

# High-resolution Iterative Feedback Network for Camouflaged Object Detection

Xiaobin Hu, Shuo Wang, Xuebin Qin, Hang Dai, Wenqi Ren, Ying Tai, Chengjie Wang, Ling Shao

arXiv:2203.11624v2 [cs.CV] 3 Feb 2023

**Abstract**—Spotting camouflaged objects that are visually assimilated into the background is tricky for both object detection algorithms and humans who are usually confused or cheated by the perfectly intrinsic similarities between the foreground objects and the background surroundings. To tackle this challenge, we aim to extract the high-resolution texture details to avoid the detail degradation that causes blurred vision in edges and boundaries. We introduce a novel HitNet to refine the low-resolution representations by high-resolution features in an iterative feedback manner, essentially a global loop-based connection among the multi-scale resolutions. In addition, an iterative feedback loss is proposed to impose more constraints on each feedback connection. Extensive experiments on four challenging datasets demonstrate that our HitNet breaks the performance bottleneck and achieves significant improvements compared with 29 state-of-the-art methods. To address the data scarcity in camouflaged scenarios, we provide an application example by employing the cross-domain learning to extract the features that can reflect the camouflaged object properties and embed the features into salient objects, thereby generating more camouflaged training samples from the diverse salient object datasets. The code will be available at: <https://github.com/HUxiaobin/HitNet>.

**Index Terms**—Camouflaged objects, High-resolution texture, Iterative feedback manner.

## I. INTRODUCTION

Camouflaged object detection (COD) is a bio-inspired research area to detect hidden objects or animals that blend with their surroundings [10]. From biological and psychological studies [5], [44], the camouflage skill helps some animals prevent being the prey of their predators, and it also can cheat the human perception system that is sensitive to the coloration and the illumination around the edges. The camouflaged studies not only provide an effective way to deeply understand human perception system, but also benefit a wide range of downstream applications, such as medical image segmentation [7], [13], [14], artistic creation [4], species discovery [39], and crack inspection [15].

In the last two decades, a growing interest is witnessed in developing algorithms capable of seeing targets through camouflage. Early methods aim to utilize the handcrafted low-level features (*e.g.*, texture and contrast [26], 3D convexity [38] and motion boundary [23]). These features still suffer from the limited capability of discriminating the foreground and the

Xiaobin Hu, Ying Tai and Chengjie Wang are with Tencent YouTu Lab, Shanghai, China.

Shuo Wang is with CVL, ETH Zurich, Zurich, Switzerland.

Hang Dai, Xuebin Qin and Ling Shao are with Mohamed bin Zayed University of Artificial Intelligence Abu Dhabi, United Arab Emirates.

Wenqi Ren is with the School of Cyber Science and Technology, Sun Yat-sen University at Shenzhen, Shenzhen 518107, China

Corresponding author: Hang Dai (hang.dai@mbzuae.ac.ae).

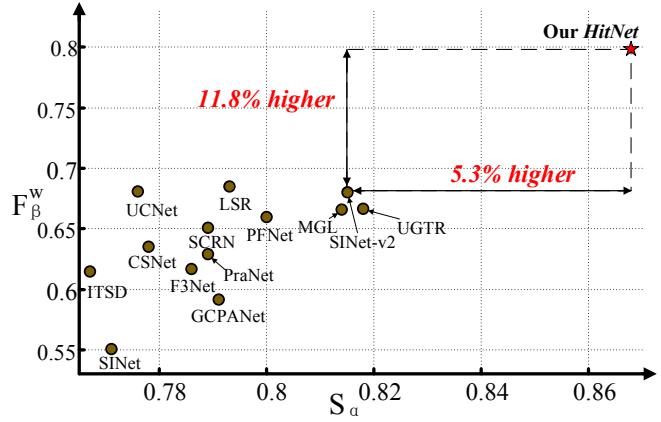


Fig. 1. Weighted F-measure ( $F_\beta^w$ ) vs. Structure-measure ( $S_\alpha$ ) of top 13 models from 29 SOTA methods and our *HitNet* on COD10k-Test dataset.  $F_\beta^w$  is a comprehensive metric to evaluate the weighted precision and recall of the prediction map, and  $S_\alpha$  aims to analyze the structural information of the prediction map. Our framework achieves a remarkable performance milestone.

background in complex scenes. Recently, some CNNs-based frameworks have been proposed to analyze the visual similarities around boundaries between the camouflaged objects and their surroundings. The auxiliary information is extracted from the shared context as the boundary guidance for COD, such as features for identification [12], classification [27], boundary detection [58] and uncertainties [55].

Although the approaches mentioned above have improved the performance, most methods discard the high-resolution details, including edges or textures, by down-sampling the high-resolution images. Fig. 2 shows an interesting phenomenon by evaluating the low-resolution (LR) and the high-resolution (HR) images on the same model well-trained on LR images, respectively. The segmentation result from HR has more details like cat beards than that from LR. This implies that the high-resolution priors are crucial to the boundary and edge detection [48], [61]. The degradation of inputs from HR to LR leads to blurry vision without capturing fine structures. To balance the trade-off between computational resources and performance, the down-sampling operation on high-resolution input is acceptable to achieve satisfactory performance to some extent. But the lose of edge details is not desirable in segmentation tasks, especially for camouflaged object detection. We find that two main aspects account for the degradation: 1) the lack of high-resolution information from input images; 2) the absence of an effective mechanism to enhance the low-

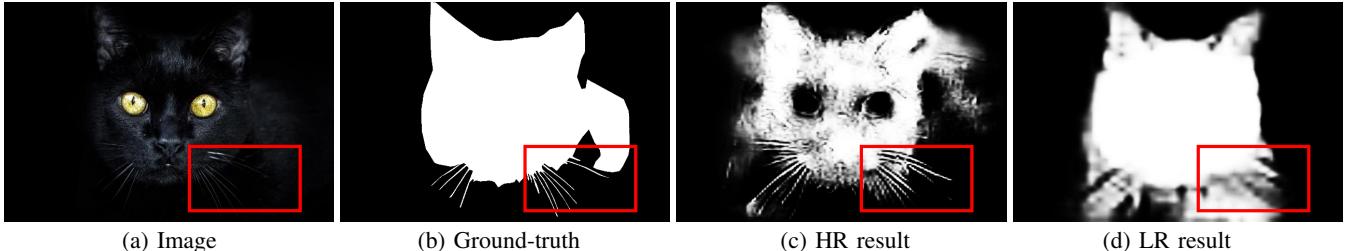


Fig. 2. Results from high-resolution (HR) and low-resolution (LR) inputs with SINet [12] trained on LR images. LR result has blurred edges (*e.g.*, cat's beard), which indicates the details loss (*e.g.*, boundaries) during image degradation process from HR to LR.

resolution features. Thus, it is promising to explore how to maintain the high-resolution information at input level and enhance the low-resolution features without sacrificing the real-time property.

To achieve this goal, we propose a High-resolution Iterative Feedback Network (**HitNet**) to sufficiently and comprehensively exploit the multi-scale HR information and refine the LR with HR knowledge via an adaptively iterative feedback approach. Specifically, HitNet includes three main components: Transformer-based Feature Extraction (TFE), Multi-resolution iterative refinement (RIR), and Iteration feature feedback (IFF). To reduce the computational cost for the HR feature maps, we adopt pyramid vision transformer [49] as an image feature encoder through a progressive shrinking pyramid and spatial-reduction attention. Then, we utilize the RIR module to recursively refine the LR feature extracted from TFE via a global and cross-scale feedback strategy. To ensure the better aggregation of feedback feature, we use iteration feature feedback (IFF) to impose constraints on feedback feature flow.

In addition, we implement an application that converts the salient objects to camouflaged objects via our cross-domain learning strategy. Results from our application can be used as additional training data for the COD task without increasing the parameters and computations of deep learning models in the inference stage.

Our main contributions are summarized as:

- We propose a novel recursive operation to refine the low-resolution feature via a cross-scale feedback mechanism. The recursive operation is simple and can be easily extended to existing COD models.
- Based on the recursive operation, we design a novel framework, termed as High-resolution Iterative Feedback Network (**HitNet**) for COD task. The corresponding iterative feedback loss with an iteration weight scheme is also proposed for **HitNet** to penalize the output of each iteration.
- Our HitNet sets a new record, as shown in Fig. 1, breaking the performance bottleneck, compared with existing cutting-edge models on four benchmarks using four standard metrics. On COD10K, HitNet achieves  $F_\beta^w$  of 0.798, which is **16.5%** higher than the second-best LSR [34]. On CHAMELEON, our HitNet achieves a mean MAE error of 0.018, which is **40.0%** better than the second-best SINet-v2 [10].

