High-Resolution Iterative Feedback Network for Camouflaged Object Detection

Xiaobin Hu[1], Shuo Wang[2], Xuebin Qin[3], Hang Dai[4]*, Wenqi Ren[5], Donghao Luo[1], Ying Tai[1], Ling Shao[6]

[1]Tencent Youtu Lab
[2]ETH Zurich
[3]Mohamed bin Zayed University of Artificial Intelligence
[4]University of Glasgow
[5]Sun Yat-sen University
[6]Terminus Group
{xiaobinhu, michaelluo, yingtai}@tencent.com, shawnwang.tech@gmail.com, xuebin@ualberta.ca, Hang.Dai@glasgow.ac.uk, renwq3@mail.sysu.edu.cn, ling.shao@ieee.org

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. To design better feedback feature flow and avoid the feature corruption caused by recurrent path, an iterative feedback strategy 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 35 state-of-the-art methods. In addition, to address the data scarcity in camouflaged scenarios, we provide an application to convert the salient objects to camouflaged objects, thereby generating more camouflaged training samples from the diverse salient objects. Code will be made publicly available.

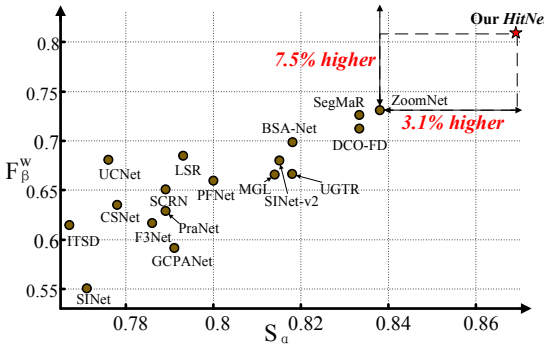


Figure 1: Weighted F-measure (F_β^w) *vs.* Structure-measure (S_α) of top 17 models from 35 SOTA methods and our *HitNet* on COD10k-Test dataset. F_β^w is a comprehensive metric to evaluate the weighted precision and recall of the prediction map, and S_α aims to analyze the structural information of the prediction map. Our framework achieves a remarkable performance milestone.

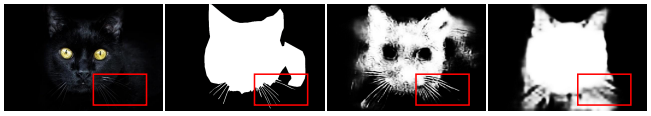
Introduction

Camouflaged object detection (COD) is a bio-inspired research area to detect hidden objects or animals that blend with their surroundings (Fan et al. 2021). From biological and psychological studies (Cuthill 2019; Stevens and Merilaita 2009), 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 (Dong et al. 2021; Fan et al. 2020b,c), artistic creation (Chu et al. 2010), species discovery (Pérez-de la Fuente et al. 2012), and crack inspection (Fang et al. 2020).

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 (Huerta et al. 2007), 3D convexity (Pan et al. 2011) and motion boundary (Yin et al. 2011)). These features still suffer from the limited capability of discriminating the foreground and the 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 (Fan et al. 2020a), classification (Le et al. 2019), boundary detection (Zhai et al. 2021) and uncertainties (Yang et al. 2021).

Although the approaches mentioned above have improved

* Corresponding author.



(a) Image (b) GT (c) HR result (d) LR result

Figure 2: Results from high-resolution (HR) and low-resolution (LR) inputs with SINet (Fan et al. 2020a) trained on LR images. In camouflaged scenario (a) where cat is assimilated into the dark background, cat's beard is very challenging to be segmented. From results analysis, LR result has blurred edges (*e.g.,* cat's beard), which indicates the high-frequency detail loss (*e.g.,* boundaries) during image degradation from HR to LR in camouflaged scenario.

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. Although HR result suffers from a bit of over-segmentation with a lot of noise, it still has more high-frequency details like cat beards than that from LR. This implies that the high-resolution priors are crucial to the boundary and edge detection (Zhang et al. 2021; Wang et al. 2021a). 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 , especially for camouflaged object detection where edges or boundaries are visually assimilated into the background.

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-resolution features. Thus, it is promising to explore how to maintain the high-resolution information 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 effective 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 in TFE stage, we adopt pyramid vision transformer (Wang et al. 2021b) as an image feature encoder. 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 (Li et al. 2017; Zhao et al. 2021, 2020) to camouflaged objects via our cross-domain learning strategy. Results from our application can be used as additional training data to further improve the segmentation accuracy of the COD task without increasing the parameters and computations of 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. To avoid the feature corruption caused by recurrent path, the corresponding iterative feedback loss with an iteration weight scheme is 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_β^w of 0.804, which is **7.5**% higher than the second-best ZoomNet22 (Youwei et al. 2022).

