Dual-Constraint Coarse-to-Fine Network for Camouflaged Object Detection

Guanghui Yue, *Member, IEEE*, Houlu Xiao, Hai Xie, Tianwei Zhou, *Associate Member, IEEE*, Wei Zhou, Weiqing Yan, Baoquan Zhao, Tianfu Wang, and Qiuping Jiang, *Member, IEEE*

***Abstract*—Camouflaged object detection (COD) is an important yet challenging task, with great application values in industrial defect detection, medical care, etc. The challenges mainly come from the high intrinsic similarities between target objects and background. In this paper, inspired by the biological studies that object detection consists of two steps, i.e., search and identification, we propose a novel framework, named DCNet, for accurate COD. DCNet explores candidate objects and extra object-related edges through two constraints (object area and boundary) and detects camouflaged objects in a coarse-to-fine manner. Specifically, we first exploit an area-boundary decoder (ABD) to obtain initial region cues and boundary cues simultaneously by fusing multi-level features of the backbone. Then, an area search module (ASM) is embedded into each level of the backbone to adaptively search coarse regions of objects with the assistance of region cues from the ABD. After the ASM, an area refinement module (ARM) is utilized to identify fine regions of objects by fusing adjacent-level features with the guidance of boundary cues. Through the deep supervision strategy, DCNet can finally localize the camouflaged objects precisely. Extensive experiments on three benchmark COD datasets demonstrate that our DCNet is superior to 12 state-of-the-art COD methods. In addition, DCNet shows promising results on two COD-related tasks, i.e., industrial defect detection and polyp segmentation.**

***Index Terms*—Camouflaged object detection, dual-constraint, coarse-to-fine, industrial defect detection, polyp segmentation.**

I. Introduction

RECENTLY, camouflaged object detection (COD) [1]–[4] has become a popular research topic due to its potential applications in various fields, such as industrial defect detection [5], [6], polyp segmentation [7]–[9], etc. Unlike traditional generic object detection [10], [11] and salient object detection [12]–[14], where the objects are easily distinguished from the background, COD aims at detecting objects that have high intrinsic similarities to their surroundings [1]. As an emerging yet challenging topic, COD has attracted increasing attentions from academia and industry in the past few years [15].

In early research, due to the lack of large-scale datasets, researchers mainly utilized handcrafted features to detect camouflaged objects via analysis of texture and color information [16]. However, the representation ability of handcrafted features is limited, and the performance of such methods is usually unsatisfactory. In 2020, Fan *et al*. [15] constructed a large-scale COD dataset and proposed a deep learning (DL) based COD method, termed SINet, which has a coarse-to-fine structure that first employs a search module to roughly localize the candidate regions and then utilizes an identification module to accurately localize the camouflaged objects. The coarse-to-fine strategy and global-local representation strategy have been widely adopted and achieved remarkable success in multiple applications, such as medical image segmentation [17], video instance segmentation [18], [19], and video captioning [20]. Integrating the design concept of SINet and these two strategies is conducive to designing effective COD networks. Recently, DL-based COD has achieved a booming development. One widely adopted idea is to upgrade the coarse-to-fine structure of SINet from different perspectives. Representative works were reported from enhancing the initial region cues during the stage of coarse region detection via multi-level information integration [1], refining features during the stage of accurate object detection via cross-scale feature fusion [2], and magnifying candidate regions via object area amplification [21].

Broadly speaking, the emerging COD task parallels traditional object detection [22] and instance segmentation [18]. A direct way for designing specific COD networks is drawing inspiration from popular networks of these two related tasks. In the literature, the feature pyramid network (FPN) and its variants [19], [23]–[26] have been validated effectively in these tasks due to their powerful ability in multi-scale feature representation and fusion, which helps to understand objects with different sizes. Inspired by this, many FPN-based COD methods have been proposed [27]. However, existing methods usually have poor performance when coping with challenging camouflaged objects. Compared to object detection and instance segmentation, COD is more difficult due to ambiguous

This work was supported in part by Guangdong Basic and Applied Basic Research Foundation (No. 2021A1515011348), in part by National Natural Science Foundation of China (Nos. 62371305, 62001302, 62103286), in part by Tencent "Rhinoceros Birds" - Scientific Research Foundation for Young Teachers of Shenzhen University, in part by Natural Science Foundation of China (No. 62271277), in part by Natural Science Foundation of Zhejiang (No. LR22F020002), and in part by Natural Science Foundation of Ningbo (No. 2022J081). (Corresponding Author: Qiuping Jiang)

G. Yue, H. Xiao, H. Xie, and T. Wang are with the School of Biomedical Engineering, Shenzhen University Medical School, Shenzhen University, Shenzhen 518060, China (email: yueguanghui@szu.edu.cn; 2100241022@email.szu.edu.cn; shehare@szu.edu.cn; tfwang@szu.edu.cn).

T. Zhou is with the College of Management, Shenzhen University, Shenzhen 518060, China (e-mail: tianwei@szu.edu.cn).

W. Zhou is with the School of Computer Science and Informatics, Cardiff University. (e-mail: zhouw26@cardiff.ac.uk).

W. Yan is with the School of Computer and Control Engineering, Yantai University, Yantai, 261400, China, and is also with the School of Computer Science Engineering, Nanyang Technological University, Singapore. (e-mail: wqyan@tju.edu.cn).

B. Zhao is with the School of Artificial Intelligence, Sun Yat-sen University, China (e-mail: zhaobaoquan@mail.sysu.edu.cn).

Q. Jiang is with the Faculty of Information Science and Engineering, Ningbo University, Ningbo 315211, China (e-mail: jiangqiuping@nbu.edu.cn).

boundaries caused by the high similarities between the objects and background. To address such a problem, one popular idea is to take the boundary prediction as an auxiliary task with the assumption that boundary constraint helps the network learn more discriminative feature representations. Remarkable works were proposed to generate the boundary map first, and then use the boundary map as the guidance of the encoder [28], [29] and the decoder [30], [31], or use both the coarse region map and boundary map as the guidance of the follow-up decoder [32] for better COD performance. In addition, to help the network focus more on boundary details, a recent attempt also predicted the object regions and boundary cues progressively at multiple stages of the decoder [33].

Although the aforementioned methods have made significant progress in the COD field, there is still much room for performance improvement. There are two possible reasons for this. On the one hand, most existing methods mainly adopt the coarse-to-fine structure from top to down, while ignore the discriminative feature exploration at each level of the network. As a result, the network has limited ability to complete object detection at each level. On the other hand, most boundary-constraint methods usually treat the predictions of object regions and boundary details as two separate tasks, while ignore their special roles in object search and identification. Therefore, the extracted features have limited representation ability, and most methods expose their weakness when dealing with challenging cases.

In this paper, inspired by the biological studies [34] that object detection consists of two steps, i.e., search and identification, we propose a Dual-constraint Coarse-to-fine Network (DCNet) for COD, which simultaneously uses boundary and region information as constraints. Specifically, DCNet takes Pyramid Vision Transformer (PVTv2) [35] as the backbone to extract contextual information at multiple levels effectively. Given that both low-level and high-level features are important for object detection, an area-boundary decoder (ABD) is introduced to mine the initial region cues and boundary cues of objects by aggregating multi-level features of the backbone. Then, an area search module (ASM) is used to adaptively search the coarse regions of objects at each level with the assistance of region cues from the ABD. After the ASM, an area refinement module (ARM) is utilized to identify the fine regions of objects by fusing adjacent-level features with the guidance of boundary cues from the ABD. Finally, DCNet can localize the camouflaged objects precisely in a coarse-to-fine manner by mimicking the search-to-identify mechanism at each level and aggregating multi-level features from top to down through the deep supervision strategy.

Compared to recent works, the proposed DCNet has the following differences. 1) Unlike most FPN-based methods that design specific modules to aggregate complementary information from adjacent features [25] or to calibrate the up-sampled features to be spatially aligned [26] during feature fusion, we propose an effective module that helps the network focus more on boundary information during feature fusion. 2) Contrary to existing coarse-to-fine work [18] that builds specific distributions to first locate instance pixels coarsely and then promote the instance boundary, we utilize the initial region and boundary cues to guide the object detection procedure at each level of the network in a coarse-to-fine manner, inspired by the search-to-identify mechanism. 3) Different from the work [19] that dynamically divides a target instance into subregions, we consider the target object as a whole. 4) Unlike the methods [20], [36] that use a global-local encoder to produce rich semantic vocabulary, we utilize the Transformer-based encoder to model long-range relations of the image for effective object detection. 5) Different from the methods [28], [37] that generate the region and boundary cues by two independent modules, we only use one module (i.e., the ABD) to output these two types of cues.

Our contributions can be summarized as follows:

- We propose a novel dual-constraint coarse-to-fine framework for COD, named DCNet. Different from existing coarse-to-fine COD frameworks that only predict object regions from top to down, DCNet also localizes object regions progressively at each level with the assistance of region and boundary cues.
- To imitate the search-to-identify mechanism, we propose a new feature exploration strategy, in which the ASM and ARM are applied to adaptively search the coarse regions with the assistance of initial region cues and identify the fine regions by integrating coarse regions with boundary cues, respectively. To obtain the initial region cues and edge cues, we propose an ABD that aggregates multi-level features from the backbone.
- Experimental results on three benchmark datasets demonstrate that the proposed DCNet obtains superior performance over 12 state-of-the-art COD methods. In addition, it also performs well on two COD-related tasks, i.e., industrial defect detection and polyp segmentation.

The rest of this paper is organized as follows. In Section II, we briefly review existing COD methods. In Section III, we detail the proposed method. Experimental settings, results, and analysis are presented in Section IV. Section V concludes this paper.

