# D2C-Net: A Dual-branch, Dual-guidance and Cross-refine Network for Camouflaged Object Detection

K. Wang, H. Bi, Y. Zhang, C. Zhang, Z. Liu, S. Zheng

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

# D$^2$C-Net: A Dual-branch, Dual-guidance and Cross-refine Network for Camouflaged Object Detection

Kang Wang, Hongbo Bi, Yi Zhang, Cong Zhang, Ziqi Liu, Shuang Zheng

***Abstract*—In this work, we propose a novel framework for camouflaged object detection (COD), named D$^2$C-Net, which contains two new modules: dual-branch features extraction (DFE) and gradually refined cross fusion (GRCF). Specifically, the DFE simulates the two-stage detection process of human visual mechanisms in observing camouflage scenes. For the first stage, a dense concatenation is employed to aggregate multi-level features and expand the receptive field. The first stage feature maps are then utilized to extract two-direction guidance information, which benefits the second stage. The GRCF consists of a self-refine attention unit and a cross-refinement unit, with the aim of combining the peer layer features and DFE features for an improved COD performance. The proposed framework outperforms 13 state-of-the-art deep learning-based methods upon three public datasets in terms of five widely used metrics. Finally, we show evidence for the successful applications of the proposed method in the fields of surface defect detection, medical image segmentation.**

***Index Terms*—Camouflaged Object Detection, Deep Learning, Receptive Field**

## I. Introduction

WHEN we look at an image or face to a scene, if the discrimination between the foreground and the background is high, we first observe the most attractive object, that is, the salient object [1]–[4]. On the contrary, if the discrimination between the foreground and the background is low, it is difficult to observe the foreground object at first sight, that is, the camouflaged object [5]–[7]. Camouflaged objects can be roughly divided into two categories, natural camouflaged objects and artificial camouflaged objects. Natural camouflage objects [8]–[10] refer to animals and plants hiding in the background environment with their own advantages (such as body shape, color, etc.) in order to adapt to the environment and protect themselves, while artificial ones (e.g., military camouflage clothes) occur in real daily life. Besides, the system equipped with camouflaged detection function has a lot of potential applications, including medical image segmentation [11]–[13], surface defect detection [14], search and rescue work [15], image change detection [16], etc.

In order to improve the accuracy of the corresponding camouflaged detection system, it is particularly important to design a COD algorithm with state-of-the-art performance. At present, the existing COD algorithms can be roughly divided into traditional and deep learning-based methods. Traditional methods [17]–[19] mainly rely on manually designed features to detect or segment camouflaged objects. Thanks to the rapid development of deep learning in the recent two years, the detection accuracy of the COD algorithm has also been effectively improved. For example, to explore the detection of human who wears camouflage pattern in complex background scenes, Zheng et al. [20] employed a bit of short connections into the framework that is beneficial to predict accurate camouflage maps. Le et al. [21] proposed a two-stream (classification stream and detection stream) framework for camouflage object detection. Recently, Fan et al. [22] proposed SINet, which simulates the human visual mechanism. Meanwhile, RF [23] structure is used to improve the receptive field of the region to be detected, and satisfactory results are achieved.

Considering the different contributions of low-level features and high-level features to generate the camouflage maps [21]–[23], and due to the high-resolution of low-level features, we employ the third layer of the backbone as the node to simulate the two stages of the human eye observing the camouflage scene. The first stage is simulated to the human observation of potential camouflage objects in the scene. The second stage is similar to the judgment and optimization of camouflage objects according to the prior knowledge learned in the previous stage. These two stages constitute the DFE module. At the same time, we introduce the bi-directional (positive attention, reverse attention) guidance information to guide each level after the node. Furthermore, in order to employ the shallow low-level semantic information to supplement the high-level semantic information, so as to make the detection results more complete. We adopt peer-to-peer features to refine the camouflage map generated in the previous stage by combining two optimization strategies (self-refine attention and cross-refinement), which constitute the GRCF module. Our main contributions are listed as follows:

1) We propose a novel approach for COD, which consists of two main components, namely, DFE module and GRCF module. We carry out a series of experiments to prove the effectiveness of our proposed model.
2) We simulate the observation process of human visual

Kang Wang, Cong Zhang and Ziqi Liu are with the School of Electrical Engineering & Information Department, Northeast Petroleum University, Daqing, 163318, China (e-mail: kangwang@stu.nepu.edu.cn, Congzhang98@stu.nepu.edu.cn and liuziqi@stu.nepu.edu.cn).

Hongbo Bi is with the School of Electrical Engineering & Information Department, Northeast Petroleum University, Daqing, 163318, China ( phone number: 136-6459-0305; e-mail: bihongbo@nepu.edu.cn).

Yi Zhang is with the IETR Lab, INSA Rennes , Rennes, 35510, France (e-mail: yi.zhang1@insa-rennes.fr)

Shuang Zheng is with the Tianjin chain number technology company, Tianjin, 300000, China (e-mail: s_zheng@linkdata.email).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

mechanism observing the camouflage scene in two stages, which employs the features generated in the first stage to extract bi-direction guidance information, and then make the second stage prediction.

3) Benefiting from the peer features can provide some cues about the location of target object regions, we employ a self-refine attention unit and cross-refine unit to conduct more accurate camouflage maps.
4) To prove the superiority of our model, we compare our model with other 13 state-of-the-art models on three benchmark datasets (CAMO [21], CHAMELEON [24], COD10K [5]). Our model acquires the best performance basing on all five public evaluation metrics.

## II. Related works

In this section, we discuss works about camouflaged object detection and residual attention.

### A. Camouflaged object detection

In the early stage, traditional methods usually employ handcrafted features to detect camouflage objects or perform the work related to camouflage objects [25]. Xue et al. [26] proposed a nonlinear feature fusion method to evaluate the camouflage degree of the target object. In [27], an unsupervised method is used to detect camouflage objects using Gray Level Co-occurrence Matrix based on the image block. The model is limited when dealing with a non-uniform backbone. Yin et al. [28] used Optical Flow to detect moving objects with camouflage colors. Pan et al. [29] proposed a camouflage object detection model based on 3D convexity, which can deal with more complex backgrounds.

In recent years, to accurately distinguish the camouflaged objects in complex backgrounds, increasing architectures apply convolutional neural networks (CNNs) to camouflaged object detection. For example, Nie et al. [21] proposed an end-to-end network (ANet) based on CNNs to precisely separate the camouflaged objects from given images. In [20], Zheng et al. presented a dense deconvolution network (DDCN) for camouflaged object detection to effectively detect the target object in the background of artificial camouflage. To more accurately locate camouflage objects in complex backgrounds, DDCN introduced the deep convolutional network to extract the semantic feature information. When the human eyes are performing some visual tasks (such as salient object detection [30], [31], pedestrian re-recognition, etc.), the processes are usually carried out in two steps, namely, search and identification. In [22], Fan et al. constructed a network (SINet) which consists of two main components, the search module and the identification module. The search module is mainly used to save the feature information of various levels and the identification module is applied to precisely locate and distinguish camouflaged objects.

### B. Residual attention

In order to make full use of the features and improve the performance, some detection [32] or segmentation [13] tasks attempt to introduce attention mechanisms (such as spatial attention, channel attention and residual attention) to obtain more complete feature clues. In [33], Chen et al. presented a network based on deep learning for salient object detection which introduced the reverse attention to capturing the side-output residual features. The reverse attention mechanism can also be applied to medical image detection. In [13], Fan et al. introduced reverse attention to Inf-Net to accurately segment the Coronavirus Disease 2019 (COVID-19) from CT images. Inf-Net adopts a method that adaptively learning reverse attention among three parallel high-level features, instead of simply integrating feature information from all levels. The residual information can achieve saliency refinement by simple convolutional parameters, meanwhile keeping the detection accuracy unchanged. To utilize the multi-scale features effectively, Gao et al. [34] presented a multi-scale backbone structure, namely, Res2Net. This method constructs a bit of hierarchical residual-like connections in a residual block which is benefiting to extract local and global features.