Related Work

Camouflaged Object Detection. COD aims to spot the camouflaged object from its high-similarity surroundings (Fan et al. 2020a). It has wide applications (Fan et al. 2020b; Chu et al. 2010; Pérez-de la Fuente et al. 2012) and many COD methods (Youwei et al. 2022; Cheng et al. 2022) 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 handcrafted features (*e.g.,* colour and intensity features (Huerta et al. 2007), 3D convexity (Pan et al. 2011), and motion boundary (Yin et al. 2011)). 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, boundaries, and location. 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) (Zamir et al. 2017; Hu et al. 2021). In image super-resolution, some studies proved certain improvements after using different feedback mechanisms, such as up-

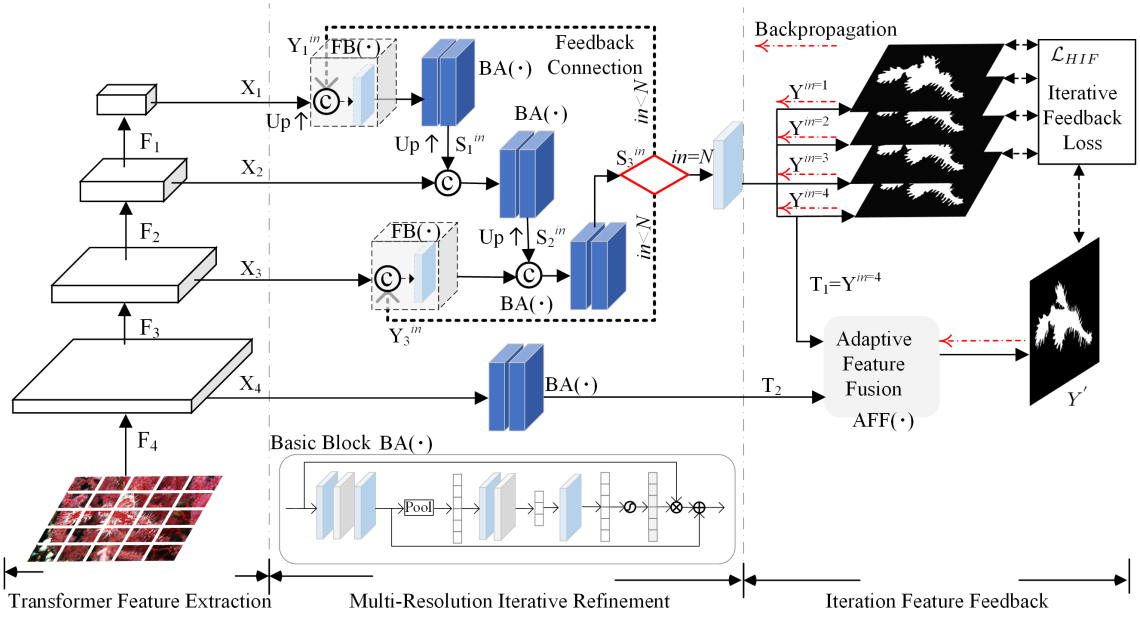


Figure 3: An overview of High-resolution Iterative Feedback Network (*HitNet*). Our HitNet consists of Transformed-based backbone for multi-scale feature extraction, multi-resolution iterative refinement to self-correct low-resolution features with high-resolution information via a cross-resolution iterative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

and down-projection units (Haris, Shakhnarovich, and Ukita 2018) and dual-state recurrent module (Han et al. 2018). However, most of these mechanisms are implemented by using recurrent structures (Li et al. 2019) while the information flows from the LR to HR images are still feed-forward. Recently, Li *et al.* (Li et al. 2019) 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.* (Li et al. 2019), our feedback connection is designed as a global connection other than a local connection (Feng, Lu, and Ding 2019) and embedded into the multi-scale framework via a feedback fusion block, which merges the information from multi-scale outputs. To avoid corruption of each iteration, we impose more constraints on each feedback connection by supervising each iteration with the corresponding loss.

Vision Transformer. The Transformer (Vaswani et al. 2017) was firstly proposed 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 (Dosovitskiy et al. 2020; Srinivas et al. 2021), semantic segmentation (Zheng et al. 2021), object detection (Dai et al. 2021), and other low-level tasks (Yang et al. 2020). Thus, we adopt a Pyramid Vision Transformer (PVT) (Wang et al. 2021b) 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.

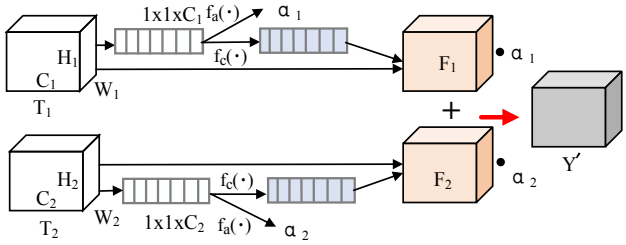


Figure 4: Adaptive Feature Fusion.