II. Related Works

A. Coarse-to-fine Camouflaged Object Detection

Humans have a two-step mechanism to find the camouflaged object [34], i.e., search and identification. In the past few years, increasing works have been reported for COD by mimicking such a mechanism. One common feature of these works is the use of a coarse-to-fine structure, in which they first search for coarse regions and then identify precise regions of the camouflaged object. For instance, Wang *et al*. [1] proposed a two-stage COD network, where a rough prediction is first obtained by fusing high-level information, based on which the accurate prediction is subsequently generated by using the self-attention and cross-refine unit. Later, they continued the coarse-to-fine structure and made further improvements in the field of COD.

Mei *et al*. [38] employed the rough prediction feature generated by a positioning module, which was designed to locate the potential target objects, to help the feature refinement of target region. Jia *et al*. [39] proposed a network to identify the rough

position of the target object and iteratively magnify and crop the image accordingly. However, their multi-stage strategy suffers from the drawback of significantly prolonging the model's inference time. He *et al*. [40] introduced a ranking COD network to locate, segment, and rank camouflaged objects concurrently. Liu *et al*. [2] introduced a coarse map guided COD network. Different from [1], they utilized a cross-scale feature fusion module to integrate multi-scale information during the stage of accurate prediction generation. Bi *et al*. [41] introduced an in-layer information enhancement module and a cross-layer information aggregation module to simulate the search and identify of human visual observation mechanism, respectively. Considering that magnifying the candidate regions helps to recognize targets more clearly, Xing *et al*. [21] utilized an object area amplification module to amplify operations on feature maps for better performance.

In summary, existing works have shown that the coarse-to-fine structure contributes to accurate COD by considering the human visual observation characteristic. Most existing methods mainly simulate the search-to-identify mechanism from top to down of the network in a progressive manner, while ignore the discriminative feature exploration at each layer of the network. Therefore, most methods do not perform well on challenging cases of COD, and there is still much room for performance improvement.

B. Boundary-guided Camouflaged Object Detection

Both texture and boundary cues are important for humans to find the camouflaged object in an image [42]. The texture cues reflect the object's internal detail information and help us quickly and roughly discover the object. The boundary cues help us distinguish the object from the background. These two kind of cues are complementary and their combination can improve the detection accuracy for challenging samples, especially those with ambiguous boundaries.

Recently, several boundary-guided COD methods have been proposed. For example, Zhai *et al*. [37] transformed feature maps into sample-dependent semantic graphs and incorporated edge guide features to improve camouflage detection accuracy and robustness. This approach captures visual dependencies for enhanced camouflage detection. Chen *et al*. [32] proposed a boundary-guided fusion module to explore the relationship between the camouflaged regions and their boundaries. With this module, the network was able to simultaneously refine the boundary and region features. Sun *et al*. [28] excavated object-related edge semantics by integrating high-level global location information and low-level local edge information under explicit boundary supervision. Specifically, they incorporated the edge semantics with the extracted features at various levels to guide the representation learning of COD as well as enforce the network to focus on the object structure and details. Likewise, Zhou *et al*. [43] extracted object-related edge information from two low-level features. Different from [28], they directly adapted edge feature concatenate with feature representation from fusion module to acquire edge-related feature.

Zhu *et al*. [44] designed a texture label with multiple cues to facilitate accurate COD. Besides, an interactive guidance framework was proposed to capture the indefinable boundaries and the texture differences via progressive interactive guidance. Tu *et al*. [14] proposed to extract hierarchical information to integrate non-local features. They also incorporated boundary prior information into the extracted hierarchical features to detect the objects with more precise boundaries. Lee *et al*. [30] aggregated local boundary information and global information at various levels to generate the boundary cues, which were used to refine the feature in the decoder for better COD performance.

In summary, boundary-guided methods have been favoured to tackle the COD problem. However, most methods either take the object region prediction and boundary prediction as separate tasks yet without any interaction, or merely use the boundary information to guide the feature encoding stage or decoding stage without considering the positive role of object regions in feature extraction. More efforts are needed for performance improvement. In this paper, inspired by the search-to-identify mechanism, we propose a novel COD framework that aggregates multi-level features from top to down and integrates features progressively at each level of the network with the region and boundary constraints.

III. Methodology

A. Motivation and Architecture Overview

Accurately localizing the camouflaged object regions is very challenging due to the high intrinsic similarities between target objects and background. In this study, we propose a novel deep neural network, named DCNet, for COD, which simultaneously uses region and boundary information as constraints. The motivations behind DCNet are two-fold. First, inspired by the biological studies [34] that object detection consists of two steps, i.e., search and identification, we propose to integrate these two steps into the network design concept. Specifically, an ASM is proposed to search the coarse regions of objects with the assistance of the initial region cues, and an ARM is introduced to identify the fine regions of objects with the help of the boundary cues. These two modules can localize the object regions at each level of the network in a coarse-to-fine manner, imitating the search-to-identify mechanism. To obtain both the initial region cues and boundary cues, we propose an ABD, which is constrained by two supervision signals during network training, i.e., the region ground truth and the boundary ground truth. Second, considering the different prediction accuracies of ARMs at different levels, we also progressively refine the prediction in a coarse-to-fine manner through top-to-down connections of ARMs. Each ARM is constrained by a region ground truth. Benefiting from two types of coarse-to-fine predictions, our DCNet can produce more accurate prediction in the COD task.

Fig. 1 presents the architecture of our proposed DCNet, which consists of four key components, i.e., the backbone (PVTv2), the ABD, the ASM, and the ARM. First, PVTv2 is leveraged to encode the input image to acquire multi-level features ($F_i, i \in \{1, 2, 3, 4\}$), which contain rich spatial details and semantic information from the shallow to high levels, respectively. Then, we utilize an ABD to simultaneously

extract the initial region cues P_o and boundary cues P_b by aggregating F_1, F_2, F_3, and F_4. Next, the feature F_i and region cues P_o are fed into an ASM to adaptively search the coarse regions $\mathcal{F}_i$ of the camouflaged object at the i-th level of the network. After that, an ARM is utilized to identify the fine regions of the camouflaged object by fusing adjacent-level features $\widetilde{F}_{i+1}$ and $\mathcal{F}_i$ with the guidance of boundary cues P_b. Finally, the output $\widetilde{F}_i$ of the ARM is processed by a 1×1 convolution operation to generate the prediction mask P_i at each level. Through the deep supervision strategy, we can aggregate multi-level features from top to down and localize the camouflaged object in a coarse-to-fine manner. For better performance, the prediction mask P_1 of the first ARM is selected as the final output.

B. Area-Boundary Decoder

Inspired by the fact that both texture and boundary cues are important for object detection in an image [42], we propose an ABD to generate the initial region cues P_o and boundary cues P_b simultaneously. Fig. 2 shows the architecture of the proposed ABD, which takes F_1, F_2, F_3, and F_4 as the input and outputs P_o and P_b. To generate the region cues P_o, we adopt a pyramid structure that integrates F_1, F_2, F_3, and F_4 in a step-wise manner. Specifically, we first adjust the channel size of F_4 to the same size as that of F_3 by a 1×1 convolution operation, resulting in F_4'. Then, we up-sample F_4' in the spatial domain and add the up-sampled feature map to F_3. Next, the addition F_{43}' is processed by a 1×1 convolution operation to adjust its channel size to that of F_2, followed by an operation of up-sampling 2 times in the spatial domain to match its spatial size to that of F_2 for the sum operation. Through the above operations, we can finally get F_{21}', which is further processed by a 1×1 convolution operation, followed by a Sigmoid function, to output P_o.

As shown in Fig. 2, to generate the boundary cues P_b, we up-sample F_2, F_3, and F_4 to the spatial size of F_1 via bilinear interpolation and concatenate the resulting feature maps with F_1 along the channel direction. After that, the concatenated feature map is processed by a sequence of a 3×3 convolution operation, a batch normalization operation, and a ReLU function (abbreviated as CBR for convenient expression) twice, followed by a 1×1 convolution and a Sigmoid function, to output P_b. The above operations can be formulated as

$$P_b = \sigma(Conv_1(\mathbb{C}^2(F_1 \copyright F_2^{\uparrow\times2} \copyright F_3^{\uparrow\times4} \copyright F_4^{\uparrow\times8}))), \quad (1)$$

where $\sigma(\cdot)$ is the Sigmoid function, $Conv_1$ denotes the 1×1 convolution, $\mathbb{C}^2$ means conducting the CBR operation twice. $\copyright$ is the concatenation operation, and $F_i^{\uparrow\times n}$ stands for the operation of up-sampling feature map n times.