## II. RELATED WORK

**Camouflaged Object Detection.** COD aims to spot the camouflaged object from its high-similarity surroundings [12], [37], [43], [45]. It has wide applications [4], [13], [39] and many COD methods [3], [56] have been proposed. These methods can be categorized into two main classes: handcrafted-based and deep-learning-based. More specifically, most of the early works were developed based on the hand-crafted features (*e.g.*, colour and intensity features [26], 3D convexity [38], and motion boundary [23]). But they are relatively less robust and prone to fail in complex scenarios. More studies resort to the powerful representation capacity of deep learning models to detect camouflaged objects in a data-driven way and have achieved impressive improvements against those handcrafted-based methods. On the one hand, deep models usually have many parameters, which ensures stronger representative capabilities for segmenting the camouflaged objects from their backgrounds. On the other hand, most of these deep models benefit by exploring the auxiliary knowledge, *e.g.*, fixations [34], boundaries [58], location [12], image-level labels [27], and uncertainty analysis [55]. Nevertheless, most of models pay much attention to regional accuracy. At the same time, few of them explore the effectiveness of high-frequency information (in high-resolution), which plays a vital role in perceiving the clear boundaries or edges of camouflaged targets. Thus, it impedes the further improvements of COD models. To address this issue, we design a novel High-resolution Iterative Feedback Network, which sets a new record on all benchmarks.

**Iterative Feedback Mechanism of Super-Resolution** allows the network to correct previous states (*i.e.*, lower-resolution) with a higher-level output (*i.e.*, higher-resolution) [24], [57]. In image super-resolution, some studies proved certain improvements after using different feedback mechanisms, such as up-and down-projection units [20] and dual-state recurrent module [19]. However, most of these mechanisms are implemented by using recurrent structures [28] while the information flows from the LR to HR images are still feed-forward. Recently, Li *et al.* [28] proposed an image super-resolution feedback network to refine LR representation with HR information by outlining the edges and contours while suppressing smooth areas. Inspired by this work, we build our transformer-based high-resolution iterative feedback for COD. Different from Li *et al.* [28], our feedback connection is designed as a global connection other than a local connection [16] and embedded into the multi-scale framework via a feedback

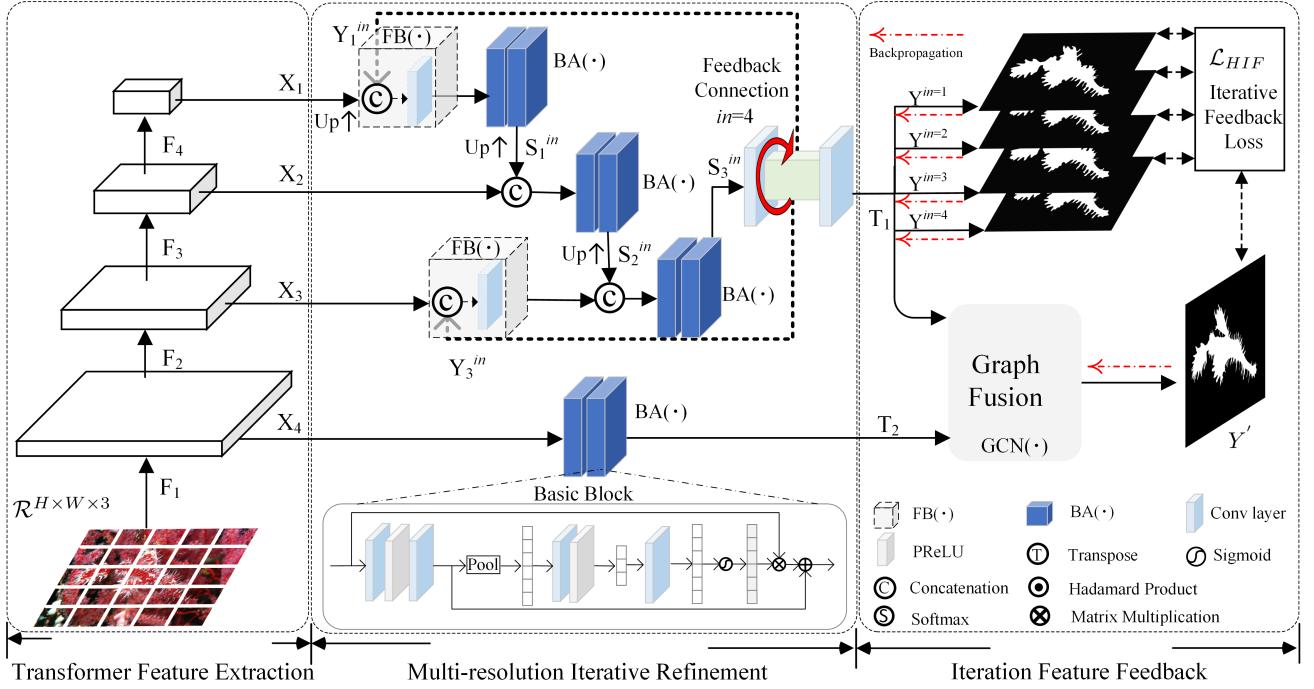


Fig. 3. An overview of Transformer-based High-resolution Iterative Feedback Network (*HitNet*). Our *HitNet* consists of Transformed-based backbone for multi-scale feature extraction, multi-resolution iterative refinement to self-correct low-resolution features with high-resolution information via a cross-resolution iterative feedback mechanism, and iteration feature feedback to impose constraint on each iteration.

fusion block, which merges the information from multi-scale outputs. To avoid the corruption of each iteration, we impose more constraints on each feedback connection by supervising each iteration with the corresponding loss.

**Vision Transformer.** The Transformer was firstly proposed by Vaswani *et al.* [47] as a powerful tool in the domain of machine translation. Considering the superiority of transformer in modeling long-term dependencies, more recent studies have tried to exploit its potentials in different vision tasks, such as image classification [8], [42], semantic segmentation [66], object detection [6], and other low-level tasks [54]. Those works verify the image data can be learned in a sequence-to-sequence approach. Unlike the convolution layer, the layer of multi-head self-attention in the Transformer has dynamic weights and a global receptive field, making the Transformer more effective and powerful in catching non-local knowledge. However, the Transformer suffers high computational and memory costs, especially for HR inputs. Thus, we adopt a Pyramid Vision Transformer (PVT) [49] that uses a progressive shrinking pyramid structure to reduce the sequence length and a spatial-reduction attention layer to decrease the computation further when learning HR features.

### III. PROPOSED METHOD

**Motivation.** Our motivation stems from the observation of degradation phenomenon, shown in Fig. 2, HR inputs generate more accurate predictions than LR inputs, especially for object boundaries. Thus, we aim to explore the feature interaction between high- and low-resolution for COD.

**Overview.** To achieve this goal, we design *HitNet*, consisting of three main modules: Transformer-based Feature Extraction

(Sec. III-A), Multi-resolution iterative refinement (Sec. III-B), and Iteration Feature Feedback (Sec. III-C).

#### A. Transformer-based Feature Extraction

Recently, the transformer [47], which was originally developed for NLP tasks, are proved to be very competitive in many vision tasks against the existing CNN-based backbones. However, many of the vision transformers are GPU memory exhaustive and our HR features will further exaggerate the problem. To alleviate this issue, we choose the Pyramid Vision Transformer (PVT) [49] as our feature extraction module, which can extract multi-scale features, and handle relatively higher resolution feature maps with less memory costs by its progressive shrinking strategy and spatial reduction attention mechanism.

**Progressive Shrinking Strategy:** Denoting the patch size of the  $i$ -th stage as  $P_i$ . Input feature  $F_{i-1} \in \mathcal{R}^{H_{i-1} \times W_{i-1} \times C_{i-1}}$  will be evenly split into  $\frac{H_{i-1} \times W_{i-1}}{P_i \times P_i}$  patches. After the linear projection, the height and width are  $\frac{1}{P_i}$  of the input features.

**Spatial Reduction Attention:** PVT employs spatial reduction attention (SRA) to decrease the computational cost. More specifically, after receiving a query  $Q$ , a key  $K$ , and a value  $V$ , SRA reduces the spatial dimension of  $K$  and  $Q$  before the attentive operation. Given a input sequence  $x \in \mathcal{R}^{H_i W_i \times C_i}$  at  $i$ -th stage, a reduction ratio of spatial dimension  $R_i$  is used to reshape the input  $x$  to the size  $\frac{H_i W_i}{R_i^2} \times (R_i^2 C_i)$ . Then, a linear projection is adopted to further reduce the dimension from input sequence (*i.e.*,  $R_i^2 C_i$ ) to  $C_i$ . Consequently, the computational cost of SRA only occupies  $\frac{1}{R_i^2}$  of the standard Multi-head attention layer [47].

**Multi-scale Feature Extraction:** PVT consists of four stages, and each stage includes a patch embedding and an encoder structure. The input features to each stage ( $\mathbf{F}_i$ ) are first divided into patches with size of  $P_i$ . After that, these features are fed into Transformer encoder structure to get the output features  $X_i$  for the  $i$ -th. Then, we get the multi-scale features ( $\mathbf{X}_1$ ,  $\mathbf{X}_2$ ,  $\mathbf{X}_3$ ,  $\mathbf{X}_4$ ) with  $(\frac{1}{32}, \frac{1}{16}, \frac{1}{8}, \frac{1}{4})$  resolution of input images for further processing.