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.

Transformer-based Feature Extraction

Currently, 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) (Wang et al. 2021b) 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.

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 ($\boldsymbol{F_i}$) 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 ($\boldsymbol{X_1}$, $\boldsymbol{X_2}$, $\boldsymbol{X_3}$, $\boldsymbol{X_4}$) with (512, 320, 128, 64) number of channels and with ($\frac{1}{32}$, $\frac{1}{16}$, $\frac{1}{8}$, $\frac{1}{4}$) resolution of input images for further processing.

Multi-Resolution Feedback Refinement

The multi-scale resolution features $\boldsymbol{X}$ extracted from the Transformer backbone are fed to a basic block $\mathrm{BA}(\cdot)$ (Zhang et al. 2018) as shown in Fig. 3:

$$\mathrm{BA}(\boldsymbol{X_i}) = C_2(\boldsymbol{X_i}) + C_b(C_2(\boldsymbol{X_i})) \cdot \boldsymbol{X_i}, \tag{1}$$

where $\boldsymbol{X_i}$ is the input feature of i-th scale produced by Transformer module, $C_2(\cdot)$ indicates two stacked convolutional layers with 3×3 filters. $C_b(\cdot)$ denotes the channel attention function (Zhang et al. 2018).

Iterative Feedback Mechanism is critical in this module to achieve high accuracy around the object boundary (see Fig. 6). The setting iterative number *in*=1 assumes the first iteration and no feedback feature transported from previous state. Thus, the $\boldsymbol{Y_1^{in}}$ and $\boldsymbol{Y_3^{in}}$ 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 $\mathrm{FB}(\cdot)$ as:

$$\mathrm{FB}(\boldsymbol{X_i} + \boldsymbol{Y_i^{in}}) = \mathrm{Sq}(\mathrm{Concat}(\boldsymbol{X_i} \uparrow, \boldsymbol{Y_i^{in}})), \tag{2}$$

where $\boldsymbol{Y_i^{in}}$ 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 $\boldsymbol{X_i}$ to $\boldsymbol{Y_i^{in}}$ to avoid degradation of the HR information. $\mathrm{Concat}(\cdot)$ indicates the channel-based concatenation operation between $\boldsymbol{X_i} \uparrow$ and $\boldsymbol{Y_i^{in}}$, and $\mathrm{Sq}(\cdot)$ is feature size and channel compression using convolution layer with large kernel and stride[1] to get 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 $\boldsymbol{X_1}$ and $\boldsymbol{Y_1^{in}}$ and the output the feature can be defined as:

$$\boldsymbol{S_1^{in}} = \mathrm{BA}(\mathrm{FB}(\boldsymbol{X_1} + \boldsymbol{Y_1^{in}})), \tag{3}$$

Then, $\boldsymbol{S_1^{in}}$ is further fed into the next scale to generate next output feature as follows:

$$\boldsymbol{S_2^{in}} = \mathrm{BA}(\mathrm{Concat}(\boldsymbol{S_1^{in}} \uparrow, \boldsymbol{X_2})), \tag{4}$$

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

$$\boldsymbol{S_3^{in}} = \mathrm{BA}(\mathrm{Concat}(\boldsymbol{S_2^{in}} \uparrow, \mathrm{FB}(\boldsymbol{X_3} + \boldsymbol{Y_3^{in}}))), \tag{5}$$

After ending at in-th iteration, $(in + 1)$-th iteration starts from the first scale to the last scale in the same way. The feedback features $\boldsymbol{Y_1^{in+1}}$ and $\boldsymbol{Y_3^{in+1}}$ at $(in+1)$-th iteration are updated as follows:

$$\boldsymbol{Y_1^{in+1}} = \boldsymbol{Y_3^{in+1}} = \mathrm{Conv}(\boldsymbol{S_3^{in}}), \tag{6}$$

[1]If i=1, kernel = 8 with stride = 4 while i=3, kernel = 1 with stride = 1.

where Conv is a convolution layer with 3 kernel size and 1 padding. The segmentation prediction map of i-th iteration ($\boldsymbol{Y^{in}}$) is obtained via two stacked convolution operation $\boldsymbol{Y^{in}} = \mathrm{Conv}(\mathrm{Conv}(\boldsymbol{S_3^{in}}))$. The low and upper index of $\boldsymbol{Y}$ indicates the scale and the iteration information.

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 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 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.