## III. Our Proposed Method

In this section, we first describe our proposed network for COD. Then, we enumerate each component in detail and illustrate its effectiveness.

### A. Motivation

As mentioned in related work, mostly traditional methods use low-level handcraft features to evaluate the performance of camouflage patterns or detect / segment camouflage objects in simple scenarios. Because the handcraft features have the bias of human prior knowledge, the results are not satisfactory. Although the performance of the model based on deep learning is much better than the traditional algorithm, most of the previous work only focuses on the detection of camouflaged objects in specific scenes. Therefore, the generalization ability and application scenarios are limited. SINet has achieved the best performance, which not only benefits from the new standard data set proposed by the author but also the novelty of the model design. In terms of its performance, it can be further improved.

Generally speaking, previous models have the following shortcomings: the bias of the handcraft features, the shortage of data sets, the singleness of application scenarios of model design, etc. Considering these problems, we rethink the problem of COD and propose $D^2$C-Net. For the COD task, the main purpose is to detect objects that are similar to the background. From the perspective of human visual mechanism, it is usually difficult to find the complete camouflaged objects in the first stage (i.e. the first observation) when we look at a scene. In order to capture a more detailed camouflage map, we attempt to carry on the second stage on the basis of a comprehensive analysis of the first stage. After two stages of observation, we can obtain a relatively accurate camouflage map. Based on this, we propose a COD framework that imitates a two-stage observation mechanism.

### B. Overview of Our Proposed Network for COD

As shown in Fig. 1, our proposed network consists of two main parts, Dual-branch Features Extraction (DFE) module and Gradually Refined Cross Fusion (GRCF) module. In the DFE module, we incorporate a two-stage observation strategy. Inspired by [22], [23], we take $Conv3$ as the bifurcation point.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

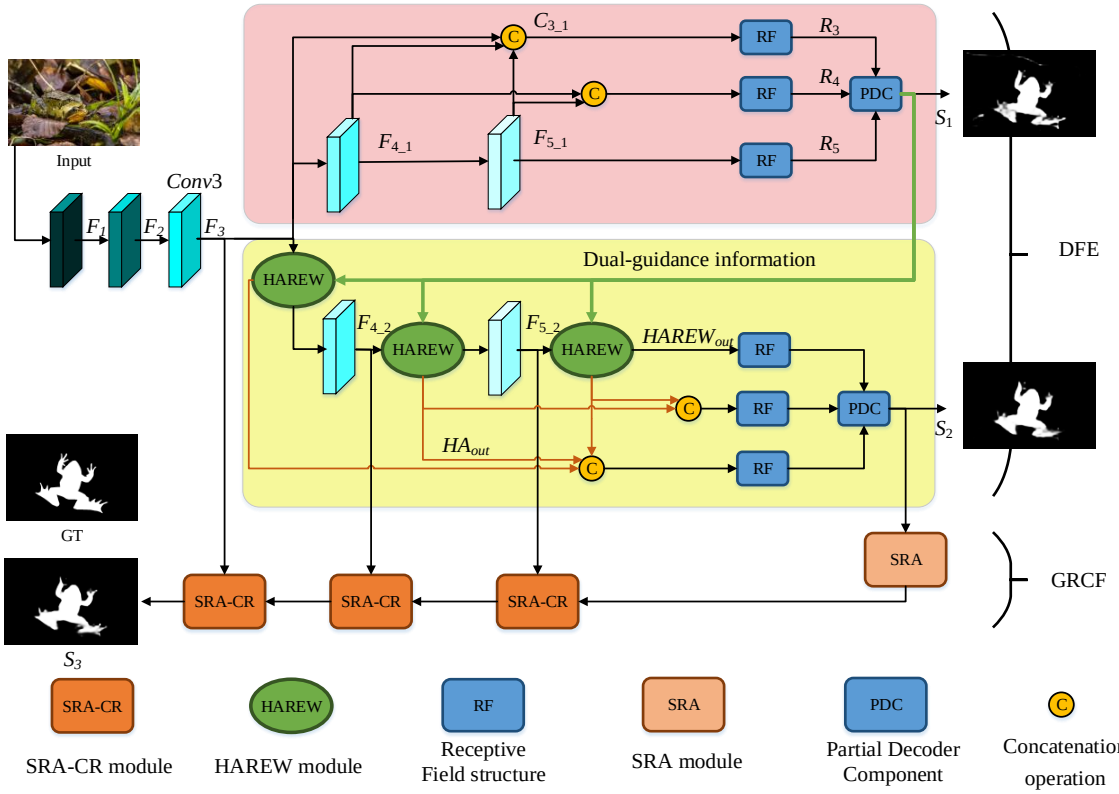


Fig. 1. Overview of our proposed network for COD, which consists of DFE module and GRCF module. In the DFE module, we simulate the mechanism of human vision and design a two-stage detection network. Details of component HAREW can be found in Fig. 2. The components in GRCF moudle see section D. Gradually Refined Cross Fusion and Fig. 3 for details. In the GRCF module, we propose a progressive cross refine mechanism for two-stream information.

The rounded rectangle with a pink background (consists of $Conv4\_1$ and $Conv5\_1$) indicates the first stage observation, and the light yellow rounded rectangle (consists of $Conv4\_2$ and $Conv5\_2$) indicates the second stage observation. In each stage, in order to provide more accurate information and a larger receptive field, we introduce RF structure on the basis of comprehensive consideration of feature reuse. The design of the RF [22] structure is theoretically consistent with our visual mechanism (the process from coarse to refine when detecting objects). Next, we employ PDC [23] components to aggregate features of different levels in parallel to generate the camouflage maps of the first stage. Compared with the first stage, in the second stage, we employ more refinement operations to optimize the detection results, including holistic attention and residual attention guidance. We integrate these operations into the Holistic Attention-Reverse-Weighted addition (HAREW) module. In the GRCF module, inspired by the conventional UNet [35] structure, we use the local information of the peer layer (from DFE) and the high-level information of the GRCF module to generate the final camouflage map.

The differences between our model and SINet [22] are as follows: 1) Our proposed model ($D^2C$-Net) adopts a U-Net structure, while SINet does not. 2) In the first stage of the DFE module, we only used three RF modules (low-level information is not considered), while SINet used four. 3) The first stage and second stage of the DFE module are symmetrical, and two-direction guidance information is added. 4) $D^2C$-Net includes a self-refine attention unit and a cross-refine unit, while SINet does not.

### C. Dual-branch Features Extraction

From [23], we can know that $Conv3$ can maintain clear object boundary information. Compared with the first two layers, this layer contains less redundant information. Therefore, we also employ $Conv3$ as the bifurcation point for two-stage prediction. As shown in Fig. 1, in the first stage, the position and shape of camouflaged objects are roughly detected. In order to refine the results of the first stage, we allow more refinement operations in the second stage, such as attention mechanism and guidance information. In particular, given an input $X^{C*H*W}$ (C, H and W represent the number of channels, height, width of the input image, respectively.), we can obtain a rough feature map of the corresponding level, which is represented by $F_i$, $(i = 1, ..., 5)$. When $i = 4, 5$, the output feature maps of the first stage are represented by $F_{i\_1}$, and the output of the second stage is represented by $F_{i\_2}$.