### B. multi-resolution feedback refinement

The multi-scale resolution features  $\mathbf{X}$  extracted from the Transformer backbone are fed to a basic block  $\text{BA}(\cdot)$  [62] as shown in Fig. 3:

$$\text{BA}(\mathbf{X}_i) = C_2(\mathbf{X}_i) + C_b(C_2(\mathbf{X}_i)) \cdot \mathbf{X}_i, \quad (1)$$

where  $\mathbf{X}_i$  is the input feature of  $i$ -th scale produced by Transformer module,  $C_2(\cdot)$  indicates two stacked convolutional layers with  $3 \times 3$  filters.  $C_b(\cdot)$  denotes the channel attention function [62].

**Iterative Feedback Mechanism** is critical in this module to achieve high accuracy around the object boundary (see Fig. 9). The setting iterative number  $in=1$  assumes the first iteration and no feedback feature transported from previous state. Thus, the  $\mathbf{Y}_1^{in}$  and  $\mathbf{Y}_1^{in}$  are the initial value (0) when  $in=1$ . For iterative number ( $in > 1$ ), the feedback features are produced by previous iteration and then passed into feedback block  $\text{FB}(\cdot)$  as:

$$\text{FB}(\mathbf{X}_i + \mathbf{Y}_i^{in}) = \text{Sq}(\text{Concat}(\mathbf{X}_i \uparrow, \mathbf{Y}_i^{in})), \quad (2)$$

where  $\mathbf{Y}_i^{in}$  is the feedback features of  $in$ -th iteration at  $i$ -th scale ( $i \neq 2$ ), Symbol  $\uparrow$  is the up-sampling operation from the size of  $\mathbf{X}_i$  to  $\mathbf{Y}_i^{in}$  to avoid degradation of the HR information.  $\text{Concat}(\cdot)$  indicates the channel-based concatenation operation between  $\mathbf{X}_i \uparrow$  and  $\mathbf{Y}_i^{in}$ , and  $\text{Sq}(\cdot)$  is size and channel compression using the convolution layer with large kernel and stride<sup>1</sup> to obtain identical size for  $i$ -th scale.

As shown in Fig. 3, with the prerequisite that the iterative number ( $in > 1$ ), the first scale structure receives  $\mathbf{X}_1$  and  $\mathbf{Y}_1^{in}$  and the output the feature can be defined as:

$$\mathbf{S}_1^{in} = \text{BA}(\text{FB}(\mathbf{X}_1 + \mathbf{Y}_1^{in})). \quad (3)$$

Then,  $\mathbf{S}_1^{in}$  is further fed into the next scale to generate next output feature as follows:

$$\mathbf{S}_2^{in} = \text{BA}(\text{Concat}(\mathbf{S}_1^{in} \uparrow, \mathbf{X}_2)), \quad (4)$$

Finally, the features of the previous scale are transported to the next scale as:

$$\mathbf{S}_3^{in} = \text{BA}(\text{Concat}(\mathbf{S}_2^{in} \uparrow, \text{FB}(\mathbf{X}_3 + \mathbf{Y}_3^{in}))). \quad (5)$$

After ending at  $in$ -th iteration, ( $in + 1$ )-th iteration starts from the first scale to the last scale in the same way. The design intuitions on different scales are mainly motivated to get a better cross-scale data flow. The feedback features are explicitly imported into the top and third top scales for the data flow. As the data flow works, the second-top scale can

<sup>1</sup> If  $i=1$ , kernel = 8 with stride = 4 while  $i=3$ , kernel = 1 with stride = 1.

get the implicit feedback features from the top scale. From our experiments, this setting can decrease the computational cost but maintaining good performance. Our HitNet breaks the performance bottleneck due to the following three indispensable mechanisms:

- In each iteration, it outputs an intermediate HR segmentation map that is supervised with a segmentation loss, enabling the feedback features to learn HR cues.
- The HR feedback features merge with inputs features in a feedback block, alleviating the degradation of HR information.
- It uses a feedback fusion mechanism to exploit the HR data flow in a multi-scale structure.

### C. Iteration Feature Feedback

To tailor satisfactory feedback feature flow and avoid the feature corruption caused by recurrent path, we present iteration feature feedback strategy to tie the each feedback feature with the segmentation ground-truth. Intuitively, the data flow of feedback features can be controlled by the loss function. Our basic loss function is defined as  $\mathcal{L} = \mathcal{L}_{IoU}^w + \mathcal{L}_{BCE}^w$ , where  $\mathcal{L}_{IoU}^w$  is the weighted intersection-over-union (IoU) loss and  $\mathcal{L}_{BCE}^w$  denotes the weighted binary cross entropy (BCE) loss. Unlike other recurrent structures [50], we compute the HR prediction loss in each iteration and use an iteration-weight scheme to penalize the output of each iteration when predicting a HR segmentation map:

$$\mathcal{L}_{HIF} = \sum_{in}^N (w \cdot in) \mathcal{L}(Y^{in}) + \mathcal{L}(Y'), \quad (6)$$

where  $in$  is the current iteration number,  $N$  is the total iteration number,  $w$  is the weight parameter,  $Y^{in}$  is the output of  $in$ -th iteration,  $Y'$  is the output of graph-based resolution fusion. In this way, our iteration-weight scheme focuses on the features of deeper iterations by assigning higher weights.

In this session, to efficiently integrate the features from the previous module, we introduce the non-local graph fusion (shown in Fig. 3) via a graph fusion module [7] and [46].

$$Y' = \text{GCN}(T_1, T_2), \quad (7)$$

where  $Y'$  is final prediction map,  $T_1$  is the  $Y^{in=4}$ , GCN is the graph fusion module. For more details regarding the graph fusion, we refer the readers to [7] and [46]. .

## IV. EXPERIMENTS

### A. Experimental Settings

**Datasets.** Our experiments are based on four widely-used COD datasets: (1) CHAMELEON [41] collects 76 high-resolution images from the Internet with the label of camouflaged animals. Each image is manually annotated with object-level GT masks. (2) CAMO [27] includes 2500 images with eight categories, and 2000 images of them for training and 500 images for testing. (3) COD10K [12] is the largest collection containing 10,000 images that divided into 10 super-classes and 78 sub-classes from multiple photography websites. (4) NC4K [34] consists of 4,121 images and is commonly used to evaluate the generalization ability of models. Following

TABLE I

QUANTITATIVE RESULTS OF OUR METHOD AND OTHER 29 STATE-OF-THE-ART METHODS ON FOUR BENCHMARK DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN **BOLD**, AND THE SECOND-BEST IS MARKED IN UNDERLINE. OUR *HitNet* OUTPERFORMS THE SECOND-BEST MODEL BY A LARGE MARGIN.