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 (Wei, Wang, and Huang 2020), 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:

$$\mathcal{L}_{HIF} = \sum_{in}^{N} (w \cdot in)\mathcal{L}(Y^{in}) + \mathcal{L}(Y^{'}), \tag{7}$$

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 design an adaptive feature fusion module (shown in Fig. 4).

$$Y^{'} = \mathrm{AFF}(T_1, T_2), \tag{8}$$

where $Y^{'}$ is final prediction map, T_1 is the $Y^{in=4}$, AFF is the Adaptive Feature Fusion module. Specifically, when given the features T_1 and T_2, the global average pooling is used to get the shrunk features with the size of $1 \times 1 \times C$ on the channel dimension. Afterwards, the operations (f_c and f_a) are implemented on the shrunk features to get the channel-wise weights and adaptive feature-wise coefficients

Baseline Models	CAMO-Test				COD10K-Test			
	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
MirrorNet (Yan et al. 2021)	0.741	0.804	0.652	0.100	‡	‡	‡	‡
MGL (Zhai et al. 2021)	0.775	0.847	0.673	0.088	0.814	0.865	0.666	0.035
PFNet (Mei et al. 2021)	0.782	0.852	0.695	0.085	0.800	0.868	0.660	0.040
UGTR (Yang et al. 2021)	0.785	0.859	0.686	0.086	0.818	0.850	0.667	0.035
UJSC (Li et al. 2021)	0.800	0.858	0.728	0.073	0.809	0.884	0.684	0.035
C^2FNet (Sun et al. 2021)	0.796	0.853	0.719	0.080	0.813	0.890	0.686	0.036
LSR (Lv et al. 2021)	0.793	0.826	0.725	0.085	0.793	0.868	0.685	0.041
SINet-V2 (Fan et al. 2021)	0.820	0.882	0.743	0.070	0.815	0.887	0.680	0.037
DCO-FD (Zhong et al. 2022)	0.828	0.884	0.747	0.069	0.833	0.907	0.711	0.033
BSA-Net (Zhu et al. 2022)	0.796	0.851	0.717	0.079	0.818	0.891	0.699	0.034
SegMaR (Jia et al. 2022)	0.815	0.872	0.742	0.071	0.833	0.895	0.724	0.033
ZoomNet (Pang et al. 2022)	0.819	0.877	0.752	0.065	0.838	0.887	0.729	0.029
HitNet (Ours)	**0.844**	**0.904**	**0.806**	**0.056**	**0.869**	**0.936**	**0.804**	**0.023**

Table 1: Quantitative results of our method and the most recent 12 of 35 state-of-the-art methods on two benchmark datasets.

(α). f_c and f_a are the stacked combinations of operators *nn.Linear+ReLU+nn.Linear+Sigmoid*. The features F_1 and F_2 are obtained after assigning the channel-wise weights on the channel features of T1 and T2. Finally, the final prediction map $Y^{'}$ can be achieved weighted by adaptive coefficients α_1 and α_2 as follows:

$$Y^{'} = F_1 \times \alpha_1 + F_2 \times \alpha_2, \tag{9}$$

Experiments

Experimental Settings

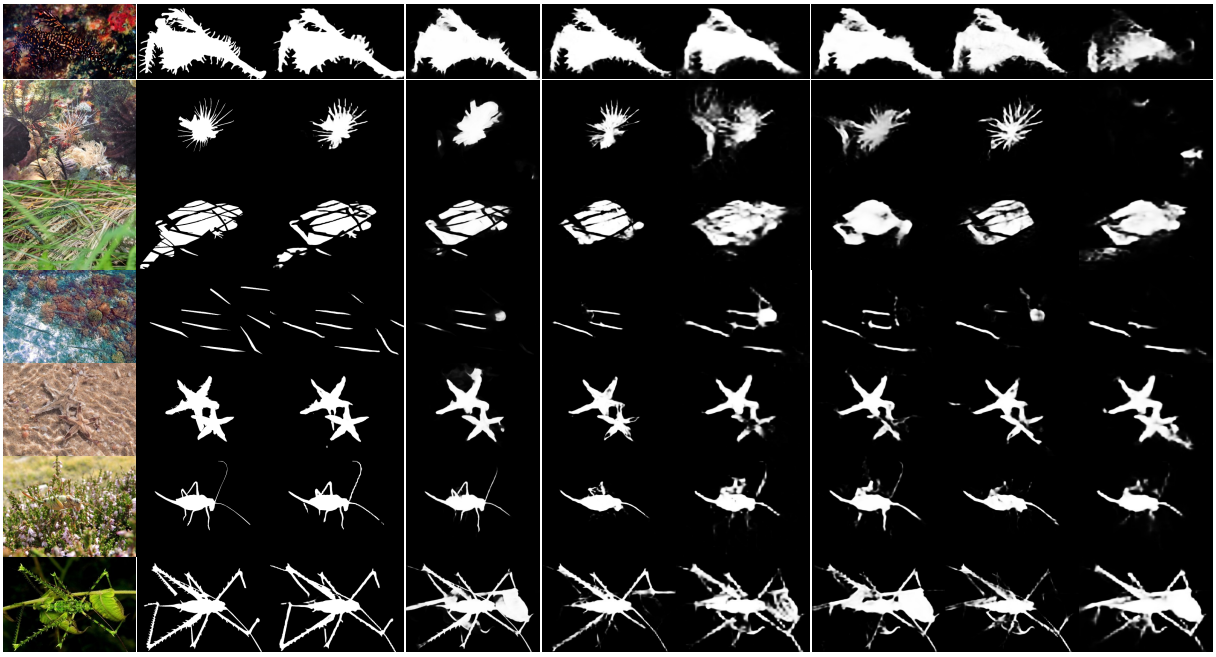
Datasets. Our experiments are based on four widely-used COD datasets: (1) CHAMELEON (Skurowski et al. 2018) collects 76 high-resolution images from the Internet with the label of camouflaged animals. (2) CAMO (Le et al. 2019) includes 2,500 images with eight categories. (3) COD10K (Fan et al. 2020a) is the largest collection containing 10,000 images that divided into 10 super-classes and 78 sub-classes from multiple photography websites. (4) NC4K (Lv et al. 2021) consists of 4,121 images and is commonly used to evaluate the generalization ability of models. Following previous studies and benchmarks (Zhai et al. 2021; Fan et al. 2020a; Yang et al. 2021; Lv et al. 2021), the training set includes 1,000 images from CAMO, and 3,040 images from COD10K. The test set consists of 76 images from CHAMELEON, 250 images from CAMO, 2,026 images from COD10K, and 4,121 images from NC4K.