C. Area Search Module

Inspired by the search-to-identify mechanism [15], we propose an ASM to search the coarse region of the camouflaged object. Fig. 3 presents the architecture of the proposed ASM, which has a dual-branch structure. In the upper branch, F_i is first multiplied by the initial region cues P_o in an element-wise manner. Such an operation helps the network focus on the candidate region of the camouflaged object. Since P_o has different spatial size as compared to F_i, we adjust its size to that of F_i using the down-sampling operation ($\downarrow$) before the multiplication operation. Then, the obtained feature map F_i^u is processed by a CBR operation, resulting in $\mathcal{F}_i^u$. After that, inspired by [45], we process $\mathcal{F}_i^u$ by a spatial attention (SA) block to further adaptively explore representative regions. Concretely, as shown by the orange dashed rectangular box in the upper branch of Fig. 3, the SA block consists of a CBR operation, a channel-wise operation (including the channel-wise mean algorithm and the channel-wise maximum algorithm in parallel), a Sigmoid function, and a 1×1 convolution operation, based on which the network can get the important degree w_s of pixels in the spatial domain. To help the network explore representative regions, the input feature map $\mathcal{F}_i^u$ of the SA block is multiplied by w_s:

$$F_i^s = \mathcal{F}_i^u \otimes \underbrace{\sigma(Conv_1(Ce(\mathbb{C}(\mathcal{F}_i^u)) \copyright Ca(\mathbb{C}(\mathcal{F}_i^u))))}_{w_s}, \quad (2)$$

where F_i^s is the output of the SA block, and $\mathcal{F}_i^u = F_i \otimes P_o^{\downarrow}$, where $P_o^{\downarrow}$ is the down-sampling operation to make P_o the same spatial size as F_i. $\otimes$ denotes the element-wise multiplication. $Ce(\cdot)$ and $Ca(\cdot)$ mean the channel-wise mean algorithm and channel-wise maximum algorithm, respectively.

The lower branch of ASM is with the similar architecture as the upper branch and only has two differences. One is the input. Instead of using the initial region cues P_o, the lower branch utilizes $(1-P_o)$ to process the feature map F_i. One reason for this is that the reverse operation helps the network distinguish the foreground and background in another view as compared to the traditional view in the upper branch. By processing F_i from two views simultaneously, the network can understand the candidate region better and finds the boundary cues more easily [46]. Another difference is that we utilize a channel attention (CA) block, instead of a SA block, in the lower branch. As shown in the green dashed rectangular box in the lower right corner of Fig. 3, the CA block has the similar structure as the SA block. The only difference is replacing the channel-wise operation by the spatial operation (including the global average pooling *GAP* and the global maximum pooling *GMP*). The output F_i^l of the lower branch can be obtained by

$$F_i^c = \mathcal{F}_i^l \otimes Conv_1(\underbrace{\sigma(GAP(\mathbb{C}(\mathcal{F}_i^l)) + GMP(\mathbb{C}(\mathcal{F}_i^l)))}_{w_c}), \quad (3)$$

where $\mathcal{F}_i^l = F_i \otimes (1 - P_o^{\downarrow})$, w_c means the weight obtained by the CA block in the channel domain.

After processing F_i with two branches in different views, the coarse region $\mathcal{F}_i$ of the camouflaged object can be obtained by

$$\mathcal{F}_i = Conv_1(F_i^s \oplus F_i^c), \quad (4)$$

where $\oplus$ denotes the sum operation.

D. Area Refinement Module

Benefiting from the ASM, the network can coarsely localize the object regions. To further refine boundary regions, we set

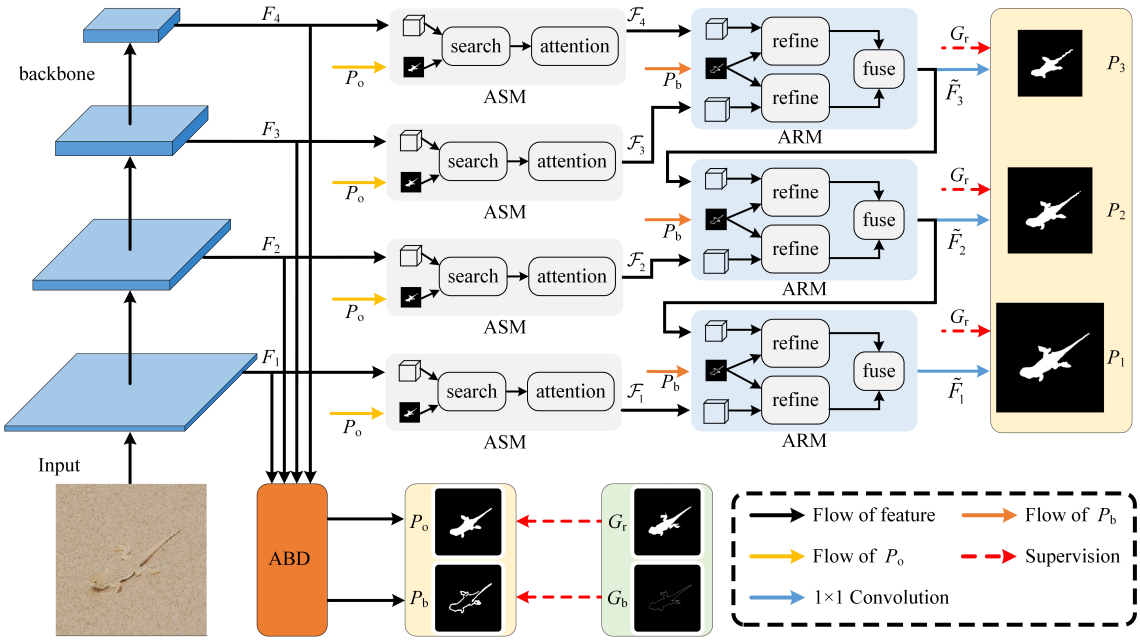


Fig. 1. Overall architecture of our proposed DCNet. An input image is first processed by the backbone. Then, the features $(F_i, i \in \{1, 2, 3, 4\})$ extracted from the backbone are fed into an ABD to generate the initial region cues P_o and boundary cues P_b simultaneously. With the assistance of P_o and P_b, we can localize the object region P_i at each level of the network by processing F_i with the ASM and ARM sequentially. Finally, the prediction mask P_1 of the first ARM is selected as the final output. G_r and G_b are the ground truth of object region and boundary, respectively.

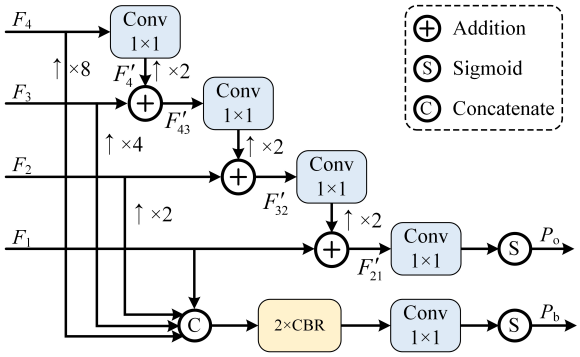


Fig. 2. Illustration of the proposed area-boundary decoder.

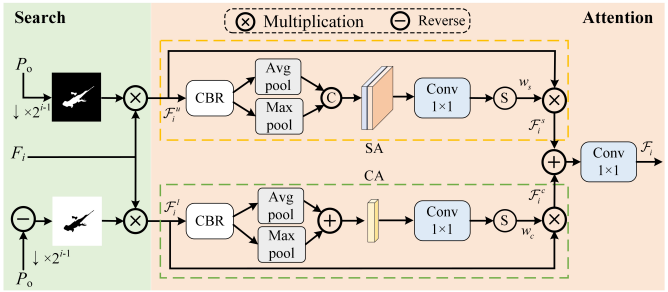


Fig. 3. Illustration of the proposed area search module.

an ARM after the ASM at each level of the network. As show in Fig. 4, the ARM has a dual-branch architecture with three inputs, i.e., $\mathcal{F}_i$, $\widetilde{F}_{i+1}$, and P_b. Specifically, in the lower branch, P_b is first processed by a down-sampling operation to adjust its spatial size to that of $\mathcal{F}_i$. Then, $\mathcal{F}_i$ is multiplied by P_b, and the product is added by $\mathcal{F}_i$. Such an operation helps the network refine the boundary regions. After that, the resulting feature maps is processed by a 3×3 convolution, a global average pooling, and a 1×1 convolution sequentially. Next, a Sigmoid function is used to obtain the channel weight w_l. The above operations can be expressed as

$$w_l = \sigma(Conv_1(GAP(Conv_3(\mathcal{F}_i \otimes P_b^{\downarrow} + \mathcal{F}_i)))), \quad (5)$$

where $P_b^{\downarrow}$ is the down-sampling operation to make P_b the same spatial size as $\mathcal{F}_i$. The upper branch has the similar operations as the lower branch but takes $\widetilde{F}_{i+1}$ as the input and outputs the channel weight w_u. Finally, we concatenate $\mathcal{F}_i$ and $\widetilde{F}_{i+1}$, w_u and w_l, along the channel direction, respectively. These two concatenated maps are multiplied to output $\widetilde{F}_i$. The above operations can be expressed as

$$\widetilde{F}_i = (\widetilde{F}_{i+1}^{\uparrow\times 2} \copyright \mathcal{F}_i) \otimes (w_u \copyright w_l). \quad (6)$$

The region prediction mask P_i can be obtained by processing $\widetilde{F}_i$ with a 1×1 convolution operation. Since there is no high-level information for the ARM at the third level, we use $\mathcal{F}_4$ as the input of ARM's upper branch.

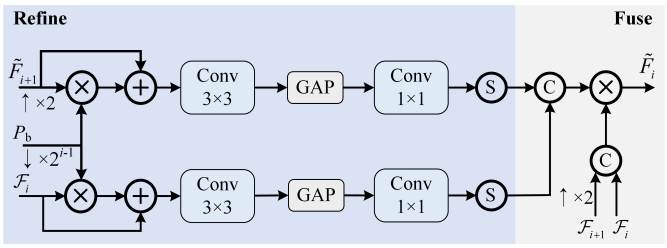


Fig. 4. Illustration of the proposed area refinement module.