Baseline Models	CHAMELEON [41]				CAMO-Test [27]				COD10K-Test [12]				NC4K [34]			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
2017 MaskRCNN [21]	0.643	0.778	0.518	0.099	<u>0.574</u>	0.715	0.430	0.151	0.613	0.748	0.402	0.080	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2017 FPN [29]	0.794	0.783	0.590	0.075	0.684	0.677	0.483	0.131	0.697	0.691	0.411	0.075	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2017 NLDF [33]	0.798	0.809	0.714	0.063	0.665	0.664	0.564	0.123	0.701	0.709	0.539	0.059	0.738	0.748	0.657	0.083
2017 PSPNet [63]	0.773	0.758	0.555	0.085	0.663	0.659	0.455	0.139	0.678	0.680	0.377	0.080	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2018 UNet++ [68]	0.695	0.762	0.501	0.094	0.599	0.653	0.392	0.149	0.623	0.672	0.350	0.086	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2018 PiCANet [31]	0.769	0.749	0.536	0.085	0.609	0.584	0.356	0.156	0.649	0.643	0.322	0.090	0.758	0.773	0.639	0.088
2019 MSRCNN [25]	0.637	0.686	0.443	0.091	0.617	0.669	0.454	0.133	0.641	0.706	0.419	0.073	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2019 PoolNet [30]	0.776	0.779	0.555	0.081	0.702	0.698	0.494	0.129	0.705	0.713	0.416	0.074	0.785	0.814	0.699	0.073
2019 BASNet [40]	0.687	0.721	0.474	0.118	0.618	0.661	0.413	0.159	0.634	0.678	0.365	0.105	0.698	0.761	0.613	0.094
2019 SCRNet [52]	0.876	0.889	0.787	0.042	0.779	0.796	0.705	0.090	0.789	0.817	0.651	0.047	0.832	0.855	0.759	0.059
2019 PFANet [65]	0.679	0.648	0.378	0.144	0.659	0.622	0.391	0.172	0.636	0.618	0.286	0.128	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2019 CPD [51]	0.853	0.866	0.706	0.052	0.726	0.729	0.550	0.115	0.747	0.770	0.508	0.059	0.790	0.810	0.708	0.071
2019 HTC [11]	0.517	0.489	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.520	0.221	0.088	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2019 EGNet [64]	0.848	0.870	0.702	0.050	0.732	0.768	0.583	0.104	0.737	0.779	0.509	0.056	0.796	0.830	0.718	0.067
2019 ANet-SRM [27]	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	0.682	0.685	0.484	0.126	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2020 CSNet [18]	0.856	0.869	0.766	0.047	0.771	0.795	0.705	0.092	0.778	0.810	0.635	0.047	0.819	0.845	0.748	0.061
2020 MirrorNet [53]	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	0.741	0.804	0.652	0.100	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2020 PraNet [13]	0.860	0.898	0.763	0.044	0.769	0.833	0.663	0.094	0.789	0.839	0.629	0.045	0.822	0.876	0.724	0.059
2020 SINet [12]	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051	0.810	0.873	0.772	0.057
2020 F3Net [50]	0.854	0.899	0.749	0.045	0.779	0.840	0.666	0.091	0.786	0.832	0.617	0.046	0.782	0.825	0.706	0.069
2020 UCNet [59]	0.880	0.930	0.836	0.036	0.739	0.787	0.700	0.094	0.776	0.857	0.681	0.042	0.813	0.872	0.777	0.055
2020 ITSD [67]	0.814	0.844	0.705	0.057	0.750	0.779	0.663	0.102	0.767	0.808	0.615	0.051	0.811	0.845	0.729	0.064
2020 SSAL [60]	0.757	0.849	0.702	0.071	0.644	0.721	0.579	0.126	0.668	0.768	0.527	0.066	0.699	0.778	0.647	0.092
2020 GCPANet [2]	0.876	0.891	0.748	0.041	0.778	0.842	0.646	0.092	0.791	0.799	0.592	0.045	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2021 MGL [58]	0.893	0.923	0.813	0.030	0.775	0.847	0.673	0.088	0.814	0.865	0.666	0.035	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2021 PFNet [36]	0.882	0.942	0.810	0.033	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040	0.829	0.887	0.745	0.053
2021 UGTR [55]	0.888	0.918	0.796	0.031	0.785	0.859	0.686	0.086	0.818	0.850	0.667	0.035	<u>‡</u>	<u>‡</u>	<u>‡</u>	<u>‡</u>
2021 LSR [34]	0.893	0.938	0.839	0.033	0.793	0.826	0.725	0.085	0.793	0.868	0.685	0.041	0.839	0.883	0.779	0.053
2021 SINet-V2 [10]	0.888	0.942	0.816	0.030	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037	0.847	0.903	0.769	0.048
2022 HitNet (Ours)	<b>0.922</b>	<b>0.970</b>	<b>0.903</b>	<b>0.018</b>	<b>0.844</b>	<b>0.902</b>	<b>0.801</b>	<b>0.057</b>	<b>0.868</b>	<b>0.932</b>	<b>0.798</b>	<b>0.024</b>	<b>0.870</b>	<b>0.921</b>	<b>0.825</b>	<b>0.039</b>

previous studies [12], [34], [55], [58], the combined training set from CAMO and COD10K is used as the training set, and others are test sets.

**Metrics.** Following [12], [27], [58], four standard metrics are used to comprehensively evaluate the model performance: mean absolute error (MAE), mean E-measure ( $E_\phi$ ) [11], S-measure  $S_\alpha$  [9], and weighted F-measure  $F_\beta^w$  [35].

**Implementation Details.** We implement our model based on PyTorch in AMD Ryzen Threadripper 3990X 2.9GHz CPU and NVIDIA RTX A6000 GPU. For the training stage, the resolution of input images is resized to 704×704, which can avoid the loss of high-resolution information to some extent, and no data augmentation is used in our model. The transformer-based feature extraction is initialized by PVT-V2 [49], and the remaining modules are initialized in a random manner. We employ the AdamW [32] optimizer with the learning rate of 1e-4, which is widely used in transformer structure, and the corresponding decay rate to 0.1 for every 30 epochs. The weight  $w$  of iterative feedback loss is 0.2, and the well-optimized iteration number ( $in$ ) is 4. The total epochs of training are 100 with a batch size of 16. For testing, the images are resized to 704×704 as the network's input, and the outputs are resized back to the original size.

**Competitors.** We compare our HitNet with recent 29 state-of-the-art (SOTA) methods, including the most recent COD, salient object detection, generic object detection, and semantic segmentation methods (Tab. I). For a fair comparison, all results are either provided by the published paper or reproduced by an open-source model re-trained on the same training set with the recommended setting.

### B. Quantitative Evaluation

**CHAMELEON.** As shown in Tab. I, we compare our *HitNet* against 27 SOTA algorithms on four standard metrics. As a performance milestone, compared with second-best models [10], [34], [58], our *HitNet* significantly lowers the MAE error by **40.0%** and improves  $F_\beta^w$  by **7.6%**.

**CAMO.** For CAMO dataset, compared with 29 models, our *HitNet* still dramatically reduces the MAE error by **18.6%** and increases  $F_\beta^w$  by **7.8%** in contrast to second-best [10].

**NC4K.** We evaluate the generalization ability of all models on NC4K dataset. From Tab. I, as the best model compared with second-best models [10], [34], our algorithm reduces the MAE error by **18.7%** and improve  $F_\beta^w$  by **5.9%**.

**COD10K.** Tab. I also compares our *HitNet* with other 27 SOTA models on the most challenging COD10K test set. From the comparisons, our *HitNet* sets a remarkable record to decrease the MAE error by **31.4%** and improve  $F_\beta^w$  by **16.5%** than the second-best models [34], [55], [58]. The performance superiority of *HitNet* on four benchmarks is mainly due to the well-exploited high-resolution information and the mitigating of high-resolution degradation at the feature level via an iterative feedback mechanism.

### C. Qualitative Evaluation

Fig. 4 shows qualitative results of our *HitNet* and other most recent models. The examples are difficult to be segmented even for manual annotation due to their complex topological structures and detailed edges from the first row to the third row. But our *HitNet* is capable of segmenting clear edges

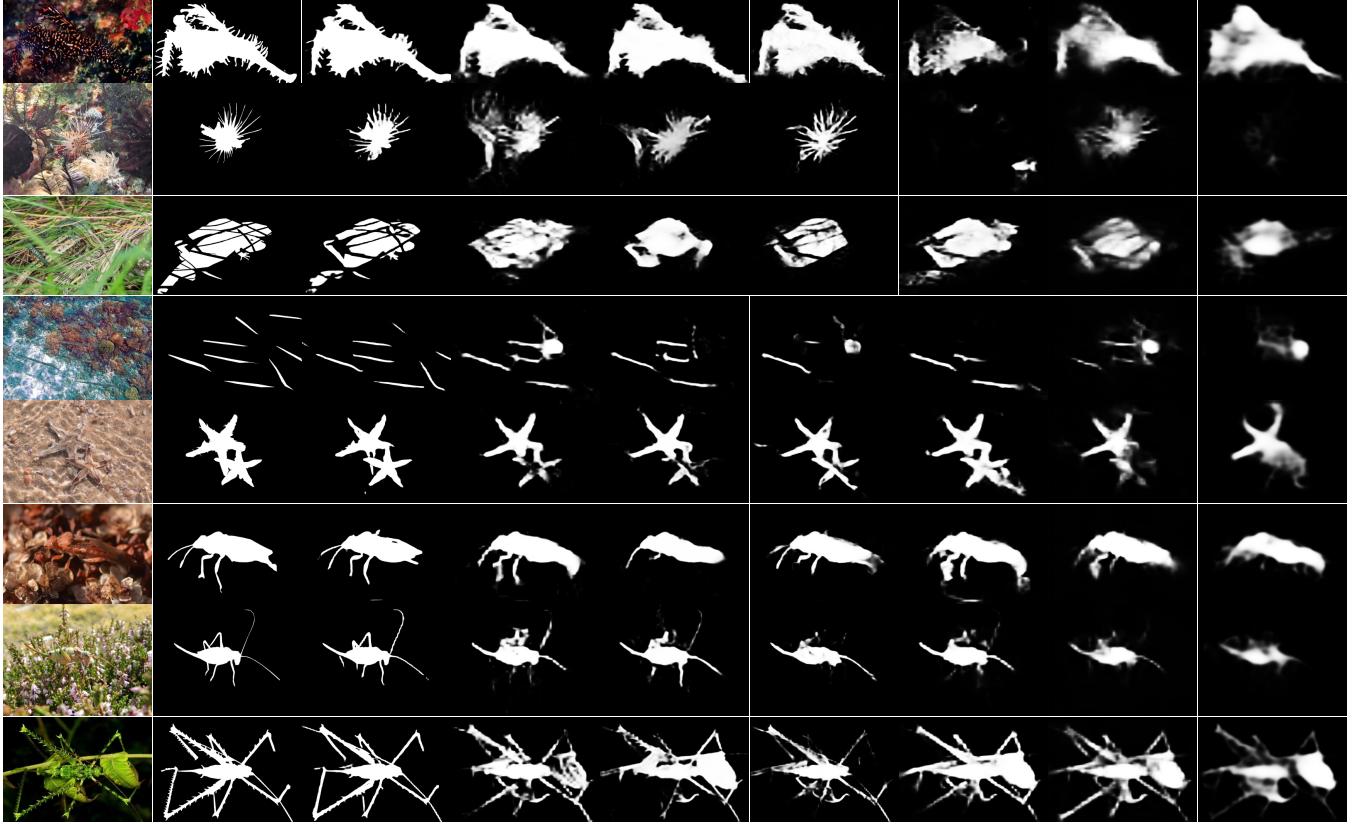


Fig. 4. Visual performance of the proposed *HitNet*. Our algorithm is capable of tackling challenging cases (e.g., complex edges with dense thorn, multiply camouflaged objects, partly occlusion, and global thin edges).