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, and no data augmentation is used in our model. The transformer-based feature extraction is initialized by PVT-V2 (Wang et al. 2021b), and the remaining modules are initialized in a random manner. We employ the AdamW (Loshchilov and Hutter 2019) 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 (N) 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.

Quantitative and Qualitative Evaluation. In Tab. 1, as a performance milestone, our model are significantly superior than other existing 35 methods on all datasets and all metrics. Compared with second-best models 2022 ZoomNet, our *HitNet* averagely lowers the relative MAE error by **16.9%** and improves relative F_β^w by **7.5%** on four datasets. Fig. 5 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 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.

Ablation Study

Effectiveness of Each Component. As shown in Tab. 2, 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 (Gao et al. 2021) used in SINet-V2 (Fan et al. 2021) as the version of 'w/o TFE'. Our algorithm without PVT backbone still achieves the best performance compared with all 35 SOTA methods. Besides, we also remove the RIR module from HitNet by substituting them with two stacked convolutional layers in each scale, expressed as 'w/o RIR'. The F_β^w performance of this variant sharply deteriorates from 0.804 to 0.712. Lastly, we also remove the constraints on feedback feature flow in IFF module as the version of 'w/o IFF-1' and replace the adaptive feature fusion (AFF) with an element-wise feature summation as the version of 'w/o IFF-2'. We find these variants also decreases the performance to some extent. But our RIR module plays a crucial role in performance improvement than others.



(a) Images (b) GT (c) HitNet (d) ZoomNet (e) SegMaR (f) SINetV2 (g) PFNet (h) LSR (i) PraNet

Figure 5: 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).

Configuration of Iteration Number. We explore the effect of iteration number in the iterative feedback mechanism on inference time and performance. To analyze the difference among the iteration number, we visualize the feature of each iteration in Fig. 6. 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.,* sharper edges).

Study of Iterative Feedback Mechanism. As discussed in §, 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 Tab. 3, any absence of three factors will fail the model to drive the data flow.

Application

The camouflaged dataset is very scarce and rare only existing in camouflaged scenarios, and almost all public camouflaged datasets have been used in our paper. In contrast, there exists a large-scale salient dataset that is almost 100 times more than the camouflaged ones. It is an open question that how to well-utilize the abundant salient dataset to improve the camouflaged object accuracy without extra annotation labor. Thus, we adopt a cross-domain learning (CDL) technique that converts salient objects to camouflage objects 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.

Cross-domain Learning. We employ the cycle-consistency structure (Zhu et al. 2017) to learn the camouflaged features and embed these features into the salient objects in an unsupervised cross-domain learning manner as shown in Fig. 7. The cycle-consistency loss can be formulated as:

$$\begin{aligned}\mathcal{L}_{cyc}(G,F) =& \mathbb{E}_x[\|(F(G(x))-x)\|_1]+ \\ & \mathbb{E}_y[\|(G(F(y))-y)\|_1]\end{aligned} \tag{10}$$

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[2] in the adversarial loss function used in CycleGAN.