E. Feature Visualization

Feature visualization, as a popular manner to improve the interpretability of the network, has been widely adopted in the COD task [47]–[49]. Meanwhile, appropriate discussions about the outcomes of the network will further promote understanding [50]. To intuitively comprehend the working mechanism of our proposed DCNet, we visualize some feature maps at critical positions of the network. Two common scenes are selected for illustration. As shown in the first row of Fig. 5, the side-outputs F_i are heterogeneous, and it is hard to distinguish the object from background in F_i. After processed by the ASM, the object regions are identified and the noise from the background is suppressed to some extent, as shown in the second row. This is because the initial cues P_o from the ABD help the ASM focus more on the candidate regions. Since the quality of P_o is limited, the ASM can only coarsely localize the object regions. This prompts us to further refine the object regions of $\mathcal{F}_i$ via the ARM. As illustrated by the last row, benefiting from the ARM, the complete object regions are highlighted, especially in the high-resolution output F_1. A possible reason for this is that the ARM utilizes the boundary cues P_b from the ABD to highlight the boundary information, providing strong clue to distinguish the object and background. Through the collaboration of ASM and ARM, our DCNet can localize the object regions at each level of the network in a coarse-to-fine manner, corresponding to the search-to-identify mechanism. This is the reason why our DCNet yields accurate predictions, with complete regions and clear boundaries, as shown later in Section IV-B2.

F. Loss Function

The proposed DCNet consists of two kinds of supervision, i.e., the region constraint L_r and the boundary constraint L_b. L_r highlights harder pixels by assigning more weight. It is composed of a weighted binary cross-entropy loss L^w_{BCE} and a weighted IoU loss L^w_{IoU} [51]:

$$L_r\left(P,G\right) = L^w_{BCE}(P,G) + L^w_{IoU}(P,G), \tag{7}$$

where P and G are the prediction mask and the ground truth, respectively. L_b is used to help the network recognize the boundary cues of the camouflaged object. In this study, we use the popular Dice loss [52] as L_b. To improve performance, we adopt a deep supervision strategy. Specifically, four region prediction masks, including one P_o from the ABD and three P_i from the ARM at three levels, are supervised by the region ground truth G_r. In addition, one boundary prediction mask P_b from the ABD is supervised by the boundary ground truth G_b. Overall, the loss function L_t of the proposed DCNet is described as follow:

$$L_t = \alpha L_r(P_o, G_r) + \lambda \sum_{i=1}^{3} L_r(P_i, G_r) + \gamma L_b(P_b, G_b), \tag{8}$$

where α, λ, and γ are weighting parameters. In this study, we set them to 2, 1, and 3, respectively, according to the ablation experiments in Section IV-C3.

IV. Experimental Results and Analysis

A. Protocols

1) Datasets: Three benchmark COD datasets, including CHAMELEON [53], COD10K [15], and CAMO [54], are selected to evaluate and compare the proposed DCNet with competing COD methods. The CHAMELEON has 76 images. The CAMO contains 1,250 images, including 1,000 images in the training set and 250 images in the testing set. The COD10K has 10,000 images, where 3,040 and 2,026 images are used for training and testing, respectively. Following previous works [2], [15], [28], [32], [55], [56], we use the training sets of CAMO and COD10K for training, and use the remaining images in three datasets for testing in our experiments.

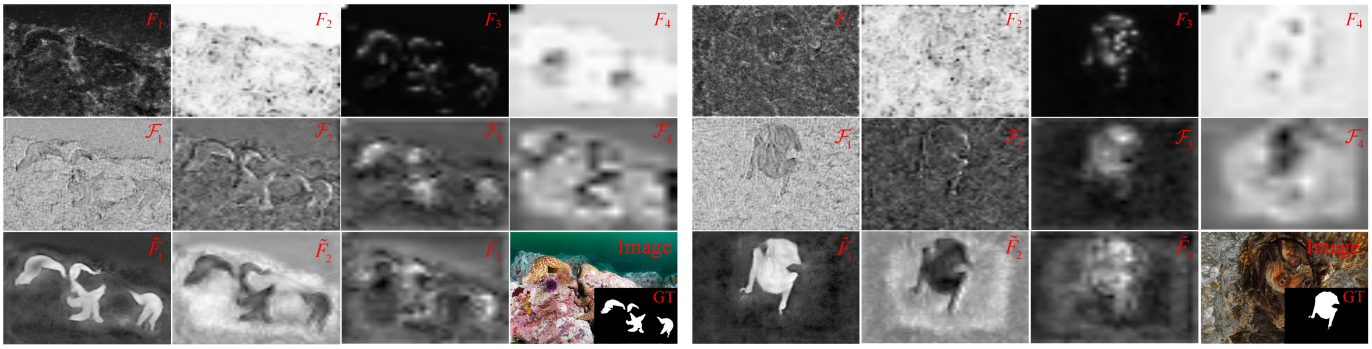
2) Evaluation Metrics: To comprehensively compare our DCNet with competing COD methods, we choose four widely-used evaluation metrics for performance comparisons, including Structure-measure (S_α) [57], E-measure (E_φ) [58], weighted F-measure (F^ω_β) [59], and mean absolute error (*MAE*) [60]. Generally, a superior COD method has a larger value of S_α, E_φ, and F^ω_β, while a smaller value of MAE.

3) Implementation Details: We implement our DCNet in PyTorch and train the model on a workstation equipped with an Nvidia GeForce RTX3090 GPU and two Intel Xeon Silver 4210R CPUs @2.40 GHz. The backbone (i.e., PVTv2) of DCNet is initialized by the parameters pre-trained on ImageNet, while other layers are randomly initialized. During the training stage, all the input images are resized into 480×480 and augmented by randomly horizontal flipping. The Adam optimization algorithm is utilized to optimize the network. The learning rate is first initialized to 1e-4 and then adjusted by a poly strategy with the power setting of 0.9. The training process is stopped at the 30-th epoch, and the batch size is set to 16.

B. Comparisons with State-of-the-Arts

We compare our proposed DCNet with 12 state-of-the-art COD methods, including SINet [15], TINet [44], PraNet [61], PFNet [38], R-MGL [37], Joint-COD [62], D2C-Net [1], BgNet [32], CANet [56], BGNet [28], SegMaR [39], and MSCAF-Net [2]. For fair comparisons, all the experimental results of these methods are either provided by the original works or obtained by retraining models from the officially released source codes.

1) Quantitative Analysis: Table I shows the quantitative comparisons between our DCNet and competing methods. As can be seen, the proposed DCNet outperforms the state-of-the-art COD methods 9 times and delivers the second best performance 3 times in a total of 12 comparisons. Additionally, compared with the recently reported method (i.e., MSCAF-Net), our DCNet has a leading edge of 0.4% and 2.1% in S_α and F^ω_β on average. This may attribute to the usage of the novel dual-constraint coarse-to-fine framework, which helps the network mine region and boundary cues of the camouflaged object sequentially at each level and localize accurate object regions progressively along different levels from top to down. We also investigate and compare the proposed DCNet with competing methods in terms of the floating point



(a) (b)

Fig. 5. Features maps obtained at different positions in the proposed DCNet. The first to third rows include the side-outputs F_i of the backbone, the outputs $\mathcal{F}_i$ of the ASM, and the outputs $\widetilde{F}_i$ of the ARM, respectively. The color image and its ground truth (GT) are presented at the lower right side of (a) and (b). For display, each feature is averaged along the channel dimension.

TABLE I
QUANTITATIVE EVALUATION RESULTS OF OUR METHOD WITH OTHER SOTAS COMPARISON ON BENCHMARK DATASETS. $\uparrow \setminus \downarrow$ REPRESENT THE LARGER OR SMALLER IS BETTER. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD.

Method	Year	CHAMELEON				COD10K-Test				CAMO-Test				FLOPs	#Params
		$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$		
SINet [15]	2020	0.869	0.891	0.740	0.044	0.771	0.806	0.551	0.051	0.751	0.771	0.606	0.100	19.55G	48.95M
TINet [44]	2021	0.874	0.916	0.783	0.038	0.793	0.861	0.635	0.042	0.781	0.848	0.678	0.087	-	-
PraNet [61]	2021	0.882	0.931	0.810	0.033	0.800	0.877	0.660	0.040	0.782	0.842	0.695	0.085	**13.15G**	30.50M
Joint-COD [62]	2021	0.894	0.943	0.848	0.030	0.817	0.892	0.684	0.035	0.803	0.853	0.728	0.076	89.89G	121.63M
PFNet [38]	2021	0.882	0.942	0.810	0.033	0.800	0.868	0.660	0.040	0.782	0.852	0.695	0.085	26.60G	46.498M
R-MGL [37]	2021	0.893	0.923	0.813	0.030	0.814	0.865	0.666	0.035	0.775	0.847	0.673	0.088	431.87G	78.47M
D2C-Net [1]	2022	0.889	0.939	0.848	0.030	0.807	0.876	0.720	0.037	0.774	0.818	0.735	0.087	-	-
BgNet [32]	2022	0.894	0.943	0.823	0.029	0.804	0.881	0.663	0.039	0.804	0.859	0.719	0.075	27.74G	60.47M
CANet [56]	2022	0.901	0.940	0.843	0.028	0.832	0.890	0.745	0.033	0.807	0.866	0.767	0.075	72.65G	57.13M
BGNet [28]	2022	0.901	0.943	0.851	0.027	0.831	0.901	0.722	0.033	0.812	0.870	0.749	0.073	58.50G	77.80M
SegMaR [39]	2022	0.906	0.954	0.860	0.025	0.831	0.901	0.722	0.033	0.815	0.872	0.742	0.071	33.65G	55.62M
MSCAF-Net [2]	2023	0.912	**0.958**	0.865	0.022	0.865	0.927	0.775	0.024	**0.873**	**0.929**	0.828	**0.046**	30.04G	**29.70M**
DCNet (Ours)	-	**0.920**	**0.958**	**0.890**	**0.019**	**0.873**	**0.934**	**0.810**	**0.022**	0.870	0.922	**0.831**	0.050	94.74G	54.43M

operations (FLOPs) and the number of parameters (#Params). As shown in the last two columns of Table I, DCNet has the FLOPs of 94.74G and #Params of 54.43M, ranking tenth and fifth among eleven competing methods, respectively. In other words, compared to competitors, our DCNet has no obvious competitive advantage in these two aspects. Despite this, DCNet has better detection accuracy, which is evident from quantitative comparisons shown in the left part of Table I and qualitative comparisons shown later in Fig. 6. It is worth noting that, the focus of our current study is not designing lightweight COD networks, but proposing an effective COD network with high detection accuracy. Generally, a COD method with higher accuracy is highly desired in practice as it contributes to more precise object detection. In the future, we will optimize the structure of DCNet to reduce its computational cost and parameter number while keeping high accuracy.