and boundaries (e.g., leaves, thorn) even for objects with occlusion (3-nd row). At the same time, other results are blurred or without correct details. For 4-th and 5-th rows, our HitNet can still clearly segment the multiply camouflaged objects significantly better than others. Similar examples from 6-th to 8-th row also show that our algorithm can segment high-resolution edges and do not ignore thin edges (e.g., the antenna or legs of insects) while other models fail in this kind of hard case. As shown from Fig. 5 to Fig. 8, we add more qualitative comparisons with recent state-of-the-art algorithms (2021 SINet-V2 [10], 2021 PFNet [36], 2021 LSR [34], 2020 PraNet [13], 2020 SINet [12], 2019 EGNet [64]) on four benchmarks (CAMO, CHAMELEON, NC4K, COD10K).

#### D. Ablation Study

**Effectiveness of Each Component.** As shown in Tab. IV, we evaluate the effectiveness of each module by removing the corresponding part from our complete (*i.e.*, TFE + RIR + IFF) HitNet. To assess the contribution of the transformer backbone, we substitute the transformer backbone with Res2Net-50 [17] used in SINet-V2 [10] as the version of ‘w/o TFE’. It is trained and then evaluated on the most challenging COD10K dataset to show its importance. Our algorithm without PVT backbone still achieves the best performance compared with all 29 SOTA methods. But compared with the Res2Net-50 backbone, PVT can achieve better performance due to its superiority of the global receptive field. Besides, we also remove the

RIR module from HitNet expressed as ‘w/o RIR’. The  $F_\beta^w$  performance of this variant sharply deteriorates from 0.798 to 0.703. Lastly, we also replace the Iteration Feature Feedback (IFF) with a convolutional fusion layer (‘w/o IFF’) and find this variant also decreases the performance to some extent. Overall, our RIR module plays a crucial role in performance improvement compared with the other two modules.

**Configuration of Iteration Number.** We explore the effect of iteration number in the iterative feedback mechanism on inference time and performance. As shown in Tab. V, the performance is gradually improved when the iterative number (*in*) increases from 1 to 5. To balance inference time and performance, we choose  $T = 4$  as default in our HitNet. Note that our *HitNet* is a real-time algorithm (39 fps), and it sets a new record that is significantly better than 29 other models on four benchmarks.

**Study of Iterative Feedback Mechanism.** As discussed in §III-B, three components enable the feedback mechanism to boost the performance: 1) Tie each iteration with loss (denoted as ‘Tie’); 2) The Feedback Block to avoid the loss of high-resolution information (denoted as ‘FB’); 3) multi-scale connection fusion (denoted as ‘Multi-fusion’); As shown in Table VII, any absence of three factors will fail the model to drive the data flow. To analyze the difference among the iteration number, we visualize the feature of each iteration in Fig. 9. We observe that the iterative feedback mechanism is a self-correcting process that the subsequent iterations can generate better representations than the previous iteration (e.g.,

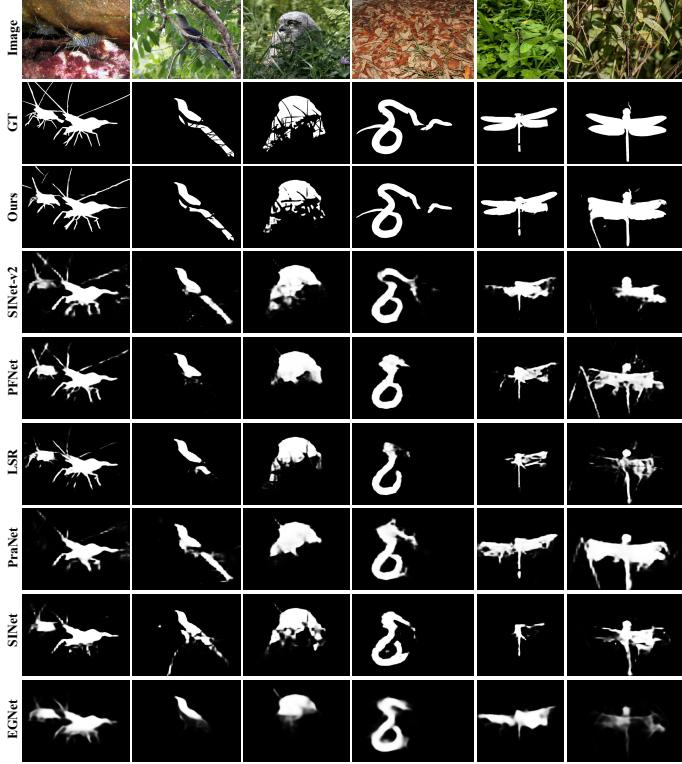


Fig. 5. Visual performance of the proposed HitNet comparison with state-of-the-art methods (SINet-V2 [10], PFNet [36], LSR [34], PraNet [13], SINet [12], EGNNet [64]) on NC4K dataset. From the left to right columns, the names of images are 54, 141, 161, 201, 597, and 601, respectively.

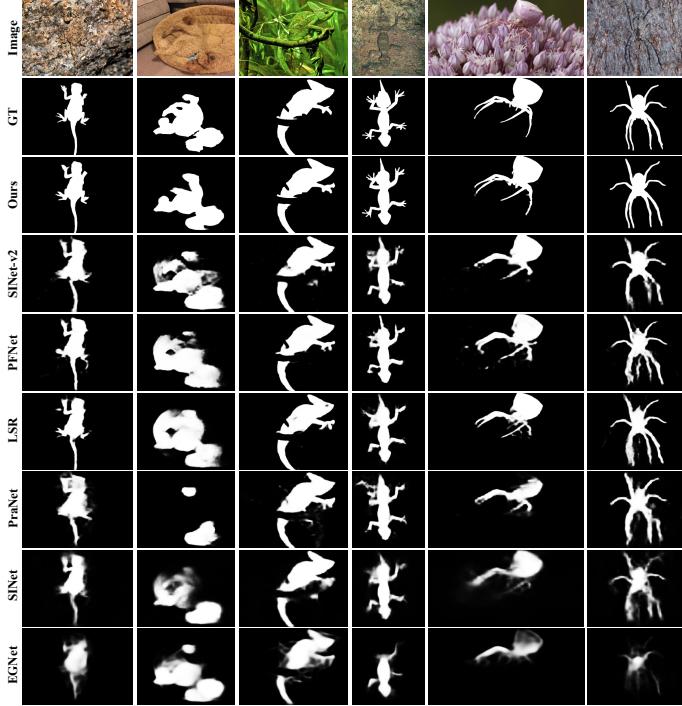


Fig. 6. Visual performance of the proposed HitNet comparison with state-of-the-art methods (SINet-V2 [10], PFNet [36], LSR [34], PraNet [13], SINet [12], EGNNet [64]) on CHAMELEON dataset. From the left to right columns, the names of images are *animal-9*, *animal-7*, *animal-23*, *animal-33*, *animal-72*, and *animal-70*, respectively.

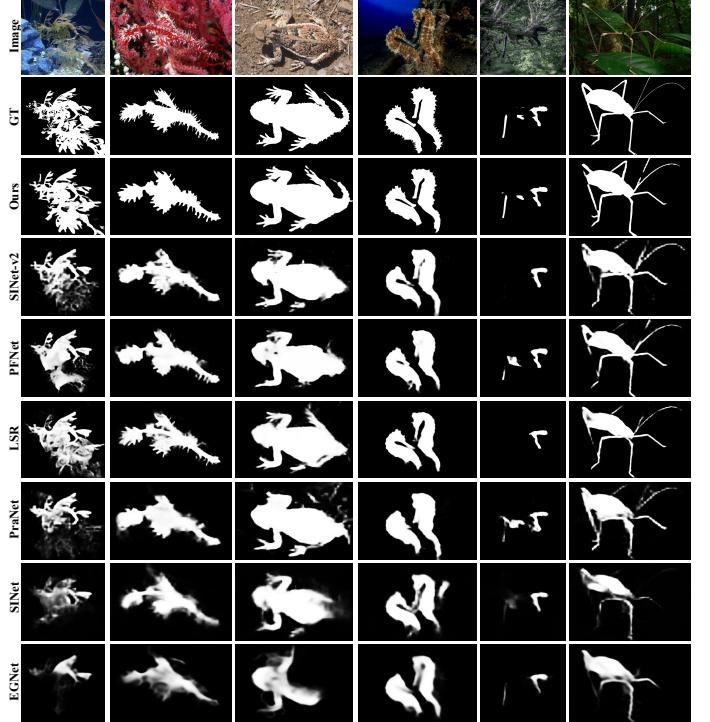


Fig. 7. Visual performance of the proposed HitNet comparison with state-of-the-art methods (SINet-V2 [10], PFNet [36], LSR [34], PraNet [13], SINet [12], EGNNet [64]) on COD10K dataset. From the left to right columns, the names of images are *COD10K-CAM-1-Aquatic-10-LeafySeaDragon-423*, *COD10K-CAM-1-Aquatic-9-GhostPipefish-350*, *COD10K-CAM-2-Terrestrial-38-Lizard-2166*, *COD10K-CAM-1-Aquatic-15-SeaHorse-1086*, *COD10K-CAM-5-Other-69-Other-5048*, and *COD10K-CAM-3-Flying-61-Katydid-4196*, respectively.