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

$$I_{sc} = \frac{1}{\mathrm{Num}} \sum_{i}^{\mathrm{Num}} \|P_i - P_{\mathrm{m}}\|_{i\in(P_{\mathrm{m}}-P_{\mathrm{std}},P_{\mathrm{m}}+P_{\mathrm{std}})}, \tag{11}$$

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_{std} is the standard deviation, and i is the pixel index that belongs to one σ rule to exclude the effect of extreme values. In Fig. 7, the car is an abandoned example detected as a high salient case by our contrastive index. Empirically, we set the threshold of I_{sc} as $I_{sc} = 20$.

[2]Minimize the generator loss while maximized the discriminator loss.

Metric	w/o TFE	w/o RIR	w/o IFF-1	w/o IFF-2	HitNet
$F_\beta^w \uparrow$	0.745	0.712	0.793	0.801	**0.804**
$S_\alpha \uparrow$	0.851	0.833	0.860	0.862	**0.869**
$M \downarrow$	0.026	0.031	0.025	0.024	**0.023**
$E_\phi \uparrow$	0.921	0.907	0.931	0.932	**0.936**

Table 2: Ablation analyses of HitNet on COD10K dataset.

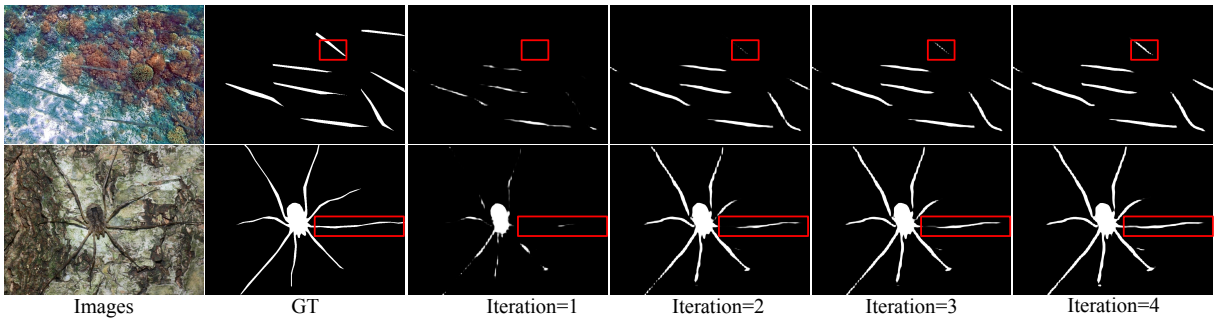


Figure 6: Visual performance of each iteration in our iterative feedback mechanism in our RIR module.

Configurations			Performance
Tie	FB	Multi-fusion	MAE ↓
×	×	×	0.0265
√	×	×	0.0252
√	√	×	0.0246
√	√	√	**0.0230**

Table 3: Ablation study on indispensable factors of Iterative Feedback Mechanism on COD10K dataset.

	COD10K (Fan et al. 2020a)			
Data Strategy	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
w/o	0.869	0.936	0.804	0.023
Salient data	0.837	0.908	0.778	0.027
CDL (Ours)	**0.879**	**0.939**	**0.812**	**0.022**

Table 4: Quantitative results on different training strategies. 'w/o' means without any data strategy, 'Salient data' means adding salient data for training.

Qualitative and Quantitative Evaluation. Our CDL strategy makes exciting and meaningful explorations on camouflage domains from the following aspects. (1) As shown in Fig. 7, it converts salient objects to camouflage objects, which bridges the gap between salient and camouflage fields. (2) As shown in Tab. 4, it finds that the usage of salient object data cannot improve but severely deteriorate the performance of COD. Meanwhile, CDL strategy can make the distribution of salient objects closer to the camouflage object distribution. Thus, it can improve the accuracy of COD and reduce the relative MAE error by 4.3%.

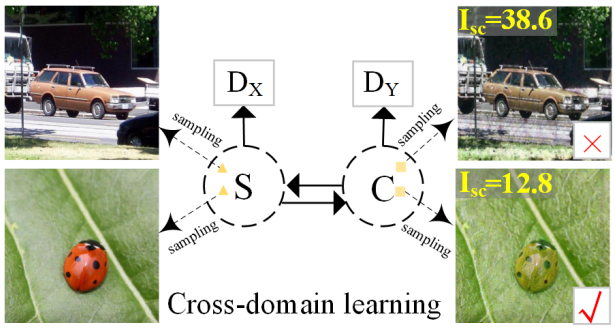


Figure 7: The overview of salient-to-camouflaged cross-domain learning pipeline. The S is the salient domain, and the C means the camouflaged domain. D_X is the discriminator of the salient domain, and D_Y is the discriminator of the camouflaged domain.

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 35 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.

References

Cheng, X.; Xiong, H.; Fan, D.-p.; Zhong, Y.; Harandi, M.; Drummond, T.; and Ge, Z. 2022. Implicit Motion Handling for Video Camouflaged Object Detection. In *CVPR*.

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

Cuthill, I. 2019. Camouflage. *JOZ*, 308(2): 75–92.