2) Qualitative Analysis: Fig. 6 provides some qualitative comparisons of our DCNet and state-of-the-art COD methods. Here, we only present the results of methods that provide the prediction masks and that release the source codes for reproduction. These images are selected from the COD10K dataset and include typically challenging cases, such as multi-objects (rows 1 and 6), small object (rows 2 and 9), occlusion (row 3), out-of-view (rows 4, 5, 8), and indefinable boundary (rows 5, 6, 8). From Fig. 6, we can observe that the proposed DCNet achieves more accurate prediction results than competing methods. Specifically, it is robust across different scenes. Although in extremely difficult samples (as shown in the second row) where almost all methods cannot localize the camouflaged object accurately, our DCNet still delivers better prediction masks compared with competing methods. Furthermore, due to the cooperative usage of region and boundary constraints, our DCNet can identify the object with the forecast boundary closer to ground truth than competing methods, as shown in the sixth and eighth rows.

C. Ablation Study

In this subsection, we further conduct ablation experiments to investigate the effectiveness of the proposed ASM and ARM. During the ablation study of each module, the other parts of our framework remain unchanged.

1) Effectiveness of ASM: To test the effectiveness of ASM, we remove it and directly multiply the side-output F_i of the backbone by the initial region cues P_o. The product is processed by a 1×1 convolution operation to generate the coarse regions $\mathcal{F}_i$. By comparing the results in the first and last rows of Table II, we can observe that the removal of ASM can cause obvious performance drop on three datasets. This

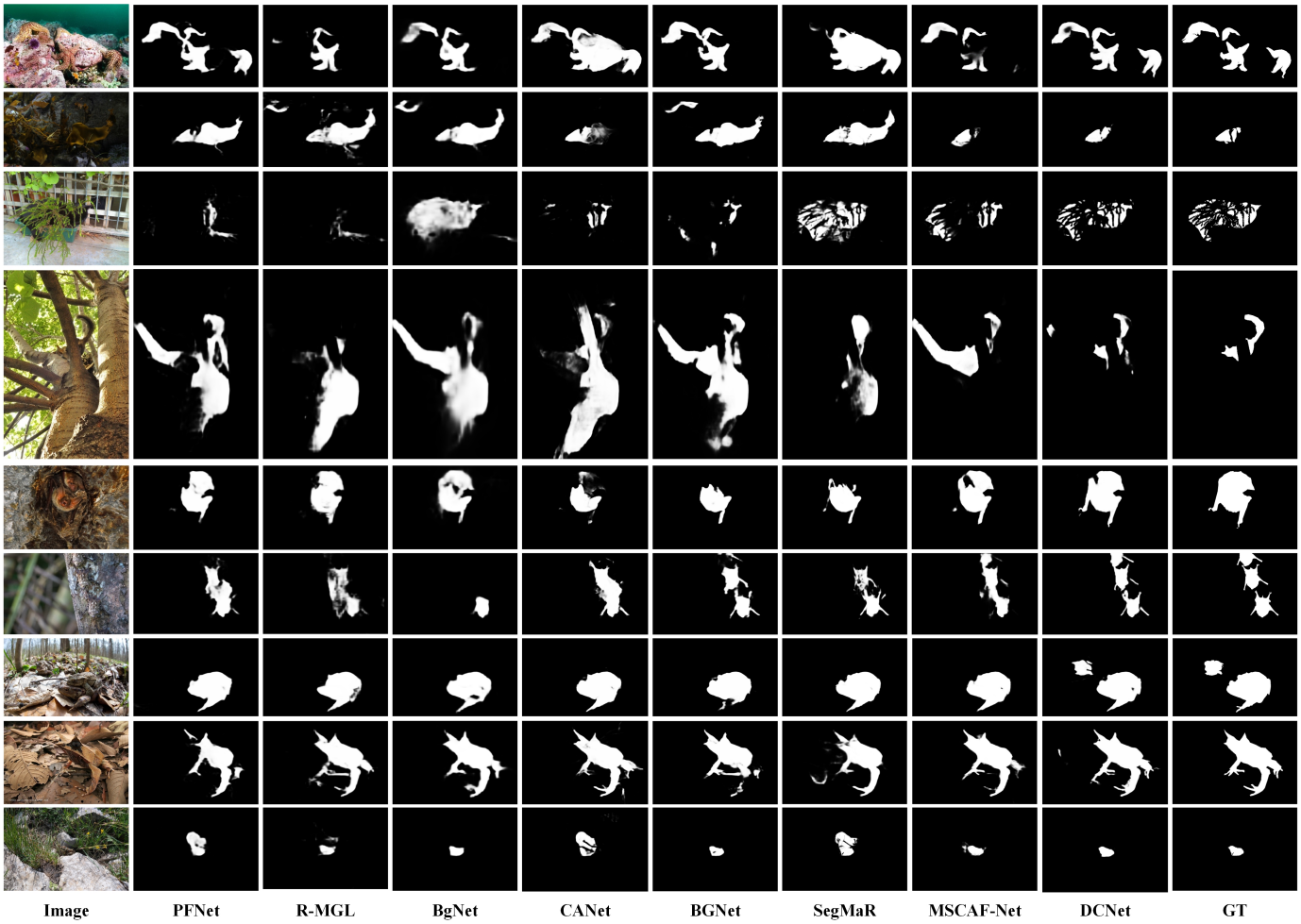


Fig. 6. Visual comparisons of our proposed DCNet with state-of-art COD methods. The ground truth (GT) of each image is presented at the last column.

TABLE II
QUANTITATIVE EVALUATION OF ABLATION STUDIES ON THREE DATASETS. WE HIGHLIGHTED THE BEST RESULTS IN BOLD.

Method	CHAMELEON				COD10K-Test				CAMO-Test			
	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
baseline	0.895	0.927	0.858	0.025	0.849	0.904	0.779	0.025	0.852	0.907	0.808	0.054
DCNet + w/o ASM	0.915	0.944	0.874	0.023	0.870	0.927	0.803	0.023	0.869	0.916	0.825	**0.050**
DCNet + w/o ARM	0.917	0.948	0.879	0.021	0.870	0.929	0.806	**0.022**	0.866	0.914	0.828	**0.050**
DCNet (Ours)	**0.920**	**0.958**	**0.890**	**0.019**	**0.873**	**0.934**	**0.810**	**0.022**	**0.870**	**0.922**	**0.831**	**0.050**

indicates that ASM plays a positive role in the COD task. In the ASM, we multiply the side-output F_i by the initial region cues P_o to help the network focus on the candidate region of the camouflaged object, and also multiply the side-output F_i by the reverse initial region cues $(1-P_o)$ to help the network distinguish the foreground and background in another view. Here, we further investigate the effectiveness of such operations. For this purpose, we remove P_o and only take F_i as the input of ASM. As shown in the first row of Table III, the performance of our DCNet is slightly decreased if we remove P_o. Additionally, we utilize two attention blocks, i.e., SA and CA, in the ASM to help the network enhance feature representation. Here, we also explore their effectiveness by removing each of them separately from ASM. By comparing the results in Table III, we can observe that the removal of SA or CA can cause performance drop on three datasets. This indicates that both SA and CA play a positive role in the COD task.

2) Effectiveness of ARM: To test the effectiveness of ARM, we remove it and directly concatenate $\mathcal{F}_i$ and the adjacent-level feature $\widetilde{F}_{i+1}$. As shown in the last two rows of Table II, there is performance drop when removing ARM from the proposed DCNet. For instance, there is an E_φ drop of 1.0%, 0.5%, and 0.8% on three datasets, respectively. This indicates that the proposed ARM contributes to improving the detection performance. In the ARM, P_b provides the boundary cues to help the network refine boundary information. Here, we further investigate the efficacy of P_b. For this purpose, we directly remove it from the ARM and keep the remaining part unchanged. As shown in the penultimate row of Table III,

the removal of P_b has resulted in slight performance drop. This indicates that P_b plays a positive role in the ARM and contributes to improving detection performance.

3) Effect of the Weights in Eq. (8)*:* In Eq. (8), there are three parameters, i.e., α, λ, and γ, to balance the contributions of each loss function. Here, we adopt the grid search to investigate the sensitivity of parameter setting. Specifically, we fix λ=1 and configure both α and γ with 5 conditions ($\alpha \in [1, 3]$ and $\gamma \in [2, 4]$ in the step of 0.5). For each combination of α, λ, and γ, a COD model is learned. Fig. 7 shows the results. Clearly, the results change with the parameter combination. In this study, we set α=2 and γ=3 due to the superior performance achieved.