In the first stage, considering that features of different levels contain various information and in order to make full use of the features of all levels after the bifurcation point, we employ dense concatenation to aggregate the features of the last three levels. We aggregate the current level with the higher level. Specifically, when $i = 3$, we first use upsampling to resize the feature maps of the fourth and fifth layers to the same resolution as the third layer. Second, we aggregate the three-level feature maps in the channel direction through the concatenation operation (the output it denoted by $C_{3\_1}$), and then feed it into the RF module, the output of RF is denoted by $R_3$. When $i = 4, 5$, we generate $R_4$ and $R_5$ in the same way.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

Next, we employ the partial decoder to combine the features $R_i$, $(i = 3, 4, 5)$ to generate the first stage camouflage map $S_1$. The process can be expressed by the following formula:

$$S_1 = PDC(R_3, R_4, R_5) \tag{1}$$

where $PDC(\cdot)$ denotes the partial decoder, the details of $PDC(\cdot)$ can be found in [23].

Since the camouflage maps in the first stage are coarse, we need to detect the camouflage scene again and further optimize $S_1$. This is also consistent with the process of our human visual mechanism looking for camouflaged objects in camouflage scenes. As shown in Fig. 1, in the second stage, the overall structure of the network is roughly the same as in the first stage and the output is denoted by $S_2$. The difference is that we enhance the camouflage map of the corresponding level by introducing guidance information from camouflage maps of the first stage. The guidance information includes forward holistic attention guidance and reverses attention guidance, which is defined as HAREW module.

Specifically, our HAREW module contains two inputs (the guidance information from $S_1$ and the features from the current convolution layer) and two outputs (one for the convolution layer of the next level, and one for the concatenation operation into the RF structure). As shown in Fig. 2, HAREW contains three parts (holistic attention, residual attention guidance and weighted addition).

As mentioned above, we need to use the information from the first stage to make a guiding prediction for the second stage. In this way, we can obtain more accurate camouflage maps than the first stage. First, we employ holistic attention to expand the coverage area of the initial camouflage map in order to improve the effectiveness of the initial camouflage map. The whole process can be expressed by the following formulas:

$$EG_{out} = MAX\left(N\left(GConv\left(\sigma\left(\nabla S_1\right), k\right)\right)\right) \tag{2}$$

$$HA_{out} = mul\left(EG_{out}, F_i\right) \tag{3}$$

where $\sigma(\cdot)$, $mul(\cdot)$, $\nabla$ represent the sigmoid function, element-wise multiplication and down-sampling, respectively. $GConv(\cdot)$ represents the typical Gaussian convolution operation with kernel $k = 32$ and zero bias, which can provide the effect of Gaussian blur, to obtain more effective camouflage maps. $N(\cdot)$ represents the normalization. $MAX(\cdot)$ represents the maximum function that aims to enhance the initial camouflage map $S_1$. $EG_{out}$ and $HA_{out}$ represent the enhanced camouflage maps and the output of holistic attention, respectively. This process can be called forward guidance prediction.

Next, we introduce a reverse attention mechanism to optimize forward camouflaged object detection by erasing the camouflage area of the current prediction, the network can explore more complementary details. This is consistent with the human visual mechanism that repeated optimization and confirmation. In particular, we can obtain the reverse attention weight by the following formula:

$$Re = 1 - \sigma(\nabla S_1) \tag{4}$$

Then, we use reverse attention weight to refine the output of holistic attention.

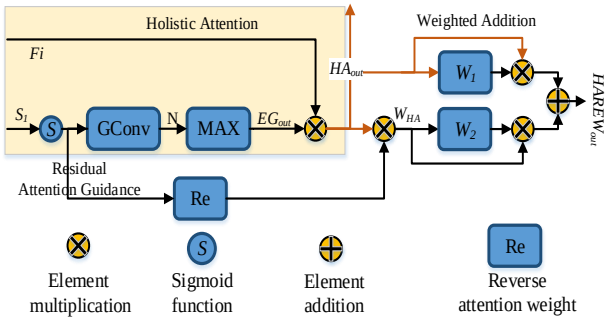


Fig. 2. The detail of our HAREW module. Re, $W_1$ and $W_2$ can be found in, Eq. (4), Eq. (8), Eq. (9), respectively.

$$W_{HA} = mul(HA_{out}, Re) \tag{5}$$

where $W_{HA}$ represents the result generated by reverse attention guidance. This process can be called reverse guidance prediction.

In the process of prediction, we can not guarantee that all pixels are predicted correctly, in order to minimize the influences of this phenomenon, we use weighted addition to fully integrate the forward guidance process and the reverse attention guidance process. The process can be expressed by the following formulas:

$$W_a = PCS(HA_{out}) \tag{6}$$

$$W_b = PCS(W_{HA}) \tag{7}$$

$$W_1 = \frac{W_a}{W_a + W_b} \tag{8}$$

$$W_2 = \frac{W_b}{W_a + W_b} \tag{9}$$

$$HAREW_{out} = mul(HA_{out}, W_1) + mul(W_{HA}, W_2) \tag{10}$$

where $PCS(\cdot)$ represents the operation with $Pooling -> Conv -> Softmax$, $HAREW_{out}$ represents the output of the HAREW module.

### D. Gradually Refined Cross Fusion

After the above process, we can obtain a relatively accurate camouflage map than the last stage. In order to further optimize the detection results of the previous stage, we employ the UNet structure to incorporate local information to refine the camouflage map $S_2$ generated by the DFE module. Unlike the traditional UNet structure [35], we first optimize the camouflage map $S_2$ generated by the decoder through the self-refine-attention unit (SRA). The features of the peer layer are further optimized through the cross refinement unit (CR). By combining two optimization strategies, accurate detection results can be obtained. SRA and CR are the main components of the GRCF module, namely SRA-CR.

As shown in Fig. 3, $F_i$ represents the features from peer layers. As for SRA unit, we first adjust the channel dimension to 256 with a $3 \times 3$ convolution layer (the output is denoted by $SR_{ori}$) and then perform parallel operations through two independent convolutions with the same kernel size ($3 \times 3$). In order to capture more complete information to the greatest extent, we multiply the output of one of the convolutions with

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

$SR_{ori}$. Then, the output is fused with the result of the other convolution unit by element-addition. In addition, considering that not each pixel is representative in the channel dimension, we introduce a new channel attention, that is, to maximize the feature through the max function in the channel dimension, namely, max-attention. The function of SRA can be expressed by the following formulas:

$$SR_{mid} = mul(Conv_a(SR_{ori}), SR_{ori}) + Conv_b(SR_{ori}) \tag{11}$$

$$SRA_{out} = SR_{mid} + MA(SR_{ori}) \tag{12}$$

where $Conv_a$ and $Conv_b$ all represent the convolution layers with the kernel size is 3. $SR_{mid}$ represents the output without max-attention, $MA(\cdot)$ represents the function that max-attention, and the output of SRA unit is denoted by $SRA_{out}$.

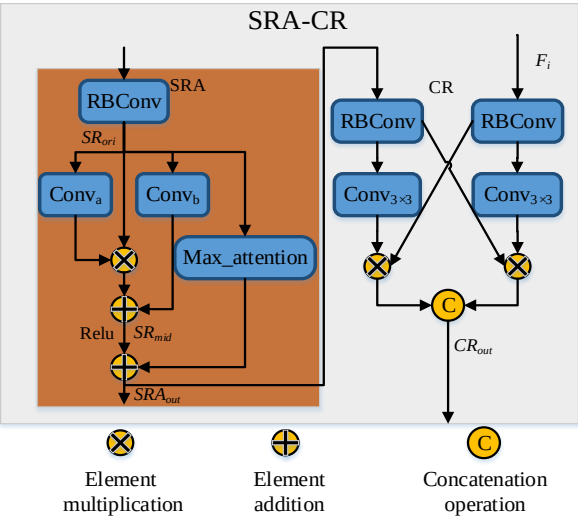


Fig. 3. The detail of SRA-CR module.