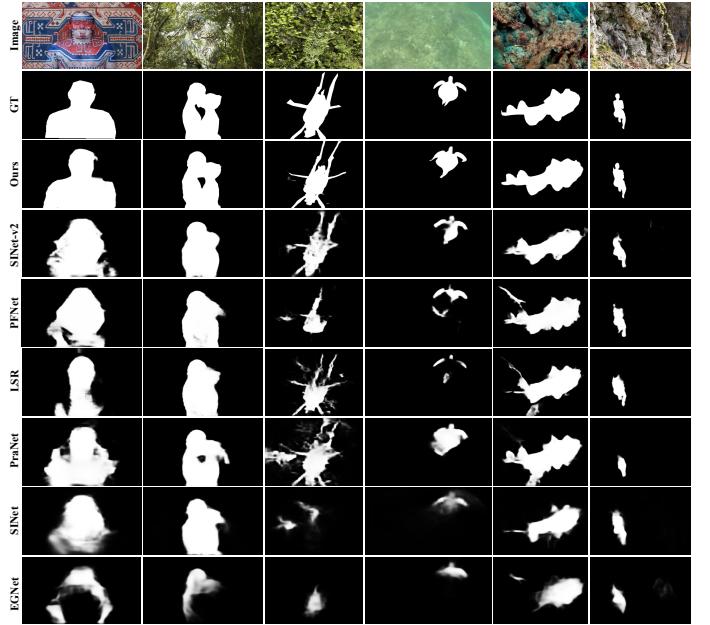


Fig. 8. Visual performance of the proposed HitNet comparison with state-of-the-art methods (SINet-V2 [10], PFNet [36], LSR [34], PraNet [13], SINet [12], EGNNet [64]) on CAMO dataset. From the left to right columns, the names of images are *camouflage-01181*, *camouflage-01240*, *camouflage-00064*, *camouflage-00135*, *camouflage-00449*, and *camouflage-01243*, respectively.

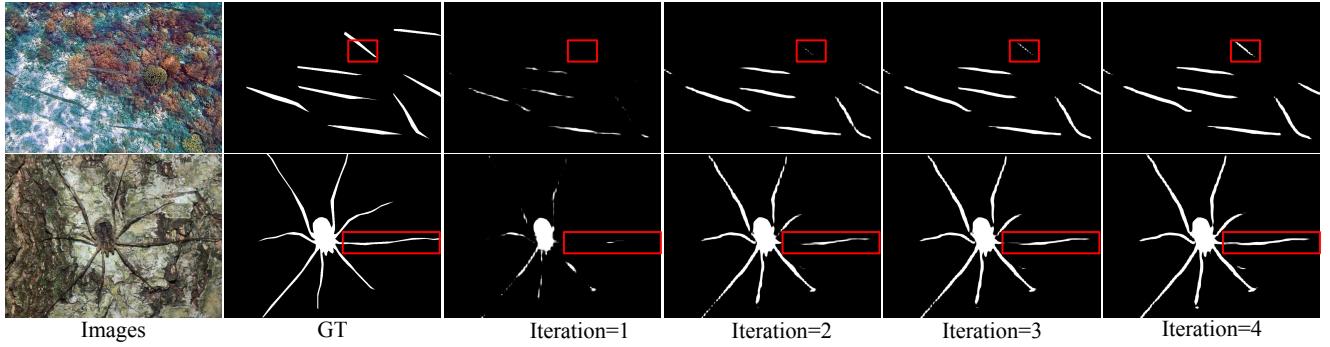


Fig. 9. Visual performance of each iteration in our iterative feedback mechanism in our RIR module (§III-B).

sharper edges).

**Analyses of Input Resolution.** Preventing the loss of HR knowledge (*e.g.*, boundaries or edges) caused by degrading input images is an efficient strategy for COD. But even given same degraded  $352 \times 352$  inputs, as shown in Tab. VIII, our HitNet is still the best algorithms and reduce the MAE error by 21.6% than second-best SINet-V2 [10].

**Evaluation of Different Backbones.** To assess the contribution of the CNN-based and Transformer-based backbones, we substitute the Transformer backbone of HitNet with Res2Net-50 [17] used in SINet-V2 [10] as the version of ‘HitNet+Res2Net-50’. In addition, we also replace Transformer backbone with ResNet-50 [22] used in PFNet [36] as the version of ‘HitNet+ResNet-50’. All models are trained and then evaluated on the most challenging COD10K dataset to show its importance. As shown in Tab. III, compared with 2021 SINet-V2 [10] and 2021 PFNet [36] models with the same backbones as ours, our HitNet achieves superior performance with high quantitative results. Our algorithm (‘HitNet+Res2Net-50’) without Transformer backbone still achieves the best performance compared with all 29 SOTA methods. But compared with the Res2Net-50 and ResNet-50 backbone, Transformer can achieve better performance due to its superiority of the global receptive field.

**Inference time.** For inference time analysis without considering I/O time, the batch size is set as 1 with the image resolution  $704 \times 704$ . The test stags are also implemented on PyTorch in AMD Ryzen Threadripper 3990X 2.9GHz CPU and NVIDIA RTX A6000 GPU. We compared our HitNet with the most recent algorithms in Tab. VI. From this comparison, our HitNet achieves the best performance with  $F_\beta^w$  0.798 (11.3% higher than the second-best 2021 LSR [34]). Meanwhile, our HitNet is also the second-fastest algorithm with the real-time property.

**Computational complexity (CC) of different backbones:** We conduct the GFLOPs analysis on the input resolution  $704 \times 704$  using four variants. As seen in Tab. II, PVT is much lower than the ViT and also has similar GFLOPs CC with a convolutional-based backbone (ResNet50, Res2Net50).

TABLE II  
GFLOPS ANALYSES OF DIFFERENT BACKBONES ON COD10K.

	ResNet50	Res2Net50	ViT	PVT
GFLOPs ↓	48	52	76	54

TABLE III  
QUANTITATIVE RESULTS BASED ON DIFFERENT BACKBONES.

Backbone	Methods	COD10K [12]			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
ResNet-50 [22]	2021 PFNet [36]	0.800	0.868	0.660	0.040
	HitNet+ResNet-50	<b>0.805</b>	<b>0.885</b>	<b>0.688</b>	<b>0.036</b>
Res2Net-50 [17]	2021 SINet-V2 [10]	0.815	0.887	0.680	0.037
	HitNet+Res2Net-50	<b>0.835</b>	<b>0.898</b>	<b>0.735</b>	<b>0.029</b>
PVT [49]	HitNet+PVT	<b>0.868</b>	<b>0.932</b>	<b>0.798</b>	<b>0.024</b>

TABLE IV  
ABLATION ANALYSES OF OUR HITNET ON COD10K DATASET [12].

Metric	w/o TFE	w/o RIR	w/o IFF	HitNet
$F_\beta^w \uparrow$	0.735	0.703	0.791	<b>0.798</b>
$S_\alpha \uparrow$	0.835	0.821	0.863	<b>0.868</b>
$M \downarrow$	0.029	0.034	0.025	<b>0.024</b>
$E_\phi \uparrow$	0.898	0.896	0.926	<b>0.932</b>

TABLE V  
ITERATIVE NUMBER ( $in$ ) ANALYSES ON COD10K DATASET.

	$in=1$	$in=2$	$in=3$	$in=4$	$in=5$
MAE ↓	0.0252	0.0248	0.0249	<b>0.0240</b>	0.0241
Test time (ms) ↓	<b>18.8</b>	21.1	23.4	25.6	27.6

TABLE VI  
INFERENCE SPEED ( $fps$ ) ANALYSES ON COD10K DATASET.

	[12]	[36]	[34]	[10]	Ours
$F_\beta^w \uparrow$	0.551	0.660	0.685	0.680	<b>0.798</b>
Test time ( $fps$ ) ↓	36	<b>40</b>	29	31	39

TABLE VII  
ABLATION STUDY ON INDISPENSABLE FACTORS OF ITERATIVE FEEDBACK MECHANISM ON COD10K DATASET.