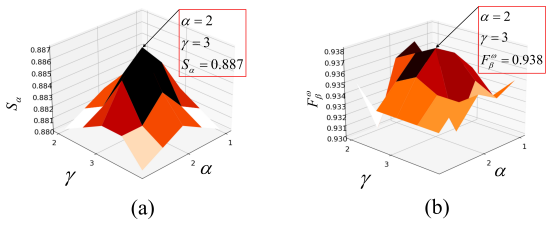


Fig. 7. Effect of the weights in Eq. (8) on the performance of the proposed DCNet: (a) the mean S_α value and (b) the mean F_β^ω value on three datasets.

D. Failure Cases

Generally, the camouflaged object is similar in pattern to its surroundings. To better distinguish the object from background, we utilize an ARM that helps the proposed DCNet focus more on boundary information, achieving considerable performance, as shown in Table I and Fig. 6. Despite this, our DCNet also produces biased predictions in some very challenging cases. As illustrated in Fig. 8, DCNet may fail to accurately localize the proper object when the object has no clear boundaries with its surroundings (see the 1st row) and when the object shares a high pattern similarity with its surroundings (see the 2nd and 3rd rows). Similarly, the boundary-constraint competitors, e.g., R-MGL, BgNet, and BGNet, also yield erroneous predictions in these cases. This indicates that, for more accurate predictions in such challenging cases, more advanced techniques apart from the boundary constraint are required. We leave the design of such techniques as a future work.

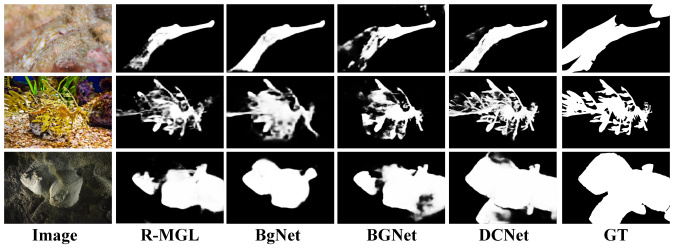


Fig. 8. Illustration of failure cases.

E. COD-related Applications

In this study, we further validate the effectiveness of our proposed DCNet on two COD-related tasks.

1) Application to Polyp Segmentation: Following the previous work [61], we select five public datasets, including Kvasir [63], CVC-ClinicDB [64], ETIS [65], CVC-300 [66], and CVC-ColonDB [67], to test the effectiveness of our method on polyp segmentation. As suggested by [61], we take 900 images from Kvasir and 550 images from CVC-ClinicDB as the training set and take the remaining images in these five datasets as the testing set. Apart from F_β^ω, S_α, E_φ, and MAE, two widely used evaluation metrics ($mDice$ and $mIoU$) in the polyp segmentation field are also selected to evaluate the performance.

Quantitative Results. Due to space limitation, we compare our DCNet with three methods, which have been validated effectively in polyp segmentation, including PraNet [61], PFNet [38], and BGNet [28]. The quantitative results are listed in Table IV. As can be seen, our proposed DCNet achieves an $mDice$ score of 0.864, an $mIou$ score of 0.803, an F_β^ω score of 0.846, an S_α score of 0.906, an E_φ score of 0.933, and an *MAE* score of 0.091 on average for five datasets. It is superior to these competing methods in terms of six evaluation metrics on Kvasir, CVC-ClinicDB, ETIS, and CVC-ColonDB, while performs the second best in terms of $mIoU$, F_β^ω, and S_α on CVC-300. Overall, our DCNet achieves elegant performance and is competent for the polyp segmentation task.

Qualitative Results. Fig. 9 illustrates the qualitative results of PraNet, PFNet, BGNet, and our DCNet on four representative scenes, including large polyps (rows 1 and 2), low light environment (row 3), multiple tiny polyps (row 3), and single tiny polyp (row 4). Moreover, these polyps vary in texture, shape and size, which brings a tough challenge for the segmentation task. As shown by Fig. 8, our proposed DCNet can identify polyps well and perform better than competing methods in these challenging cases.

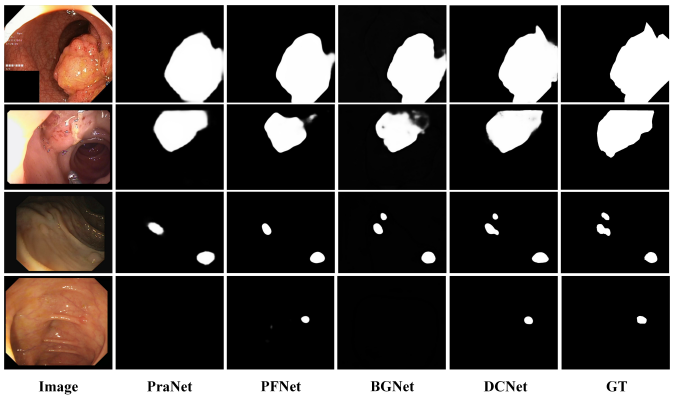


Fig. 9. Visual comparison of the proposed DCNet with three methods.

2) Application to Industrial Defect Detection: In this study, we choose the popular industrial defect detection dataset MVTEC AD [68] to further validate the superiority of our proposed DCNet in COD-related tasks. Since MVTEC AD is mainly used for testing unsupervised methods and its training set has no ground truth, we only choose its testing set (1,258

TABLE III
FURTHER ABLATION STUDIES OF THE ATTENTION BLOCKS IN THE ASM AND THE BOUNDARY GUIDANCE P_b IN THE ARM. WE HIGHLIGHTED THE BEST RESULTS IN BOLD.

Method	CHAMELEON				COD10K-Test				CAMO-Test			
	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$F_\beta^\omega \uparrow$	$MAE \downarrow$
DCNet + w/o P_o	0.919	0.955	0.886	0.021	0.869	0.930	0.802	0.023	**0.871**	0.918	**0.832**	**0.050**
DCNet + w/o P_o + w/o SA	0.916	0.944	0.881	0.021	0.868	0.929	0.802	0.023	0.869	**0.922**	0.830	**0.050**
DCNet + w/o P_o + w/o CA	0.916	0.947	0.880	0.022	0.869	0.928	0.801	0.023	0.869	0.921	0.830	**0.050**
DCNet + w/o P_b	0.918	0.955	0.889	0.020	0.872	0.934	**0.810**	**0.022**	0.864	0.914	0.827	0.052
DCNet (Ours)	**0.920**	**0.958**	**0.890**	**0.019**	**0.873**	**0.934**	**0.810**	**0.022**	0.870	**0.922**	0.831	**0.050**

TABLE IV
QUANTITATIVE RESULTS ON FIVE POLYP SEGMENTATION DATASETS.

Method	Kvasir					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	0.899	0.847	0.886	0.913	0.944	0.028
PFNet [38]	0.902	0.848	0.891	0.911	0.928	0.028
BGNet [28]	0.842	0.850	0.788	0.904	0.929	0.034
DCNet (Ours)	**0.917**	**0.874**	**0.909**	**0.926**	**0.946**	**0.023**
Method	CVC-ClinicDB					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	0.906	0.861	0.894	0.933	0.963	0.009
PFNet [38]	0.921	0.872	0.914	0.931	0.971	0.013
BGNet [28]	0.840	0.866	0.772	0.920	0.967	0.019
DCNet (Ours)	**0.929**	**0.884**	**0.919**	**0.944**	**0.974**	**0.007**
Method	ETIS					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	0.660	0.601	0.644	0.809	0.834	0.014
PFNet [38]	0.707	0.620	0.663	0.819	0.846	0.019
BGNet [28]	0.526	0.554	0.460	0.759	0.817	0.023
DCNet (Ours)	**0.782**	**0.713**	**0.751**	**0.867**	**0.885**	**0.019**
Method	CVC-300					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	**0.897**	**0.839**	**0.878**	**0.940**	0.960	**0.006**
PFNet [38]	0.892	0.823	0.867	0.929	0.957	0.008
BGNet [28]	0.795	0.833	0.711	0.914	**0.973**	0.013
DCNet (Ours)	**0.897**	0.832	0.876	0.935	0.960	0.009
Method	CVC-ColonDB					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	0.708	0.634	0.692	0.818	0.858	0.039
PFNet [38]	0.730	0.656	0.716	0.826	0.886	0.038
BGNet [28]	0.599	0.604	0.543	0.784	0.818	0.048
DCNet (Ours)	**0.794**	**0.712**	**0.774**	**0.856**	**0.899**	**0.033**

TABLE V
QUANTITATIVE RESULTS ON MVTEC AD.

Method	MVTEC AD					
	$mDice \uparrow$	$mIoU \uparrow$	$F_\beta^\omega \uparrow$	$S_\alpha \uparrow$	$E_\varphi \uparrow$	$MAE \downarrow$
PraNet [61]	0.496	0.435	0.467	0.692	0.851	0.024
PFNet [38]	0.467	0.428	0.450	0.669	0.899	0.026
BGNet [28]	0.267	0.452	0.199	0.605	0.887	0.067
DCNet (Ours)	**0.600**	**0.559**	**0.582**	**0.737**	**0.948**	**0.021**

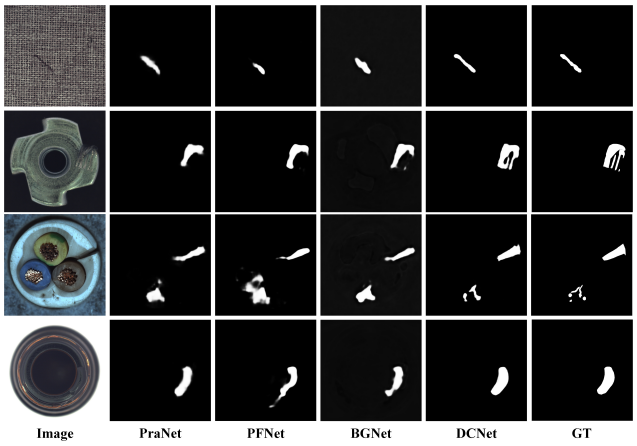


Fig. 10. Visual comparison of the proposed DCNet with three methods on representative images of MVTEC AD. Zoom in for more details.

images) in our experiments. Specifically, we randomly divide 1,258 images into the training set and testing set by a ratio of 4:1. Similar to the polyp segmentation task, six evaluation metrics are used here.