Inspired by [36], we propose a CR unit to merge the local information $F_i$ $(i = 3, 4, 5)$ of the peer layer with the generated camouflage map. The process can be expressed by the following formulas:

$$CR_{11} = RBConv(SRA_{out}) \tag{13}$$

$$CR_{21} = RBConv(F_i) \tag{14}$$

$$CR_{12} = mul(Conv(CR_{11}), CR_{21}) \tag{15}$$

$$CR_{22} = mul(Conv(CR_{21}), CR_{11}) \tag{16}$$

$$CR_{out} = Concat(CR_{12}, CR_{22}) \tag{17}$$

where $CR_{1j}$ $(j = 1, 2)$ and $CR_{2j}$ represent the intermediate output, and $RBConv(\cdot)$ denotes the operation with $Conv -> BN -> Relu$. $Concat(\cdot)$ represents the operation with concatenation.

### E. Loss Function

Binary cross entropy loss (BCE) is widely used to measure the difference between the prediction and label, which pays more attention to the pixel-level error and does not consider the correlation between each pixel. IOU loss is often used in segmentation tasks, which aims to optimize the global structure. Inspired by [37]–[39], we employ weighted BCE loss and weighted IOU loss as the combined loss, which can be defined by the following formula:

$$L = L^w_{IOU} + L^w_{BCE} \tag{18}$$

Compared with the standard BCE loss and IOU loss, $L^w_{IOU}$ and $L^w_{BCE}$ (the definitions can be found in [37], [38]) pay more attention to hard samples. We employ the same parameter definition and setting as [37], [38], and the effectiveness has been proved. In this paper, we adopt deep supervision for the three stages, and their locations are shown in Fig. 1. Therefore, the final loss can be expressed by the following formula:

$$L_{all} = L_{S1} + L_{S2} + L_{S3} \tag{19}$$

$L_{S1}$, $L_{S2}$ and $L_{S3}$ represent the loss between the output of each stage and GT, respectively.

## IV. Experiments and Results

In this section, we will describe the benchmark dataset in the COD field, the evaluation metrics, experimental setting, comparisons with other models, and ablation study in detail.

### A. Datasets

CAMO [21] was proposed in 2019 as the first dataset for camouflaged object segmentation, which consists of 1250 images and is divided into two categories (artificially camouflaged objects and naturally camouflaged objects). 1000 images are used for training and the rest for testing.

CHAMELEON [24] is collected from Google using the keyword Camouflaged Animal, which contains only 76 images and manually object-level labels.

COD10K [22] is the latest and largest data set for camouflaged object detection, with pixel-level annotation. The dataset contains 10K images, divided into five superclasses, and 6000 images were randomly selected for training and the rest for testing.

### B. Evaluation Metrics

**F-measure** [40] is widely used to evaluate the similarity of the two images (the output generated by the model and the corresponding GT of the input), which can be formulated as:

$$F_\beta = \frac{\left(\beta^2 + 1\right) PR}{\beta^2 P + R} \tag{20}$$

According to [41], $\beta^2$ is set to 0.3 to balance $P$ and $R$, where $P$ represents precision and $R$ denotes recall.

**S-measure** [42] is to measure the structural similarity between the prediction map and the corresponding label, which can be defined as:

$$S = \alpha S_o + (1 - \alpha) S_r \tag{21}$$

where $S_o$ represents structural similarity measurement based on object-level, and $S_r$ represents a region-based similarity measurement. According to [42], we set $\alpha$ as 0.5.

**E-measure** [43] is proposed for binary map evaluation, and takes into account local information and global information, defined as:

$$E = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi_{FM}(x, y) \tag{22}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

where $\phi_{FM}$ ($x$, $y$ is the corresponding pixel) represents the enhanced alignment term, which is used to capture image-level statistics and pixel-level matching. $W$ and $H$ represent the width and height of the images that we input, respectively.

**MAE** is to measure the average absolute difference between the output by the model and the ground-truth of input. MAE is a pixel-level error evaluation index, defined as:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x,y) - G(x,y)| \tag{23}$$

where $H$ and $W$ represent the heigh and width of the input. $S$ denotes the generated map, and $G$ denotes the ground truth.

**PR curve** is drawn with Precision and Recall as variables. Recall is the transverse coordinate and Precision is the vertical coordinate. We can employ different thresholds to divide the foreground and background of generated maps, and then we can calculate the corresponding Precision score and Recall score. Finally, we can draw a PR curve to describe the performance of the model by combining the obtained scores.

### C. Experimental Setting

We employ Res2Net [34] as our backbone and the parameters are pre-trained on ImageNet [53]. Our proposed model was implemented on Google Cloud Platform in Pytorch toolbox, and an NVIDIA Tesla P100 was utilized for acceleration. We train our proposed model on the dataset of CAMO+COD10K (contains 4040 images), which comes from https://github.com/DengPingFan/SINet. For training, the spatial resolution of input images is resized into 320*320. Our model employs Adam as optimizer with the initial learning rate 0.001 and the batchsize is set to 12. The code will be available at https://github.com/MS-KangWang.

### D. Comparison with start-of-the-art (SOTA)

As far as we know, there is almost no publicly available code in the field of COD, so we select several classic depth models as the comparison model, which contains PiCANet [44], UNet++ [45], ANet [21], BASNet [46], CPD-R [23], HTC [47], MSRCNN [48], FPANet [49], PoolNet [50], EGNet-R [51], GCPANet [52], SCRN [36], SINet [22]. The selection criteria are the same as SINet [22], including achieving SOTA performance in the special field, classical architectures and recently published. As for camouflage maps, parts of compared models are from https://github.com/DengPingFan/SINet, and parts are retrained and tested based on the open-source code (based on the recommended parameters).

*1) Quantitative Evaluation:* As listed in TABLE I and TABLE II, our model achieves the best score in all datasets under four public map quality evaluation indicators. Specifically, for the current largest camouflaged object dataset COD10K, the $Fmean$ score of our model achieves the highest score of 0.720, which is 0.086 (13.56%) higher than second ranked model. Besides, $MAE$ also decreases by 0.014 (27.45%). For its subset, our D$^2$C-Net achieves the best performance compared to any other model. As for the other two datasets, CAMO and CHAMELEON, our model achieves still different degrees of improvement compared with other models. Overall, our D$^2$C-Net achieves SOTA performance.

*2) Qualitative Evaluation:* As shown in Fig. 4, in order to highlight the performance of our proposed model, we make a series of more detailed visual contrast experiments, and provide the corresponding visual contrast maps. The red rectangle in Fig. 4 marks the difference between the detection results of different models and the corresponding GT. From the figure, we can see that the detection result of our model is closer to GT. In other words, our model performs better than other comparative models. Specifically, from the first row and the second row, we can see that the results of our model are almost identical to GT. However, the other models are missing in varying degrees or the detection results of camouflage objects are not complete. In the fourth row, our model can detect the camouflaged object completely, while SINet [22] can only detect the boundary of the target roughly. Although SCRN [36] and GCPANet [52] can detect the main part of the target, the edge of the object is not clear and has a serious trailing phenomenon. In addition, as can be seen from Fig. 5, the overall performance of our model is better than other models. In general, our model can detect camouflage objects more accurately than other models, whether in the target body or its boundary, which may benefit from our optimization strategy (gradually refined cross fusion, see section III $D$ for details).

We also provide some examples of artificial camouflage objects. From Fig. 6, we can know that our model can detect camouflage objects more accurately than SINet. This also proves that our model can achieve good performance on artificial camouflage objects.

### E. Ablation Study

In this section, in order to prove the effectiveness of each component in our model, we have conducted experiments on three benchmark datasets. The details are summarized as follows.