Dai, Z.; Cai, B.; Lin, Y.; and Chen, J. 2021. Up-detr: Unsupervised pre-training for object detection with transformers. In *CVPR*.

Dong, B.; Wang, W.; Fan, D.-P.; Li, J.; Fu, H.; and Shao, L. 2021. Polyp-PVT: Polyp Segmentation with PyramidVision Transformers. *arXiv preprint arXiv:2108.06932*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*.

Fan, D.-P.; Ji, G.-P.; Cheng, M.-M.; and Shao, L. 2021. Concealed Object Detection. *IEEE TPAMI*.

Fan, D.-P.; Ji, G.-P.; Sun, G.; Cheng, M.-M.; Shen, J.; and Shao, L. 2020a. Camouflaged object detection. In *CVPR*.

Fan, D.-P.; Ji, G.-P.; Zhou, T.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020b. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*.

Fan, D.-P.; Zhou, T.; Ji, G.-P.; Zhou, Y.; Chen, G.; Fu, H.; Shen, J.; and Shao, L. 2020c. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE TMI*, 39(8): 2626–2637.

Fang, F.; Li, L.; Gu, Y.; Zhu, H.; and Lim, J.-H. 2020. A novel hybrid approach for crack detection. *Pattern Recognition*, 107: 107474.

Feng, M.; Lu, H.; and Ding, E. 2019. Attentive feedback network for boundary-aware salient object detection. In *CVPR*.

Gao, S.; Cheng, M.-M.; Zhao, K.; Zhang, X.-Y.; Yang, M.-H.; and Torr, P. H. 2021. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 43(02): 652–662.

Han, W.; Chang, S.; Liu, D.; Yu, M.; Witbrock, M.; and Huang, T. S. 2018. Image super-resolution via dual-state recurrent networks. In *CVPR*.

Haris, M.; Shakhnarovich, G.; and Ukita, N. 2018. Deep back-projection networks for super-resolution. In *CVPR*.

Hu, X.; Yan, Y.; Ren, W.; Li, H.; Bayat, A.; Zhao, Y.; and Menze, B. 2021. Feedback Graph Attention Convolutional Network for MR Images Enhancement by Exploring Self-Similarity Features. In *MIDL*.

Huerta, I.; Rowe, D.; Mozerov, M.; and Gonzàlez, J. 2007. Improving background subtraction based on a casuistry of colour-motion segmentation problems. In *IbPRIA*.

Jia, Q.; Yao, S.; Liu, Y.; Fan, X.; Liu, R.; and Luo, Z. 2022. Segment, Magnify and Reiterate: Detecting Camouflaged Objects the Hard Way. In *CVPR*, 4713–4722.

Le, T.-N.; Nguyen, T. V.; Nie, Z.; Tran, M.-T.; and Sugimoto, A. 2019. Anabranch network for camouflaged object segmentation. *CVIU*, 184: 45–56.

Li, A.; Zhang, J.; Lv, Y.; Liu, B.; Zhang, T.; and Dai, Y. 2021. Uncertainty-aware joint salient object and camouflaged object detection. In *CVPR*, 10071–10081.

Li, X.; Yang, F.; Cheng, H.; Chen, J.; Guo, Y.; and Chen, L. 2017. Multi-scale cascade network for salient object detection. In *ACM MM*, 439–447.

Li, Z.; Yang, J.; Liu, Z.; Yang, X.; Jeon, G.; and Wu, W. 2019. Feedback network for image super-resolution. In *CVPR*.

Loshchilov, I.; and Hutter, F. 2019. Decoupled weight decay regularization. In *ICLR*.

Lv, Y.; Zhang, J.; Dai, Y.; Li, A.; Liu, B.; Barnes, N.; and Fan, D.-P. 2021. Simultaneously localize, segment and rank the camouflaged objects. In *CVPR*.

Mei, H.; Ji, G.-P.; Wei, Z.; Yang, X.; Wei, X.; and Fan, D.-P. 2021. Camouflaged object segmentation with distraction mining. In *CVPR*.

Pan, Y.; Chen, Y.; Fu, Q.; Zhang, P.; and Xu, X. 2011. Study on the camouflaged target detection method based on 3D convexity. *MAS*, 5(4): 152–157.

Pang, Y.; Zhao, X.; Xiang, T.-Z.; Zhang, L.; and Lu, H. 2022. Zoom in and Out: A Mixed-Scale Triplet Network for Camouflaged Object Detection. In *CVPR*, 2160–2170.

Pérez-de la Fuente, R.; Delclòs, X.; Peñalver, E.; Speranza, M.; Wierzchos, J.; Ascaso, C.; and Engel, M. S. 2012. Early evolution and ecology of camouflage in insects. *PNAS*, 109(52): 21414–21419.

