multimedia Tools and Applications* 75, 7 (2016), 4065–4082.
[49] Chuan Yang, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. 2013. Saliency detection via graph-based manifold ranking. In *CVPR*.
[50] Shuquan Ye, Dongdong Chen, Songfang Han, and Jing Liao. 2021. Learning with Noisy Labels for Robust Point Cloud Segmentation. *ICCV* (2021).
[51] Jianqin Yin, Yanbin Han, Wendi Hou, and Jinping Li. 2011. Detection of the Mobile Object with Camouflage Color Under Dynamic Background Based on Optical Flow. *Procedia Engineering* 15 (2011), 2201 – 2205. https://doi.org/10.1016/j.proeng.2011.08.412 CEIS 2011.
[52] Qing Zhang, Gelin Yin, Yongwei Nie, and Wei-Shi Zheng. 2020. Deep Camouflage Images. In *AAAI*.
[53] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. 2018. Progressive attention guided recurrent network for salient object detection. In *CVPR*.
[54] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. 2017. Pyramid scene parsing network. In *CVPR*.
[55] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. 2019. EGNet:Edge Guidance Network for Salient Object Detection. In *ICCV*.
[56] Ting Zhao and Xiangqian Wu. 2019. Pyramid feature attention network for saliency detection. In *ICCV*. 3085–3094.
[57] Ting Zhao and Xiangqian Wu. 2019. Pyramid Feature Attention Network for Saliency detection. In *CVPR*.
[58] Xiaoqi Zhao, Youwei Pang, Lihe Zhang, Huchuan Lu, and Lei Zhang. 2020. Suppress and Balance: A Simple Gated Network for Salient Object Detection. In *ECCV*.
[59] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. 2019. UNet++: Redesigning Skip Connections to Exploit Multiscale Features in Image Segmentation. *IEEE TMI* (2019).
[60] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. 2014. Saliency optimization from robust background detection. In *CVPR*.
[61] Xueyan Zou, Fanyi Xiao, Zhiding Yu, and Yong Jae Lee. 2020. Delving Deeper into Anti-aliasing in ConvNets. In *BMVC*.