From TABLE III, we can see that the performance of the model can be further improved when combining different components, especially on COD10k (the largest data set at present). Compared with model B, although the performance growth of the model C is not obvious on the CAMO dataset (small dataset), it has been improved on other datasets. From the overall comprehensive performance (D), this is the best compared with any stage. Comparing with the backbone on COD10K dataset, the mean F-measure of our model increased 0.079 (12.32%), and MAE score decreased 0.011 (22.92%).

We also provide intermediate results in different stages to prove the effectiveness of our proposed model. As shown in Fig. 6, when we only employ the backbone (A) to extract features and the output of the last layer is upsampled as the final camouflage maps, our network can roughly lock the position of the camouflaged object. However, shadow exists at the edge of the generated camouflage maps of model (A), which is unsatisfactory. Next, when we combine the structure of the first stage (B) as illustrated in Fig. 1, the detected coverage area of camouflage maps is more accurate. However, there are still some objects that are wrongly detected or incomplete, such as the fifth row and third row. Then, we introduce dual-guidance information, and through the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

TABLE I

QUANTITATIVE COMPARISON BETWEEN OUR MODEL AND OTHERS UNDER FOUR PUBLIC IMAGE QUALITY EVALUATION INDEX, WHICH CONTAINS S-MEASURE ($S$), MEAN ABSOLUTE ERROR ($MAE$), MEAN E-MEASURE ($E_\phi$) AND MEAN F-MEASURE ($F_\alpha$), ↑ DENOTES THAT THE HIGHER THE BETTER AND ↓ DENOTES THAT THE LOWER THE BETTER. THE BEST SCORES WERE MARKED RED, FOLLOWED BY GREEN AND BLUE. THE EVALUATION CODE IS FROM HTTPS://GITHUB.COM/DENGPINGFAN/CODTOOLBOX/.

| model | CAMO | | | | CHAMELEON | | | | COD10K | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ |
| 2018-CVPR-PiCANet [44] | 0.609 | 0.156 | 0.584 | 0.419 | 0.769 | 0.085 | 0.749 | 0.615 | 0.649 | 0.090 | 0.643 | 0.411 |
| 2018-DLMIA-UNet++ [45] | 0.599 | 0.149 | 0.653 | 0.461 | 0.695 | 0.094 | 0.762 | 0.557 | 0.623 | 0.086 | 0.672 | 0.409 |
| 2019-CVIU-ANet [21] | 0.682 | 0.126 | 0.685 | 0.541 | * | * | * | * | * | * | * | * |
| 2019-CVPR-BASNet [46] | 0.618 | 0.159 | 0.661 | 0.475 | 0.687 | 0.118 | 0.721 | 0.528 | 0.634 | 0.105 | 0.678 | 0.417 |
| 2019-CVPR-CPD-R [23] | 0.726 | 0.115 | 0.729 | 0.613 | 0.853 | 0.052 | 0.866 | 0.752 | 0.747 | 0.059 | 0.770 | 0.581 |
| 2019-CVPR-HTC [47] | 0.476 | 0.172 | 0.442 | 0.206 | 0.517 | 0.129 | 0.489 | 0.236 | 0.548 | 0.088 | 0.520 | 0.253 |
| 2019-CVPR-MSRCNN [48] | 0.617 | 0.133 | 0.669 | 0.527 | 0.637 | 0.091 | 0.686 | 0.505 | 0.641 | 0.073 | 0.706 | 0.478 |
| 2019-CVPR-PFANet [49] | 0.659 | 0.172 | 0.622 | 0.464 | 0.679 | 0.144 | 0.648 | 0.450 | 0.636 | 0.128 | 0.618 | 0.375 |
| 2019-CVPR-PoolNet [50] | 0.702 | 0.129 | 0.698 | 0.563 | 0.776 | 0.081 | 0.779 | 0.632 | 0.705 | 0.074 | 0.713 | 0.500 |
| 2019-ICCV-EGNet [51] | 0.732 | 0.104 | 0.768 | 0.647 | 0.848 | 0.050 | 0.870 | 0.750 | 0.737 | 0.056 | 0.779 | 0.573 |
| 2020-AAAI-GCPANet [52] | 0.749 | 0.111 | 0.745 | 0.630 | 0.850 | 0.056 | 0.846 | 0.733 | 0.766 | 0.057 | 0.773 | 0.599 |
| 2020-ICCV-SCRN [36] | 0.745 | 0.105 | 0.743 | 0.644 | 0.866 | 0.046 | 0.869 | 0.770 | 0.776 | 0.051 | 0.788 | 0.625 |
| 2020-CVPR-SINet [22] | 0.751 | 0.100 | 0.771 | 0.675 | 0.869 | 0.044 | 0.891 | 0.790 | 0.771 | 0.051 | 0.806 | 0.634 |
| **Ours** | **0.774** | **0.087** | **0.818** | **0.735** | **0.889** | **0.030** | **0.939** | **0.848** | **0.807** | **0.037** | **0.876** | **0.720** |

TABLE II

QUANTITATIVE COMPARISON BETWEEN OUR MODEL AND OTHERS UNDER FOUR PUBLIC IMAGE QUALITY EVALUATION INDEX, WHICH CONTAINS S-MEASURE ($S$), MEAN ABSOLUTE ERROR ($MAE$), MEAN E-MEASURE ($E_\phi$) AND MEAN F-MEASURE ($F_\alpha$), ↑ DENOTES THAT THE HIGHER THE BETTER AND ↓ DENOTES THAT THE LOWER THE BETTER. THE BEST SCORES WERE MARKED RED, FOLLOWED BY GREEN AND BLUE. AMPHIBIAN, AQUATIC, TERRESTRIAL AND FLYING ARE THE SUBCLASSES OF COD10K.

| Model | Amphibian | | | | Aquatic | | | | Terrestrial | | | | Flying | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ | $S\uparrow$ | $MAE\downarrow$ | $E_\phi\uparrow$ | $F_\alpha\uparrow$ |
| 2018-CVPR-PiCANet [44] | 0.704 | 0.086 | 0.689 | 0.494 | 0.629 | 0.120 | 0.623 | 0.423 | 0.625 | 0.084 | 0.628 | 0.359 | 0.677 | 0.076 | 0.663 | 0.440 |
| 2018-DLMIA-UNet++ [45] | 0.677 | 0.079 | 0.725 | 0.496 | 0.599 | 0.121 | 0.659 | 0.418 | 0.593 | 0.081 | 0.637 | 0.340 | 0.659 | 0.068 | 0.708 | 0.455 |
| 2019-CVPR-BASNet [46] | 0.708 | 0.087 | 0.741 | 0.535 | 0.620 | 0.134 | 0.666 | 0.431 | 0.601 | 0.109 | 0.645 | 0.350 | 0.664 | 0.086 | 0.710 | 0.454 |
| 2019-CVPR-CPD-R [23] | 0.794 | 0.051 | 0.823 | 0.659 | 0.739 | 0.082 | 0.770 | 0.606 | 0.714 | 0.058 | 0.735 | 0.516 | 0.777 | 0.046 | 0.796 | 0.616 |
| 2019-CVPR-HTC [47] | 0.606 | 0.088 | 0.596 | 0.365 | 0.507 | 0.129 | 0.494 | 0.223 | 0.530 | 0.078 | 0.484 | 0.196 | 0.582 | 0.070 | 0.558 | 0.308 |
| 2019-CVPR-MSRCNN [48] | 0.722 | 0.055 | 0.784 | 0.613 | 0.614 | 0.107 | 0.685 | 0.464 | 0.611 | 0.070 | 0.671 | 0.417 | 0.674 | 0.058 | 0.742 | 0.522 |
| 2019-CVPR-PFANet [49] | 0.690 | 0.119 | 0.661 | 0.460 | 0.629 | 0.162 | 0.614 | 0.404 | 0.609 | 0.123 | 0.600 | 0.323 | 0.657 | 0.113 | 0.632 | 0.393 |
| 2019-CVPR-PoolNet [50] | 0.766 | 0.064 | 0.769 | 0.598 | 0.689 | 0.102 | 0.705 | 0.507 | 0.677 | 0.070 | 0.688 | 0.442 | 0.733 | 0.062 | 0.733 | 0.534 |
| 2019-ICCV-EGNet [51] | 0.788 | 0.048 | 0.837 | 0.660 | 0.725 | 0.080 | 0.775 | 0.597 | 0.704 | 0.054 | 0.748 | 0.509 | 0.768 | 0.044 | 0.803 | 0.607 |
| 2020-AAAI-GCPANet [52] | 0.819 | 0.048 | 0.840 | 0.681 | 0.766 | 0.075 | 0.779 | 0.635 | 0.734 | 0.058 | 0.736 | 0.533 | 0.790 | 0.045 | 0.795 | 0.627 |
| 2020-ICCV-SCRN [36] | 0.817 | 0.043 | 0.836 | 0.697 | 0.767 | 0.071 | 0.787 | 0.649 | 0.746 | 0.051 | 0.756 | 0.566 | 0.803 | 0.040 | 0.812 | 0.656 |
| 2020-CVPR-SINet [22] | 0.827 | 0.042 | 0.866 | 0.724 | 0.758 | 0.073 | 0.803 | 0.650 | 0.743 | 0.050 | 0.778 | 0.578 | 0.798 | 0.040 | 0.828 | 0.662 |
| **Ours** | **0.848** | **0.032** | **0.911** | **0.783** | **0.805** | **0.053** | **0.873** | **0.744** | **0.773** | **0.039** | **0.848** | **0.657** | **0.835** | **0.027** | **0.902** | **0.755** |