Skurowski, P.; Abdulameer, H.; Błaszczyk, J.; Depta, T.; Kornacki, A.; and Kozieł, P. 2018. Animal Camouflage Analysis: CHAMELEON Database. Unpublished Manuscript.

Srinivas, A.; Lin, T.-Y.; Parmar, N.; Shlens, J.; Abbeel, P.; and Vaswani, A. 2021. Bottleneck transformers for visual recognition. In *CVPR*.

Stevens, M.; and Merilaita, S. 2009. Animal camouflage: current issues and new perspectives. *PTR:BS*, 364(1516): 423–427.

Sun, Y.; Chen, G.; Zhou, T.; Zhang, Y.; and Liu, N. 2021. Context-aware Cross-level Fusion Network for Camouflaged Object Detection. In *IJCAI*, 1025–1031.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. In *NeurIPS*.

Wang, J.; Sun, K.; Cheng, T.; Jiang, B.; Deng, C.; Zhao, Y.; Liu, D.; Mu, Y.; Tan, M.; Wang, X.; Liu, W.; and Xiao, B. 2021a. Deep High-Resolution Representation Learning for Visual Recognition. *IEEE TPAMI*, 43(10): 3349–3364.

Wang, W.; Xie, E.; Li, X.; Fan, D.-P.; Song, K.; Liang, D.; Lu, T.; Luo, P.; and Shao, L. 2021b. Pyramid Vision Transformer: A Versatile Backbone for Dense Prediction Without Convolutions. In *ICCV*.

Wei, J.; Wang, S.; and Huang, Q. 2020. F^3Net: Fusion, Feedback and Focus for Salient Object Detection. In *AAAI*.

Yan, J.; Le, T.-N.; Nguyen, K.-D.; Tran, M.-T.; Do, T.-T.; and Nguyen, T. V. 2021. Mirrornet: Bio-inspired camouflaged object segmentation. *IEEE Access*, 9: 43290–43300.

Yang, F.; Yang, H.; Fu, J.; Lu, H.; and Guo, B. 2020. Learning texture transformer network for image super-resolution. In *CVPR*.

Yang, F.; Zhai, Q.; Li, X.; Huang, R.; Luo, A.; Cheng, H.; and Fan, D.-P. 2021. Uncertainty-Guided Transformer Reasoning for Camouflaged Object Detection. In *ICCV*.

Yin, J.; Han, Y.; Hou, W.; and Li, J. 2011. Detection of the mobile object with camouflage color under dynamic background based on optical flow. *PE*, 15: 2201–2205.

Youwei, P.; Xiaoqi, Z.; Tian-Zhu, X.; Lihe, Z.; and Huchuan, L. 2022. Zoom In and Out: A Mixed-scale Triplet Network for Camouflaged Object Detection. In *CVPR*.

Zamir, A. R.; Wu, T.-L.; Sun, L.; Shen, W. B.; Shi, B. E.; Malik, J.; and Savarese, S. 2017. Feedback networks. In *CVPR*.

Zhai, Q.; Li, X.; Yang, F.; Chen, C.; Cheng, H.; and Fan, D.-P. 2021. Mutual graph learning for camouflaged object detection. In *CVPR*.

Zhang, P.; Liu, W.; Zeng, Y.; Lei, Y.; and Lu, H. 2021. Looking for the detail and context devils: High-resolution salient object detection. *IEEE TIP*, 30: 3204–3216.

Zhang, Y.; Li, K.; Li, K.; Wang, L.; Zhong, B.; and Fu, Y. 2018. Image super-resolution using very deep residual channel attention networks. In *ECCV*.

Zhao, J.; Zhao, Y.; Li, J.; and Chen, X. 2020. Is depth really necessary for salient object detection? In *ACM MM*, 1745–1754.

Zhao, Z.; Xia, C.; Xie, C.; and Li, J. 2021. Complementary Trilateral Decoder for Fast and Accurate Salient Object Detection. In *ACM MM*, 4967–4975.

Zheng, S.; Lu, J.; Zhao, H.; Zhu, X.; Luo, Z.; Wang, Y.; Fu, Y.; Feng, J.; Xiang, T.; Torr, P. H.; et al. 2021. Rethinking semantic segmentation from a sequence-to-sequence perspective with transformers. In *CVPR*.

Zhong, Y.; Li, B.; Tang, L.; Kuang, S.; Wu, S.; and Ding, S. 2022. Detecting Camouflaged Object in Frequency Domain. In *CVPR*, 4504–4513.

Zhu, H.; Li, P.; Xie, H.; Yan, X.; Liang, D.; Chen, D.; Wei, M.; and Qin, J. 2022. I can find you! Boundary-guided Separated Attention Network for Camouflaged Object Detection. AAAI.

Zhu, J.-Y.; Park, T.; Isola, P.; and Efros, A. A. 2017. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *ICCV*.