Quantitative Results. As shown in Table V, our proposed DCNet surpasses three competing methods by a large margin in terms of six evaluation metrics. To be specific, compared with the second best method PraNet, our DCNet has a performance gain of 10.4% in $mDice$, 12.4% in $mIoU$, 11.5% in F_β^ω, 4.5% in S_α, 9.7% in E_φ, and 0.3% in *MAE*.

Qualitative Results. Fig. 10 presents the visual results of three competing methods as well as our proposed DCNet on four challenging cases, including the allochroic carpet, bent metal nut, combined cable, and contaminated bottle. As can be seen, DCNet can accurately distinguish the target defect from the background and performs better than PraNet, PFNet, and BGNet. This indicates that our proposed DCNet has greater potential for the industrial defect detection task than these three competing methods.

V. CONCLUSION

In this paper, we propose a novel framework, named DCNet, for COD by considering both region and boundary constraints. Specifically, the proposed DCNet is inspired by the search-to-identify mechanism and applies a coarse-to-fine manner, with three key modules. First, an ABD module is introduced to explicitly predict the initial region cues and boundary cues by integrating both low-level and high-level features from the backbone. Then, at each level of the network, we embed an ASM and an ARM. The ASM is used to search the coarse region maps with the guidance of the initial region cues from the ABD, and the ARM is utilized to identify the fine regions with the guidance of the boundary cues from the ABD. Through the deep supervision strategy, we can fuse multi-level features from top to down and finally accurately localize the regions of the camouflaged object. Extensive experiments on three benchmark COD datasets indicate that our proposed DCNet surpasses 12 state-of-the-art COD methods in terms of four evaluation metrics. In addition, our proposed DCNet is also competent for polyp segmentation and defect detection tasks with good performance.

REFERENCES

[1] K. Wang, H. Bi, Y. Zhang, C. Zhang, Z. Liu, and S. Zheng, "D^{2}c-net: A dual-branch, dual-guidance and cross-refine network for camouflaged object detection," *IEEE Transactions on Industrial Electronics*, vol. 69, no. 5, pp. 5364–5374, 2022.

[2] Y. Liu, H. Li, J. Cheng, and X. Chen, "Mscaf-net: A general framework for camouflaged object detection via learning multi-scale context-aware features," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[3] G. Chen, S.-J. Liu, Y.-J. Sun, G.-P. Ji, Y.-F. Wu, and T. Zhou, "Camouflaged object detection via context-aware cross-level fusion," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6981–6993, 2022.

[4] J. Ren, X. Hu, L. Zhu, X. Xu, Y. Xu, W. Wang, Z. Deng, and P.-A. Heng, "Deep texture-aware features for camouflaged object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 33, no. 3, pp. 1157–1167, 2023.

[5] X. Zhou, Y. Wang, Q. Zhu, J. Mao, C. Xiao, X. Lu, and H. Zhang, "A surface defect detection framework for glass bottle bottom using visual attention model and wavelet transform," *IEEE Transactions on Industrial Informatics*, vol. 16, no. 4, pp. 2189–2201, 2020.

[6] Q. Luo, X. Fang, L. Liu, C. Yang, and Y. Sun, "Automated visual defect detection for flat steel surface: A survey," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 3, pp. 626–644, 2020.

[7] G. Yue, S. Li, R. Cong, T. Zhou, B. Lei, and T. Wang, "Attention-guided pyramid context network for polyp segmentation in colonoscopy images," *IEEE Transactions on Instrumentation and Measurement*, vol. 72, pp. 1–13, 2023.

[8] G. Yue, W. Han, B. Jiang, T. Zhou, R. Cong, and T. Wang, "Boundary constraint network with cross layer feature integration for polyp segmentation," *IEEE Journal of Biomedical and Health Informatics*, vol. 26, no. 8, pp. 4090–4099, 2022.

[9] G. Yue, S. Li, T. Zhou, M. Wang, J. Du, Q. Jiang, W. Gao, T. Wang, and J. Lv, "Adaptive context exploration network for polyp segmentation in colonoscopy images," *IEEE Transactions on Emerging Topics in Computational Intelligence*, vol. 7, no. 2, pp. 487–499, 2023.

[10] L. Liu, W. Ouyang, X. Wang, P. Fieguth, J. Chen, X. Liu, and M. Pietikäinen, "Deep learning for generic object detection: A survey," *International Journal of Computer Vision*, vol. 128, pp. 261–318, 2020.

[11] W. Liu, G. Lin, T. Zhang, and Z. Liu, "Guided co-segmentation network for fast video object segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 4, pp. 1607–1617, 2021.

[12] L. Zhang, Q. Zhang, and R. Zhao, "Progressive dual-attention residual network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5902–5915, 2022.

[13] C. Zhang, S. Gao, D. Mao, and Y. Zhou, "Dhnet: Salient object detection with dynamic scale-aware learning and hard-sample refinement," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7772–7782, 2022.

[14] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 582–593, 2020.

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

[16] A. Mondal, "Camouflaged object detection and tracking: A survey," *International Journal of Image and Graphics*, vol. 20, no. 04, p. 2050028, 2020.

[17] J. Wan, S. Yue, J. Ma, and X. Ma, "A coarse-to-fine full attention guided capsule network for medical image segmentation," *Biomedical Signal Processing and Control*, vol. 76, p. 103682, 2022.

[18] Z. Qin, X. Lu, X. Nie, D. Liu, Y. Yin, and W. Wang, "Coarse-to-fine video instance segmentation with factorized conditional appearance flows," *IEEE/CAA Journal of Automatica Sinica*, vol. 10, no. 5, pp. 1192–1208, 2023.

[19] D. Liu, Y. Cui, W. Tan, and Y. Chen, "Sg-net: Spatial granularity network for one-stage video instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 9816–9825.

[20] L. Yan, Q. Wang, Y. Cui, F. Feng, X. Quan, X. Zhang, and D. Liu, "Gl-rg: Global-local representation granularity for video captioning," *arXiv preprint arXiv:2205.10706*, 2022.

[21] H. Xing, Y. Wang, X. Wei, H. Tang, S. Gao, and W. Zhang, "Go closer to see better: Camouflaged object detection via object area amplification and figure-ground conversion," *IEEE Transactions on Circuits and Systems for Video Technology*, accepted, in press, DOI: 10.1109/TCSVT.2023.3255304, 2023.

[22] Y. Cui, L. Yang, and H. Yu, "Dq-det: Learning dynamic query combinations for transformer-based object detection and segmentation," *arXiv preprint arXiv:2307.12239*, 2023.

[23] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 2117–2125.

[24] Y. Cui, L. Yan, Z. Cao, and D. Liu, "Tf-blender: Temporal feature blender for video object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 8138–8147.

[25] M. Hu, Y. Li, L. Fang, and S. Wang, "A2-fpn: Attention aggregation based feature pyramid network for instance segmentation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 15 343–15 352.

[26] D. Liu, J. Liang, T. Geng, A. Loui, and T. Zhou, "Tripartite feature enhanced pyramid network for dense prediction," *IEEE Transactions on Image Processing*, vol. 32, pp. 2678–2692, 2023.

[27] H. Bi, C. Zhang, K. Wang, J. Tong, and F. Zheng, "Rethinking camouflaged object detection: Models and datasets," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5708–5724, 2022.

[28] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," *arXiv preprint arXiv:2207.00794*, 2022.

[29] H. Zhu, P. Li, H. Xie, X. Yan, D. Liang, D. Chen, M. Wei, and J. Qin, "I can find you! boundary-guided separated attention network for camouflaged object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 36, no. 3, 2022, pp. 3608–3616.

[30] M. Lee, S. Cho, C. Park, D. Lee, J. Lee, and S. Lee, "Global-local aggregation with deformable point sampling for camouflaged object detection," *arXiv preprint arXiv:2211.12048*, 2022.

[31] Z. Qiu, Z. Wang, M. Zhang, Z. Xu, J. Fan, and L. Xu, "Bdg-net: boundary distribution guided network for accurate polyp segmentation," in *Medical Imaging 2022: Image Processing*, vol. 12032. SPIE, 2022, pp. 792–799.

[32] T. Chen, J. Xiao, X. Hu, G. Zhang, and S. Wang, "Boundary-guided network for camouflaged object detection," *Knowledge-Based Systems*, vol. 248, p. 108901, 2022.

[33] X. Xu, M. Zhu, J. Yu, S. Chen, X. Hu, and Y. Yang, "Boundary guidance network for camouflage object detection," *Image and Vision Computing*, vol. 114, p. 104283, 2021.

[34] J. R. Hall, I. C. Cuthill, R. Baddeley, A. J. Shohet, and N. E. Scott-Samuel, "Camouflage, detection and identification of moving targets," *Proceedings of the Royal Society B: Biological Sciences*, vol. 280, no. 1758, p. 20130064, 2013.

[35] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pvt v2: Improved baselines with pyramid vision transformer," *Computational Visual Media*, vol. 8, no. 3, pp. 415–424, 2022.