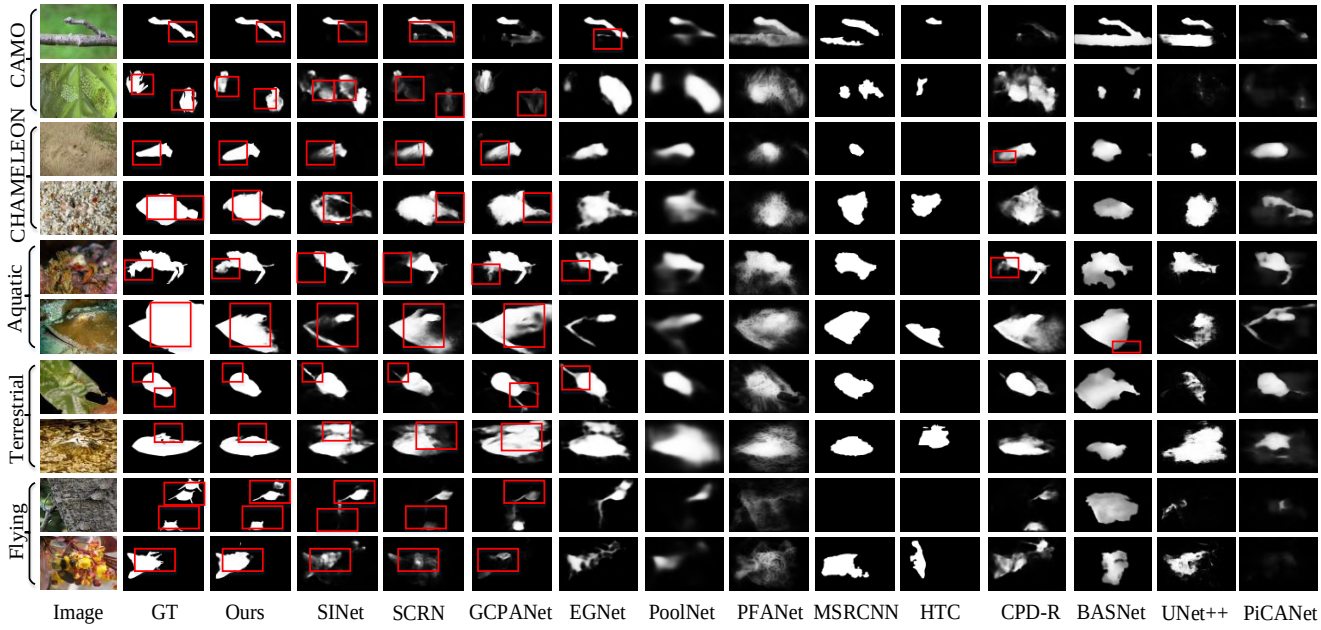


Fig. 4. Visual comparison of our proposed model and others on three benchmark datasets. The first column, the second column and the third column represent the input image, Ground Truth(GT) and our model results, respectively. From the figure, we can see that our results are closest to GT. In addition, the completely black results in the figure indicate the corresponding model does not detect any objects.

combination (C) of holistic attention and residual attention guidance, we can further modify the detection results. As can be seen from the sixth and seventh rows of Fig. 7, the redundant background is suppressed and the missing details are completely supplemented. Finally, when combined with GRCF module (D), it makes full use of some supplementary

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

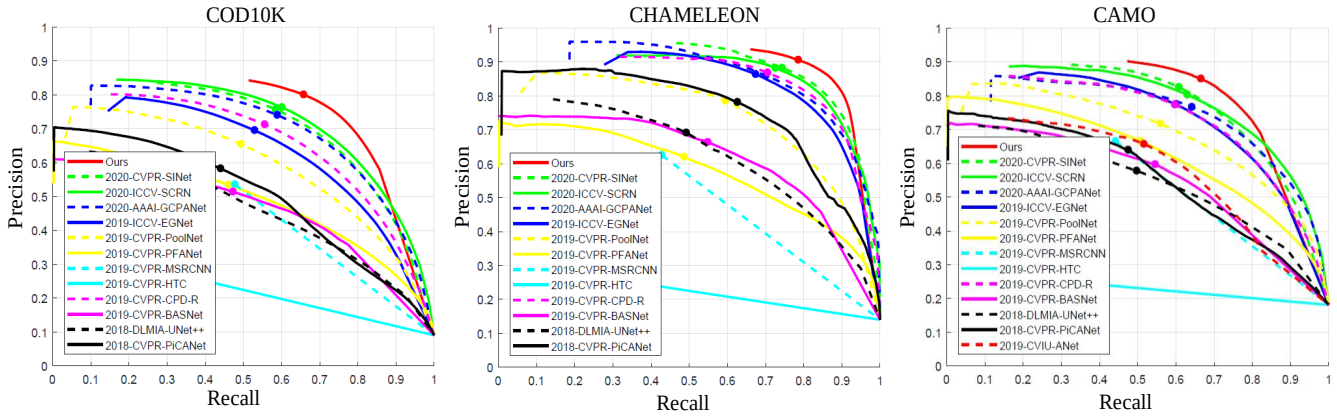


Fig. 5. The PR Curves of our model and others on three benchmark datasets.

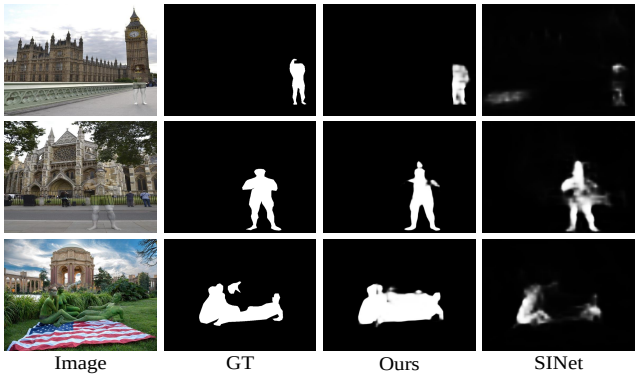


Fig. 6. The performance comparison on artificial camouflage objects.

