Boundary-Aware Distracted Attention Network for Camouflaged Object Detection

Yihan Shang†, Lin Wang†, Junyu Dong, *Member, IEEE*, and Xinghui Dong, *Member, IEEE*

***Abstract*—Camouflaged Object Detection (COD) involves precisely segmenting objects, which are seamlessly blended into the surroundings, due to the similarities between the surroundings and them in color, texture and other camouflaging techniques. To address the challenges that such similarities pose, distraction mining approaches have been proposed. However, they normally struggle with boundary delineation and extraction of relevant features. To get rid of this dilemma, we introduce a Boundary-Aware Distracted Attention Network (BADANet), which exploits both the boundary cue and the distraction mining strategy. Specifically, we first use a pre-trained network as the encoder for feature extraction. Then the features extracted are sent to a Boundary Shrinking Module (BSM) that we design. After the output is processed by a Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP) module, a boundary map is produced. Given an encoder block, the features extracted are also fed into a Multi-Branch Bidirectional Fusion Block (MBBFB), which performs the bidirectional fusion at the channel dimension, followed by the multiple fusion conducted at the spatial dimension via an individual MDASPP module. We further propose a Boundary-Aware Distracted Attention (BADA) block, which receives both the features fused and the boundary map. With regard to the encoder blocks, a series of BADA blocks are comprised of a distracted attention decoder. Finally, the detection mask is generated by the last BADA block. We have conducted a series of experiments on four popular COD data sets. Experimental results demonstrate that the proposed BADANet normally outperforms 18 baseline methods. These promising results should be due to the boundary-aware distracted attention mechanism that we design.[1]**

***Impact Statement*—Distraction mining has been applied to the field of Camouflaged Object Detection (COD), which aims to separate the foreground from the background of an image for feature extraction and coding individually. However, the majority of existing distraction mining methods ignored the boundary characteristics of objects and did not use the global attention mechanism. In contrast, the BADANet that we propose overcomes this problem by introducing a Boundary-Aware Distracted Attention (BADA) mechanism. Experimental results show that the proposed method normally performs better than 18 baselines on four publicly available COD data sets. It is suggested that our method is competent for the COD task. In addition, this method can be potentially used for other subtle visual anomaly detection tasks, such as defect detection and lesion detection.**

***Index Terms*—Camouflaged object detection, object detection, distraction mining, boundary-aware, pyramidal vision transformer.**

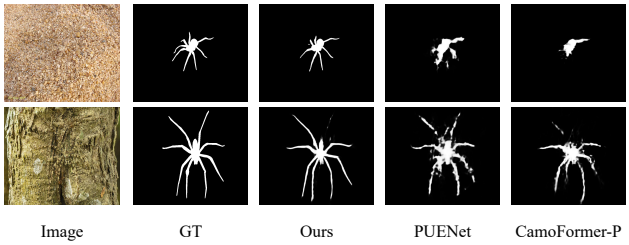

Fig. 1. Visual comparison of the results of three COD approaches in two scenes with blurred boundaries. Here, "-P" means that PVTv2 [19] was used as the backbone. Compared with the recently developed PUENet [20] and CamoFormer [21], which were also designed on top of distraction mining, our approach distinguishes the boundaries more accurately and achieves the finer object segmentation.

I. Introduction

CAMOUFLAGED Object Detection (COD) [1], [2], [3], [4] aims to discern objects which are concealed in the challenging conditions, where the color, texture and patterns of the object closely resemble those of the surroundings. In contrast to conventional object detection [5], [6], [7], [8], [9], COD normally encounters the greater challenge because the process of camouflaging enables objects to be seamlessly integrated into the surroundings which results in the indistinct boundaries and foreground-background confusion phenomenon [10]. Due to the similarity between the foreground and background in a camouflaged image, COD has also been widely utilized in various domains with the analogous data characteristics, such as polyp segmentation [11], lung infection segmentation [12], product defect detection [13], visual object tracking [14] and pest detection [15].

Many recent studies have been delved into the application of deep learning methods [16], [3], [17] to COD. To overcome the challenge of the highly similar camouflaged foreground and the background, an increasing number of studies were performed on top of the distraction mining strategy [2], [18], which was designed to extract the features of the camouflaged foreground and the background separately. Although these studies achieved promising results, accurately detecting objects with the complex scene and intricate shapes remains challenging (see Fig. 1).

This dilemma should be attributed to two significant issues that current distraction mining methods normally encounter. First, the extraction method of distraction features is relatively simple, which only focuses on local features and lacks

†Equal contribution

This study was supported by the National Natural Science Foundation of China (NSFC) (No. 42176196) (Corresponding author: Xinghui Dong).

Y. Shang, L. Wang, J. Dong and X. Dong are with the State Key Laboratory of Physical Oceanography and the Faculty of Information Science and Engineering, Ocean University of China, Qingdao, 266100. (e-mail: shangyihan@stu.ouc.edu.cn, wanglin7089@stu.ouc.edu.cn, dongjunyu@ouc.edu.cn, xinghui.dong@ouc.edu.cn).