Configurations			Performance
Tie	FB	Multi-fusion	MAE ↓
×	×	×	0.0268
✓	✗	✗	0.0255
✓	✓	✗	0.0248
✓	✓	✓	<b>0.0240</b>

## V. APPLICATION

As studied by Fan *et al.* [12], the term “salient” is essentially the opposite of “camouflaged”. We are interested in implementing an application that converts salient objects to camouflage objects. We adopt a cross-domain learning (CDL)

TABLE VIII  
INPUT RESOLUTION ANALYSES OF HITNET ON COD10K DATASET.

Resolution	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
Input: 352×352	0.827	0.907	0.727	0.029
Input: 704×704	<b>0.868</b>	<b>0.932</b>	<b>0.798</b>	<b>0.024</b>

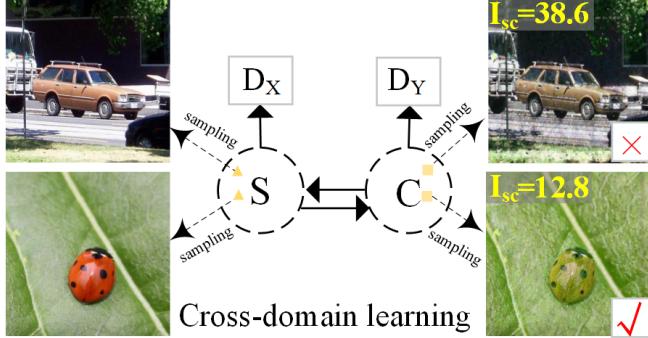


Fig. 10. The overview of salient-to-camouflaged cross-domain learning pipeline. The  $S$  is the salient domain, and the  $C$  means the camouflaged domain.  $D_X$  is the discriminator of the salient domain, and  $D_Y$  is the discriminator of the camouflaged domain.

technique to achieve this goal. In addition, we propose a contrastive index to evaluate the camouflaged level. This index can be acted as the criterion to discard some hard cases with unchangeable intrinsic salient objects.

#### A. Cross-domain Learning

We employ the cycle-consistency structure [69] to learn the camouflaged features and embed these features into the salient objects in an unsupervised cross-domain learning manner as shown in Fig. 10. The cycle-consistency loss can be formulated as:

$$\mathcal{L}_{cyc}(G, F) = \mathbb{E}_x[\|(F(G(x)) - x)\|_1] + \mathbb{E}_y[\|(G(F(y)) - y)\|_1] \quad (8)$$

where  $G$  aims to construct fake images  $\{G(x)\}$  from salient samples  $\{x\}$  to get close to camouflaged domain  $Y$  while  $D(Y)$  tries to distinguish between the translated camouflaged samples  $\{G(x)\}$  and real camouflaged samples  $\{y\}$ .  $F$  is another translator from camouflaged to salient objects. The procedure is concluded as a min-max optimization task<sup>2</sup> in the adversarial loss function used in CycleGAN [69].

To better select the converted camouflaged objects, we propose a contrastive index, considering the pixel-level similarity between object and its surroundings:

$$I_{sc} = \frac{1}{\text{Num}} \sum_i^{\text{Num}} \|P_i - P_m\|_{i \in (P_m - P_{\text{std}}, P_m + P_{\text{std}})}, \quad (9)$$

where  $I_{sc}$  is the index of camouflaged level,  $P_i$  is  $i$ -th pixel intensity value,  $P_m$  is the mean value of images,  $P_{\text{std}}$  is the standard deviation, and  $i$  is the pixel index that belongs to one  $\sigma$  rule to exclude the effect of extreme values. As shown in

<sup>2</sup> Minimize the generator loss while maximized the discriminator loss.



Fig. 11. Failure cases of CDL detected by our contrastive index.



Fig. 12. Application Results of convert salient object (i.e., a & c) to camouflaged object (i.e., b & d).

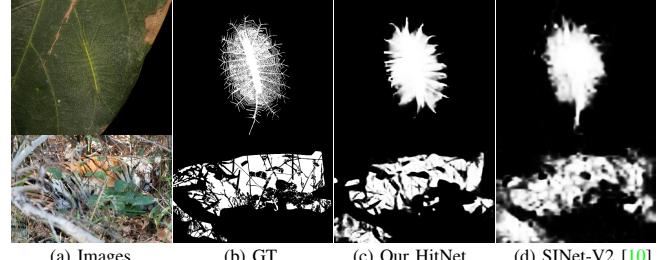


Fig. 13. Failure cases of our HitNet.

TABLE IX  
QUANTITATIVE RESULTS ON DIFFERENT TRAINING STRATEGIES.  
'W/O' MEANS WITHOUT ANY DATA STRATEGY, 'SALIENT DATA'  
MEANS ADDING SALIENT DATA FOR TRAINING.

Data Strategy	COD10K [12]			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
w/o	0.868	0.932	0.798	0.024
Salient data	0.844	0.911	0.782	0.026
<b>CDL (Ours)</b>	<b>0.871</b>	<b>0.935</b>	<b>0.806</b>	<b>0.023</b>

Fig. 10, the car is an abandoned example that can be detected as a high salient case by our contrastive index. Empirically, we set the threshold of  $I_{sc}$  between the camouflaged and salient object as  $I_{sc} = 20$  after plenty of observations. With this index, we can discard some failure converted cases, as shown in Fig. 11.

**Qualitative and Quantitative Evaluation.** As shown in Fig. 12, our Cross-Domain Learning (CDL) method can merge the salient objects into the background as camouflaged objects. As shown in Tab. IX, although there exist numerous salient object data, it fails to boost the performance in camouflaged scenarios directly. Instead, they severely deteriorate  $S_\alpha$  from 0.868 to 0.844 and  $E_\phi$  from 0.932 to 0.911. Meanwhile, the proposed CDL can improve the  $F_\beta^w$  from 0.798 to 0.806 and reduce the MAE error by 4.2%.

## VI. FAILURE CASE

Although our HitNet sets a new record in the COD task, there are still some very challenging examples (*e.g.*, caterpillar) that HitNet fails to address. As shown in Fig. 13, these cases usually contain complicated topological structures with lots of dense edges or details.

**Limitation of Multi-resolution Iterative Refinement.** For some very challenging cases, *e.g.*, one single camouflaged object with complex topological structures, global long-range edges, or the multiply camouflaged objects where one of objects is very small to be neglected, our algorithm still has some space to be improved. The bottleneck of the Iterative Feedback Mechanism mainly causes the potential reasons. This mechanism segments the dense edges or multiplies small objects by self-correcting low-resolution features with high-resolution information. Unfortunately, when the iteration number  $> 4$ , the performance only improves slightly but increases time.

## VII. CONCLUSION

We propose a novel high-resolution iterative feedback network (**HitNet**) to extract the informative and high-resolution representations for tackling the degradation issue of segmentation details on the COD task. HitNet can adaptively refine the low-resolution features with high-resolution information in an iterative feedback manner. More importantly, our approach achieves remarkable performance improvements and significantly outperforms 29 cutting-edge models on four challenging datasets. Finally, we introduce the cross-domain learning strategy to implement an application that converts the salient object to the camouflaged object, potentially enlarging the diversity of the COD dataset.

**Broader Impacts.** As a performance milestone, our HitNet has a great potential to be deployed in real scenarios of camouflaged object detection (*e.g.*, species discovery, helicopter rescue, polyp segmentation). In academia, this work may give some inspiration to explore high-resolution features to facilitate the finer segmentation. Our cross-domain strategy builds a bridge between camouflaged and salient domains, where the camouflaged domain can benefit from the diverse salient dataset to some extent.