TABLE III
THE ABLATION STUDY OF OUR PROPOSED NETWORK ON THREE DATASET. A, B, C AND D REPRESENT THE MODULE BACKBONE, A+STAGE1, B+STAGE2 AND C+GRCF, RESPECTIVELY.

| Ablation Study | | A | B | C | D |
|---|---|---|---|---|---|
| CAMO | $S\uparrow$ | 0.758 | 0.760 | 0.760 | **0.774** |
| | $M\downarrow$ | 0.098 | 0.097 | 0.094 | **0.087** |
| | $F_\alpha\uparrow$ | 0.697 | 0.706 | 0.705 | **0.735** |
| CHAMELEON | $S\uparrow$ | 0.857 | 0.881 | 0.888 | **0.889** |
| | $M\downarrow$ | 0.044 | 0.035 | 0.032 | **0.030** |
| | $F_\alpha\uparrow$ | 0.784 | 0.826 | 0.841 | **0.848** |
| COD10K | $S\uparrow$ | 0.772 | 0.793 | 0.801 | **0.807** |
| | $M\downarrow$ | 0.048 | 0.043 | 0.038 | **0.037** |
| | $F_\alpha\uparrow$ | 0.641 | 0.687 | 0.702 | **0.720** |

texture information provided by peer-to-peer features, so that the detection results are further improved. As can be seen from the last column in Fig. 7, the camouflage maps of the final model (D) are closer to GT.

We conduct further analysis (as listed in TABLE IV) of our method and SINet (published COD model with the best performance) from the aspects of model size (number of parameters) and inference speed (frames per seconds). Since there is a trade-off between model complexity and performance, we would like to emphasize that, by introducing a two-stage, dual-guidance and cross-refine network based on human visual mechanism, our method outperforms all the counterparts by a large margin (see Table I and Table II for details), which indicates that the main purpose of this work has been fulfilled. Of equal importance, although little discussed in this work, is the integrated solution considering both the accuracy and

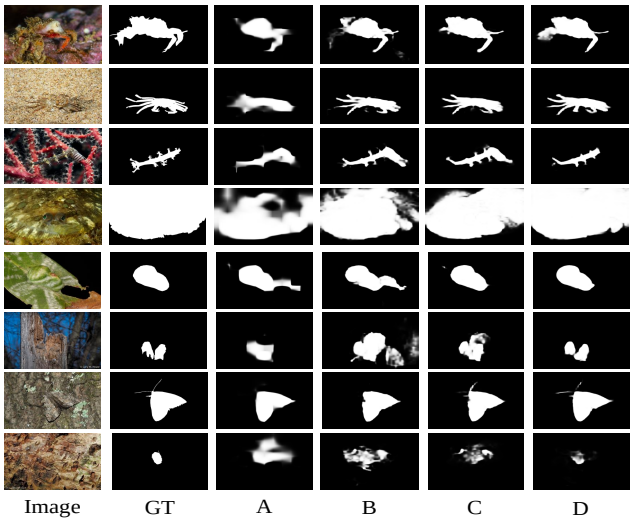


Fig. 7. Visual comparisons of D$^2$C-Net in four stages. A, B, C and D denotes the backbone, A+stage1, B+stage2 and C+GRCF, respectively.

TABLE IV
COMPARISON OF ALGORITHM EFFICIENCY.

| Model | Parameters | Frames per second | | |
|---|---|---|---|---|
| | | CAMO | CHAMELEON | COD10K |
| SINet | 48946851 | 29.28 | 29.35 | 29.75 |
| Ours | 55020186 | 17.35 | 17.00 | 16.93 |

efficiency, which is going to be the main focus of future work.

## V. POTENTIAL APPLICATIONS

Camouflaged object detection can be applied to a bit of fields, such as surface defect detection, medical segmentation, wildlife monitoring/detection and underwater biological detection. In this section, we have carried out some experiments to verify the possibility of our model applied to other fields (the first two applications). As shown in Fig. 8, we can see that our model achieves satisfactory detection results.

On the one hand, when a system is equipped with a camouflaged object detection function, the system can detect whether there are defects on the rail surface. The red box in the figure indicates the defect location and detection results. On the left of the blue line in Fig. 8, we can see that the location of the flaw can be well detected. On the other hand, we apply our model to the polyp segmentation, and the detection results are shown on the right of the blue line in Fig. 8. It is worth

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

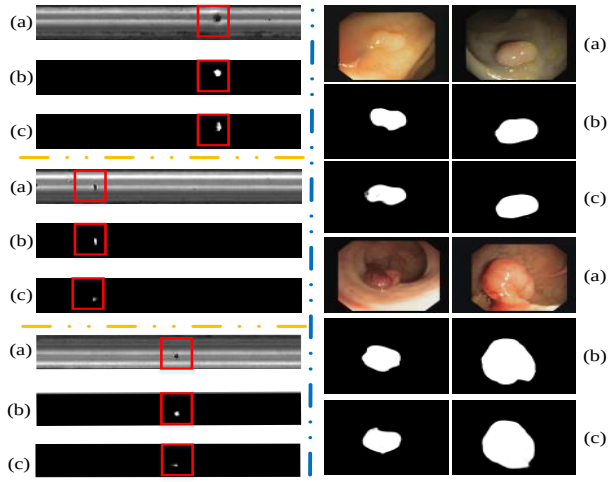


Fig. 8. Potential applications. (a), (b), (c) represent input images, GT and the output of our model, respectively. The left side of the blue line is the rail surface defect detection, and the right is the polyp detection.

mentioning that, our training sets do not contain images related to rail defect detection and polyp segmentation.

## VI. Conclusion

In this paper, we propose a novel model for COD detection, which contains the DFE module and the GRCF module. In the DFE module, we use two stages to detect camouflaged objects to simulate the human visual mechanism. The first stage can be understood as a rough prediction, while the second stage is to make an accurate guidance forecast on the basis of the first stage. In the GRCF module, we optimize the detection results of DFE by using peer-to-peer characteristics through two strategies: self-refine attention and cross-refine unit. Compared with others, our model can achieve the SOTA performance when we combine the proposed module. Besides, in order to prove the effectiveness of our model, we carry on a series of experiments on three datasets. From the experimental results, we can see that the performance of our model is gradually improved when combining different modules. In all, the effectiveness and superiority of our model have been proved. Furthermore, we hope that our model can be applied to more fields, such as industrial defect detection, medical image segmentation and detection.

## References

[1] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.*, pp. 186–202, 2018.

[2] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *arXiv preprint arXiv:2101.07663*, 2021.

[3] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *arXiv preprint arXiv:2009.03075*, 2020.

[4] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suàrez, M. Jagersand, and L. Shao, "Boundary-aware segmentation network for mobile and web applications," *arXiv preprint arXiv:2101.04704*, 2021.

[5] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *arXiv preprint arXiv:2102.10274*, 2021.

[6] Y. Lv, J. Zhang, Y. Dai, L. Aixuan, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[7] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2021.

[8] D. Osorio and I. C. Cuthill, "Camouflage and perceptual organization in the animal kingdom," *The Oxford handbook of perceptual organization. Oxford library of psychology. Oxford University Press, Oxford*, pp. 843–862, 2015.

[9] S. Merilaita, N. E. Scott-Samuel, and I. C. Cuthill, "How camouflage works," *Philos. Trans. R. Soc. Lond., B, Biol. Sci.*, vol. 372, no. 1724, p. 20160341, 2017.

[10] I. C. Cuthill, M. Stevens, J. Sheppard, T. Maddocks, C. A. Párraga, and T. S. Troscianko, "Disruptive coloration and background pattern matching," *Nature*, vol. 434, no. 7029, pp. 72–74, 2005.

[11] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Pranet: Parallel reverse attention network for polyp segmentation," *MICCAI*, vol. 12266, pp. 263–273, 2020.