[36] L. Yan, S. Ma, Q. Wang, Y. Chen, X. Zhang, A. Savakis, and D. Liu, "Video captioning using global-local representation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 10, pp. 6642–6656, 2022.

[37] Q. Zhai, X. Li, F. Yang, C. Chen, H. Cheng, and D.-P. Fan, "Mutual graph learning for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 12 997–13 007.

[38] H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, "Camouflaged object segmentation with distraction mining," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 8772–8781.

[39] Q. Jia, S. Yao, Y. Liu, X. Fan, R. Liu, and Z. Luo, "Segment, magnify and reiterate: Detecting camouflaged objects the hard way," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4713–4722.

[40] H. He, X. Li, G. Cheng, J. Shi, Y. Tong, G. Meng, V. Prinet, and L. Weng, "Enhanced boundary learning for glass-like object segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 15 859–15 868.

[41] H. Bi, C. Zhang, K. Wang, and R. Wu, "Towards accurate camouflaged object detection with in-layer information enhancement and cross-layer information aggregation," *IEEE Transactions on Cognitive and Developmental Systems*, accepted, in press, DOI: 10.1109/TCDS.2022.3172331, 2022.

[42] L. Chen, "Topological structure in visual perception," *Science*, vol. 218, no. 4573, pp. 699–700, 1982.

[43] T. Zhou, Y. Zhou, C. Gong, J. Yang, and Y. Zhang, "Feature aggregation and propagation network for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 31, pp. 7036–7047, 2022.
[44] J. Zhu, X. Zhang, S. Zhang, and J. Liu, "Inferring camouflaged objects by texture-aware interactive guidance network," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 4, 2021, pp. 3599–3607.
[45] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "Eca-net: Efficient channel attention for deep convolutional neural networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 11 534–11 542.
[46] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, "Bilateral attention network for rgb-d salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 1949–1961, 2021.
[47] Y. Lv, J. Zhang, Y. Dai, A. Li, B. Liu, N. Barnes, and D.-P. Fan, "Simultaneously localize, segment and rank the camouflaged objects," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 11 591–11 601.
[48] F. Yang, Q. Zhai, X. Li, R. Huang, A. Luo, H. Cheng, and D.-P. Fan, "Uncertainty-guided transformer reasoning for camouflaged object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 4146–4155.
[49] M. Zhang, S. Xu, Y. Piao, D. Shi, S. Lin, and H. Lu, "Preynet: Preying on camouflaged objects," in *Proceedings of the 30th ACM International Conference on Multimedia*, 2022, pp. 5323–5332.
[50] W. Wang, C. Han, T. Zhou, and D. Liu, "Visual recognition with deep nearest centroids," *arXiv preprint arXiv:2209.07383*, 2022.
[51] J. Wei, S. Wang, and Q. Huang, "F^3net: fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 34, no. 07, 2020, pp. 12 321–12 328.
[52] E. Xie, W. Wang, W. Wang, M. Ding, C. Shen, and P. Luo, "Segmenting transparent objects in the wild," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIII 16*. Springer, 2020, pp. 696–711.
[53] P. Skurowski, H. Abdulameer, J. Błaszczyk, T. Depta, A. Kornacki, and P. Kozieł, "Animal camouflage analysis: Chameleon database," *Unpublished manuscript*, vol. 2, no. 6, p. 7, 2018.
[54] T.-N. Le, T. V. Nguyen, Z. Nie, M.-T. Tran, and A. Sugimoto, "Anabranch network for camouflaged object segmentation," *Computer Vision and Image Understanding*, vol. 184, pp. 45–56, 2019.
[55] Y. Sun, G. Chen, T. Zhou, Y. Zhang, and N. Liu, "Context-aware cross-level fusion network for camouflaged object detection," *arXiv preprint arXiv:2105.12555*, 2021.
[56] J. Liu, J. Zhang, and N. Barnes, "Modeling aleatoric uncertainty for camouflaged object detection," in *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 2022, pp. 1445–1454.
[57] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4548–4557.
[58] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *arXiv preprint arXiv:1805.10421*, 2018.
[59] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 1597–1604.
[60] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 733–740.
[61] 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," in *23rd International Conference on Medical Image Computing and Computer Assisted Intervention*. Springer, 2020, pp. 263–273.
[62] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 10 071–10 081.
[63] D. Jha, P. H. Smedsrud, M. A. Riegler, P. Halvorsen, T. de Lange, D. Johansen, and H. D. Johansen, "Kvasir-seg: A segmented polyp dataset," in *MultiMedia Modeling: 26th International Conference, MMM 2020, Daejeon, South Korea, January 5–8, 2020, Proceedings, Part II 26*. Springer, 2020, pp. 451–462.
[64] J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, D. Gil, C. Rodríguez, and F. Vilariño, "Wm-dova maps for accurate polyp highlighting in colonoscopy: Validation vs. saliency maps from physicians," *Computerized Medical Imaging and Graphics*, vol. 43, pp. 99–111, 2015.
[65] J. Silva, A. Histace, O. Romain, X. Dray, and B. Granado, "Toward embedded detection of polyps in wce images for early diagnosis of colorectal cancer," *International Journal of Computer Assisted Radiology and Surgery*, vol. 9, pp. 283–293, 2014.
[66] D. Vázquez, J. Bernal, F. J. Sánchez, G. Fernández-Esparrach, A. M. López, A. Romero, M. Drozdzal, and A. Courville, "A benchmark for endoluminal scene segmentation of colonoscopy images," *Journal of Healthcare Engineering*, vol. 2017, 2017.
[67] N. Tajbakhsh, S. R. Gurudu, and J. Liang, "Automated polyp detection in colonoscopy videos using shape and context information," *IEEE Transactions on Medical Imaging*, vol. 35, no. 2, pp. 630–644, 2015.
[68] P. Bergmann, M. Fauser, D. Sattlegger, and C. Steger, "Mvtec ad–a comprehensive real-world dataset for unsupervised anomaly detection," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 9592–9600.

Guanghui Yue received the B.S. degree in communication engineering from Tianjin University in 2014, and the Ph.D. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2019. He was a joint Ph.D. student with the School of Computer Science and Engineering, Nanyang Technological University, Singapore, from September 2017 to January 2019.

He is currently an Associated Professor with the School of Biomedical Engineering, Health Science Center, Shenzhen University. His research interests include bioelectrical signal processing, multimedia quality assessment, 3D image visual discomfort prediction, pattern recognition, and medical image analysis.

HouLu Xiao received the B.S. degree in biomedical engineering from Jiujiang University, Jiangxi Province, China, in 2021.

He is currently working toward the master's degree in biomedical engineering with Shenzhen University, China. His research interests include deep learning and medical image analysis.

Hai Xie received the Ph.D. degree in School of Electronics and Information Engineering, Shenzhen University, Shenzhen, China, in 2020. He is currently with School of Biomedical Engineering, Health Science Center, Shenzhen University, China. His search interest is medical image analysis and deep learning.

Tianwei Zhou received the B.S. degree in automation and the Ph.D. degree in control science and engineering from Tianjin University, Tianjin, China, in 2014 and 2019, respectively. She was a joint Ph.D. student with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore, from August 2017 to August 2018.

She is currently an Assistant Professor with the College of Management, Shenzhen University, Shenzhen, China. Her current research interests include event-triggered control, intelligent scheduling, image processing, and medical image analysis.

Wei Zhou is an Assistant Professor at Cardiff University, United Kingdom. Dr Zhou was a Post-doctoral Fellow at University of Waterloo, Canada. Wei received the PhD degree from the University of Science and Technology of China in 2021, joint with the University of Waterloo from 2019 to 2021. Dr Zhou was a visiting scholar at National Institute of Informatics, Japan, a research assistant with Intel, and a research intern at Microsoft and Alibaba. Wei's research interests span multimedia computing, perceptual image processing, and computational vision.

Weiqing Yan received the PhD. degree in information and communication engineering from Tianjin University, Tianjin, China, in 2017. She was a visiting PhD student at visual spatial perceived lab, University of California, Berkeley, CA, USA from September 2015 to September 2016.

She is currently an associate professor with the School of Computer and Control Engineering, Yantai University, Yantai City, China, she is also a research fellow at School of Computer Science and Engineering, Nanyang Technological University, Singapore. Her research interests include Multi-view Representation Learning, 3D Vision.

Baoquan Zhao (Member, IEEE) is currently an associate professor at the School of Artificial Intelligent, Sun Yat-sen University, China. Prior to his current position, he was a Research Fellow at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. He received his PhD in computer science and technology from Sun Yat-sen University in 2017. His research interests include visual information analysis, multimedia systems and applications, and point cloud processing and compression, in which he has been published over 20 conference and journal papers, and filed 6 patents. He has been actively serving as the technical program committee member and reviewer for a dozen international conferences and journals.

Tianfu Wang received the Ph.D. degree in biomedical engineering from Sichuan University, Chengdu,China, in 1997.

He is currently a Professor with Shenzhen University, Shenzhen, China. His current research interests include ultrasound image analysis, medical image processing, pattern recognition, and medical imaging.

Qiuping Jiang (Member, IEEE) received the Ph.D. degree in signal and information processing from Ningbo University, Ningbo, China, in June 2018. He is currently an Professor with Ningbo University. From January 2017 to May 2018, he was a Visiting Student with Nanyang Technological University, Singapore. His research interests include image processing, visual perception modelling, and deep learning with applications in computer vision. He is the Associate Editor for IET Image Processing, Journal of Electronic Imaging, and APSIPA Transactions on Information and Signal Processing.