## REFERENCES

- [1] Kai Chen, Jiangmiao Pang, Jiaqi Wang, Yu Xiong, Xiaoxiao Li, Shuyang Sun, Wansen Feng, Ziwei Liu, Jianping Shi, Wanli Ouyang, et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019. [5](#)
- [2] Zuyao Chen, Qianqian Xu, Runmin Cong, and Qingming Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, 2020. [5](#)
- [3] Xuelian Cheng, Huan Xiong, Deng-ping Fan, Yiran Zhong, Mehrtash Harandi, Tom Drummond, and Zongyuan Ge. Implicit motion handling for video camouflaged object detection. In *CVPR*, 2022. [2](#)
- [4] Hung-Kuo Chu, Wei-Hsin Hsu, Niloy J Mitra, Daniel Cohen-Or, Tien-Tsin Wong, and Tong-Yee Lee. Camouflage images. *ACM TOG*, 29(4):51–1, 2010. [1, 2](#)
- [5] IC Cuthill. Camouflage. *JOZ*, 308(2):75–92, 2019. [1](#)
- [6] Zhigang Dai, Bolun Cai, Yugeng Lin, and Junying Chen. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*, 2021. [3](#)
- [7] Bo Dong, Wenhai Wang, Deng-Ping Fan, Jinpeng Li, Huazhu Fu, and Ling Shao. Polyp-pvt: Polyp segmentation with pyramidvision transformers. *arXiv preprint arXiv:2108.06932*, 2021. [1, 4](#)
- [8] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. [3](#)
- [9] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A New Way to Evaluate Foreground Maps. In *ICCV*, 2017. [5](#)
- [10] Deng-Ping Fan, Ge-Peng Ji, Ming-Ming Cheng, and Ling Shao. Concealed object detection. *IEEE TPAMI*, 2021. [1, 2, 5, 6, 7, 8, 9](#)
- [11] Deng-Ping Fan, Ge-Peng Ji, Xuebin Qin, and Ming-Ming Cheng. Cognitive vision inspired object segmentation metric and loss function. *SSI*, 6, 2021. [5](#)
- [12] Deng-Ping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *CVPR*, 2020. [1, 2, 4, 5, 6, 7, 8, 9](#)
- [13] Deng-Ping Fan, Ge-Peng Ji, Tao Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Planet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, 2020. [1, 2, 5, 6, 7](#)
- [14] Deng-Ping Fan, Tao Zhou, Ge-Peng Ji, Yi Zhou, Geng Chen, Huazhu Fu, Jianbing Shen, and Ling Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8):2626–2637, 2020. [1](#)
- [15] Fen Fang, Liyuan Li, Ying Gu, Hongyuan Zhu, and Joo-Hwee Lim. A novel hybrid approach for crack detection. *Pattern Recognition*, 107:107474, 2020. [1](#)
- [16] Mengyang Feng, Huchuan Lu, and Errui Ding. Attentive feedback network for boundary-aware salient object detection. In *CVPR*, 2019. [2](#)
- [17] Shanghua Gao, Ming-Ming Cheng, Kai Zhao, Xin-Yu Zhang, Ming-Hsuan Yang, and Philip HS Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(02):652–662, 2021. [6, 8](#)
- [18] Shang-Hua Gao, Yong-Qiang Tan, Ming-Ming Cheng, Chengze Lu, Yunpeng Chen, and Shuiceng Yan. Highly efficient salient object detection with 100k parameters. In *ECCV*, 2020. [5](#)
- [19] Wei Han, Shiyu Chang, Ding Liu, Mo Yu, Michael Witbrock, and Thomas S Huang. Image super-resolution via dual-state recurrent networks. In *CVPR*, 2018. [2](#)
- [20] Muhammad Haris, Gregory Shakhnarovich, and Norimichi Ukita. Deep back-projection networks for super-resolution. In *CVPR*, 2018. [2](#)
- [21] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. In *ICCV*, 2017. [5](#)
- [22] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. [8](#)
- [23] Jianqin Yin Yanbin Han Wendi Hou and Jinping Li. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *PE*, 15:2201–2205, 2011. [1, 2](#)
- [24] Xiaobin Hu, Yanyang Yan, Wenqi Ren, Hongwei Li, Amirhossein Bayat, Yu Zhao, and Bjoern Menze. Feedback graph attention convolutional network for mr images enhancement by exploring self-similarity features. In *MIDL*, 2021. [2](#)
- [25] Zhaojin Huang, Lichao Huang, Yongchao Gong, Chang Huang, and Xinggang Wang. Mask scoring r-cnn. In *CVPR*, 2019. [5](#)
- [26] Iván Huerta, Daniel Rowe, Mikhail Mozerov, and Jordi González. Improving background subtraction based on a casuistry of colour-motion segmentation problems. In *IbPRIA*, 2007. [1, 2](#)
- [27] Trung-Nghia Le, Tam V Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabanch network for camouflaged object segmentation. *CVIU*, 184:45–56, 2019. [1, 2, 4, 5](#)
- [28] Zhen Li, Jinglei Yang, Zheng Liu, Xiaomin Yang, Gwanggil Jeon, and Wei Wu. Feedback network for image super-resolution. In *CVPR*, 2019. [2](#)
- [29] Tsungyi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. [5](#)
- [30] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019. [5](#)
- [31] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018. [5](#)
- [32] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. [5](#)
- [33] Zhiming Luo, Akshaya Mishra, Andrew Achkar, Justin Eichel, Shaozi Li, and Pierre-Marc Jodoin. Non-local deep features for salient object detection. In *CVPR*, 2017. [5](#)
- [34] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*, 2021. [2, 4, 5, 6, 7, 8](#)
- [35] Ran Margolin, Lihhi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, 2014. [5](#)
- [36] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-

- Ping Fan. Camouflaged object segmentation with distraction mining. In *CVPR*, 2021. 5, 6, 7, 8
- [37] Andrew Owens, Connelly Barnes, Alex Flint, Hanumant Singh, and William Freeman. Camouflaging an object from many viewpoints. In *CVPR*, 2014. 2
- [38] Yuxin Pan, Yiwang Chen, Qiang Fu, Ping Zhang, and Xin Xu. Study on the camouflaged target detection method based on 3d convexity. *MAS*, 5(4):152–157, 2011. 1, 2
- [39] Ricardo Pérez-de la Fuente, Xavier Delclòs, Enrique Peñalver, Mariela Speranza, Jacek Wierzchos, Carmen Ascaso, and Michael S Engel. Early evolution and ecology of camouflage in insects. *PNAS*, 109(52):21414–21419, 2012. 1, 2
- [40] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019. 5
- [41] P Skurowski, H Abdulameer, J Błaszczyk, T Depta, A Kornacki, and P Koziel. Animal camouflage analysis: Chameleon database. Unpublished Manuscript, 2018. 4, 5
- [42] Aravind Srinivas, Tsung-Yi Lin, Niki Parmar, Jonathon Shlens, Pieter Abbeel, and Ashish Vaswani. Bottleneck transformers for visual recognition. In *CVPR*, 2021. 3
- [43] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *PTRS:BS*, 364(1516):423–427, 2008. 2
- [44] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *PTR:BS*, 364(1516):423–427, 2009. 1
- [45] Martin Stevens and Sami Merilaita. *Animal camouflage: mechanisms and function*. Cambridge University Press, 2011. 2
- [46] Gusi Te, Yinglu Liu, Wei Hu, Haifan Shi, and Tao Mei. Edge-aware graph representation learning and reasoning for face parsing. In *ECCV*, 2020. 4
- [47] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 3
- [48] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang, Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, Wenyu Liu, and Bin Xiao. Deep high-resolution representation learning for visual recognition. *IEEE TPAMI*, 43(10):3349–3364, 2021. 1
- [49] Wenhui Wang, Enze Xie, Xiang Li, Deng-Ping Fan, Kaitao Song, Ding Liang, Tong Lu, Ping Luo, and Ling Shao. Pyramid vision transformer: A versatile backbone for dense prediction without convolutions. In *ICCV*, 2021. 2, 3, 5, 8
- [50] Jun Wei, Shuhui Wang, and Qingming Huang. F<sup>3</sup>net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020. 4, 5
- [51] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019. 5
- [52] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, 2019. 5
- [53] Jinnan Yan, Trung-Nghia Le, Khanh-Duy Nguyen, Minh-Triet Tran, Thanh-Toan Do, and Tam V Nguyen. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9:43290–43300, 2021. 5
- [54] Fuzhi Yang, Huan Yang, Jianlong Fu, Hongtao Lu, and Baining Guo. Learning texture transformer network for image super-resolution. In *CVPR*, 2020. 3
- [55] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *ICCV*, 2021. 1, 2, 5
- [56] Pang Youwei, Zhao Xiaoqi, Xiang Tian-Zhu, Zhang Lihe, and Lu Huchuan. Zoom in and out: A mixed-scale triplet network for camouflaged object detection. In *CVPR*, 2022. 2
- [57] Amir R Zamir, Te-Lin Wu, Lin Sun, William B Shen, Bertram E Shi, Jitendra Malik, and Silvio Savarese. Feedback networks. In *CVPR*, 2017. 2
- [58] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *CVPR*, 2021. 1, 2, 5
- [59] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgbd saliency detection via conditional variational autoencoders. In *CVPR*, 2020. 5
- [60] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *CVPR*, 2020. 5
- [61] Pingping Zhang, Wei Liu, Yi Zeng, Yinjie Lei, and Huchuan Lu. Looking for the detail and context devils: High-resolution salient object detection. *IEEE TIP*, 30:3204–3216, 2021. 1
- [62] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *ECCV*, 2018. 4
- [63] Hengshuang Zhao, Jianping Shi, Xiaojuan Qi, Xiaogang Wang, and Jiaya Jia. Pyramid scene parsing network. In *CVPR*, 2017. 5
- [64] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet:edge guidance network for salient object detection. In *ICCV*, 2019. 5, 6, 7
- [65] Ting Zhao and Xiangqian Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019. 5
- [66] Sixiao Zheng, Jiachen Lu, Hengshuang Zhao, Xiatian Zhu, Zekun Luo, Yabiao Wang, Yanwei Fu, Jianfeng Feng, Tao Xiang, Philip HS Torr, et al. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*, 2021. 3
- [67] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *CVPR*, 2020. 5
- [68] Zongwei Zhou, Md Mahfuzur Rahman Siddiquee, Nima Tajbakhsh, and Jianming Liang. Unet++: A nested u-net architecture for medical image segmentation. In *DLMIA*, 2018. 5
- [69] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*, 2017. 9