[12] Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, C.-W. Zhao, and M.-M. Cheng, "Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation," *IEEE T. Image Process.*, 2021.

[13] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao, "Inf-net: Automatic covid-19 lung infection segmentation from ct images," *IEEE T. Med. Imag.*, vol. 39, no. 8, pp. 2626–2637, 2020.

[14] X. Le, J. Mei, H. Zhang, B. Zhou, and J. Xi, "A learning-based approach for surface defect detection using small image datasets," *Neurocomputing*, vol. 408, pp. 112–120, 2020.

[15] T. Lidbetter, "Search and rescue in the face of uncertain threats," *Eur. J. Oper. Res.*, vol. 285, no. 3, pp. 1153–1160, 2020.

[16] Q. Wang, Z. Yuan, Q. Du, and X. Li, "Getnet: A general end-to-end two-dimensional cnn framework for hyperspectral image change detection," *arXiv preprint arXiv:1905.01662*, 2019.

[17] N. U. Bhajantri and P. Nagabhushan, "Camouflage defect identification: A novel approach," in *Int. Conf. Inf. Technol.*, pp. 145–148, 2007.

[18] L. Song and W. Geng, "A new camouflage texture evaluation method based on wssim and nature image features," in *Int. Conf. Multimedia and Technol.*, pp. 1–4, 2010.

[19] P. Sengottuvelan, A. Wahi, and A. Shanmugam, "Performance of de-camouflaging through exploratory image analysis," pp. 6–10, 2008.

[20] Y. Zheng, X. Zhang, F. Wang, T. Cao, M. Sun, and X. Wang, "Detection of people with camouflage pattern via dense deconvolution network," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 29–33, 2018.

[21] T. Le, T. V. Nguyen, Z. Nie, M. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," vol. 184, pp. 45–56, 2019.

[22] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2774–2784, 2020.

[23] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3907–3916, 2019.

[24] Y. Shen, R. Ji, S. Zhang, W. Zuo, and W. Yan, "Generative adversarial learning towards fast weakly supervised detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 5764–5773, 2018.

[25] S. K. Singh, C. A. Dhawale, and S. Misra, "Survey of object detection methods in camouflaged image," *IERI Procedia*, vol. 4, no. 4, pp. 351–357, 2013.

[26] F. Xue, C. Yong, S. Xu, H. Dong, Y. Luo, and W. Jia, "Camouflage performance analysis and evaluation framework based on features fusion," *Multimed. Tools. Appl.*, pp. 4065–4082, 2016.

[27] P. Sengottuvelan, A. Wahi, and A. Shanmugam, "Performance of de-camouflaging through exploratory image analysis," in *Int. Conf. Emerg. Trends Eng. Technol.*, pp. 6–10, 2008.

[28] J. Yin, Y. Han, W. Hou, and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Procedia Engineering*, vol. 15, no. none, pp. 2201–2205, 2011.

[29] Y. Pan, Y. Chen, Q. Fu, P. Zhang, and X. Xu, "Study on the camouflaged target detection method based on 3d convexity," *Modern Applied ence*, vol. 5, no. 4, pp. 152–157, 2011.

[30] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE T. Pattern Anal. Mach. Intell.*, 2021.

[31] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, and M.-M. Cheng, "Taking a deeper look at co-salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 2919–2929, 2020.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIE.2021.3078379, IEEE Transactions on Industrial Electronics

[32] Q. Wang, T. Han, Z. Qin, J. Gao, and X. Li, "Multitask attention network for lane detection and fitting," *IEEE T. Neural Netw. Learn. Syst.*, 2020.
[33] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," *Eur. Conf. Comput. Vis.*, pp. 236–252, 2018.
[34] S. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. H. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE T. Pattern Anal. Mach. Intell.*, pp. 1–1, 2019.
[35] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *MICCAI*, pp. 234–241. Springer, 2015.
[36] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Int. Conf. Comput. Vis.*, pp. 7263–7272, 2019.
[37] J. Wei, S. Wang, and Q. Huang, "F3net: Fusion, feedback and focus for salient object detection," *AAAI Conf. Art. Intell.*, vol. 34, no. 7, pp. 12 321–12 328, 2020.
[38] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7479–7489, 2019.
[39] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function (in chinese)," *Sci Sin Inform*, 2021.
[40] P. Arbelaez, M. Maire, C. C. Fowlkes, and J. Malik, "Contour detection and hierarchical image segmentation," *IEEE T. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 898–916, 2011.
[41] D. Zhang, J. Han, C. Li, J. Wang, and X. Li, "Detection of co-salient objects by looking deep and wide," *Int. J. Comput. Vis.*, vol. 120, no. 2, pp. 215–232, 2016.
[42] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *Int. Conf. Comput. Vis.*, pp. 4558–4567, 2017.
[43] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Int. Jt. Conf. Artif. Intell.*, pp. 698–704, 2018.
[44] N. Liu, J. Han, and M. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3089–3098, 2018.
[45] Z. Zhou, M. M. R. Siddiquee, N. Tajbakhsh, and J. Liang, "Unet++: A nested u-net architecture for medical image segmentation," *IEEE T. Med. Imag.*, pp. 3–11, 2019.
[46] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 7479–7489, 2019.
[47] K. Chen, J. Pang, J. Wang, Y. Xiong, X. Li, S. Sun, W. Feng, Z. Liu, J. Shi, W. Ouyang *et al.*, "Hybrid task cascade for instance segmentation," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 4974–4983, 2019.
[48] Z. Huang, L. Huang, Y. Gong, C. Huang, and X. Wang, "Mask scoring r-cnn," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 6409–6418, 2019.
[49] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3085–3094, 2019.
[50] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, pp. 3917–3926, 2019.
[51] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *Int. Conf. Comput. Vis.*, pp. 8779–8788, 2019.
[52] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," *AAAI Conf. Art. Intell.*, vol. 34, no. 7, pp. 10 599–10 606, 2020.
[53] Krizhevsky, Alex, Sutskever, Ilya, Hinton, and E. Geoffrey, "Imagenet classification with deep convolutional neural networks." *Adv. Neural Inform. Process. Syst.*, pp. 1106–1114, 2012.

**K** ang Wang is pursuing his master degree at the Northeast Petroleum University, Daqing, China. His current research interests include Co-saliency detection, camouflaged object detection, video saliency detection and deep learning.

**H** ongbo Bi (China, 1979) received his bachelor degree in 2001 and master degree in 2004 respectively from NorthEast Petroleum University both in communications engineering. He received his Ph.D in 2013 from Beijing University of Posts and Telecommunications and worked as a PDF(PostDoc Fellow) in Harbin Engineering University in 2014-2017, worked as a visiting scholar in University of Waterloo (Canada) in 2014-2015. Currently, he is an associate professor in School of Electrical Information Engineering in NorthEast Petroleum University. His main research interests focus on saliency detection, compressive sensing, deep learning, computer vision, signal processing, etc.

**Y** i Zhang received his bachelor degree in 2016 and master degree in 2019 respectively from Southeast University both in biomedical engineering. He is pursuing his PhD degree at INSA Rennes, France. His current research interests include omnidirectional vision, salient object detection and deep learning.

**C** ong Zhang is pursuing her master degree at the Northeast Petroleum University, Daqing, China. Her current research interests include Camouflaged object detection, salient object detection and deep learning.

**Z** iqi Liu is pursuing her master degree at the Northeast Petroleum University, Daqing, China. Her current research interests include RGB-D salient object detection, RGB salient object detection and deep learning.

**S** huang Zheng has participated in the development of mine truck unmanned vehicle project, responsible for the visual recognition part, and participated in the basic algorithm research of visdrone 2018 and visdrone 2019 competitions. At present, he is the CTO of Tianjin chain number technology company, responsible for the research of computer vision algorithm.