[1]Upon acceptance of this paper, we will make our models and code publicly available.

global attention. Second, the boundary cue of an object is often neglected during the distraction process, leading to the ambiguous boundary captured by the model.

To get out of the dilemma, we propose a Boundary-Aware Distracted Attention Network (BADANet) on top of both the boundary cue and the distraction mining strategy. Given a camouflaged object, this network can not only explicitly detect the boundary of this object but also implicitly refine the segmentation mask of it during the distraction mining process. To be specific, a pre-trained network is used as the encoder for feature extraction. The features extracted are fed into a Boundary Shrinking Module (BSM) that we design in order to accurately detect the boundary of an object. We also propose a Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP) module. Since this module constructs a series of spatial pyramids from the features received, the more precise and detailed feature representation can be derived, which is useful for segmenting objects within the complex scene. The output of the BSM is then processed by an MDASPP module and the result is a boundary map.

Regarding an encoder block, the features extracted are also sent to a Multi-branch Bidirectional Fusion Block (MBBFB) for the purpose of obtaining the richer feature representation. This module fulfills the bidirectional fusion at the channel dimension. The multiple fusion is then performed at the spatial dimension through a single MDASPP module. To leverage the attention mechanism to bolster the mining of distraction features and refine the foreground and background features extracted during the distraction process, we further propose a Boundary-Aware Distracted Attention (BADA) block. In terms of the encoder blocks, a series of BADA blocks are used, which are comprised of a distracted attention decoder. Each BADA block receives both the features fused and the boundary map. In particular, the boundary data is used to refine the foreground and background masks of the object during the distraction process. As a result, the boundary of the object in the feature maps is explicitly strengthened. Finally, the detection mask is generated by the last BADA block.

To our knowledge, the COD task has not been performed using such a boundary-aware distracted attention network. The contributions of this study can be summarized as threefold.

1) We introduce a Boundary-Aware Distracted Attention Network (BADANet) by jointly exploiting both the boundary cue and the distraction mining strategy. As a result, the foreground and background can be segmented with the fine boundary.
2) We propose a Boundary Shrinking Module (BSM), which can be used to extract the boundary of the camouflaged object. The boundary can be further injected into the Boundary-Aware Distracted Attention (BADA) block that we design. This block utilizes both the boundary information and the features extracted using the encoder during the distraction mining process. Consequently, the segmentation precision of the foreground from the background can be enhanced even objects are camouflaged within the surroundings.
3) We have conducted extensive experiments on four publicly available COD data sets. The results provide the community with a set of new benchmarks.

The rest of this paper is arranged as follows. We review the related work in Section II. Then the proposed network is introduced in Section III. The experimental setup and results are reported in Section IV. Finally, we draw our conclusion in Section V.

II. Related Work

A. Camouflaged Object Detection

The conventional COD methods [22], [23] normally accomplished the segmentation of camouflaged objects by investigating and devising effective hand-crafted features which were used to distinguish camouflaged objects from the background. These methods achieved favorable results in straightforward scenes, but usually exhibited significant performance deterioration when confronted with intricate scenes.

With the rapid development of deep learning techniques, many COD methods were introduced based on deep learning in recent years. Inspired by the biological process of predators prey on prey, Fan et al. [1] proposed a two-stage search-identification COD method and introduced a large COD data set, namely, COD10K. Lv et al. [24] investigated the level of camouflage through localization and ranking and published a data set, i.e., NC4K. Both the data sets greatly promoted the development of the deep learning-based COD methods.

Sun et al. [25] introduced object boundary cues and used boundary semantics to force the model to highlight the boundary structural features of the camouflaged object. In [26], Qin et al. directed the model to implicitly focus on the boundary information of the camouflaged object through designing a composite loss function. To provide supervision on the texture branch, Ji et al. [17] constructed object-level gradient labels, enabling the model to focus on extracting the richer feature information from gradients.

In [3], Pang et al. used the image at multiple scales as the input to improve detection accuracy. Zhong et al. [27] incorporated the frequency domain feature information in order to extract the clues of camouflaged targets in the frequency domain. Feature alignment was also used to integrate the spatial and frequency domain features. For the purpose of mining subtle cues which can be used to distinguish the foreground from the background, He et al. [10] decomposed the features of the foreground and background into different frequency bands using learnable wavelets.

B. Vision Transformer

Transformer [28] was initially developed for Natural Language Processing (NLP) tasks. Then it attracted the attention from researchers in the field of computer vision. For example, DEtection TRansformer (DETR) [29] and Vision Transformer (ViT) [30] incorporated the self-attention mechanism into object detection and image classification tasks, respectively. Consequently, Transformer was widely applied to various vision tasks, including image classification [31], [32], object detection [33], semantic segmentation [34] and object tracking [35].

Recently, ViT has also been extensively used in the task of COD. Yang et al. [36] utilized uncertainty within the Transformer framework in order to guide the attention of the model towards areas of the higher uncertainty. In [37], Zhang et al. used Transformer to progressively refine high-level semantic features extracted using the Res2Net [38] backbone. Liu et al. [39] designed a dual-task Transformer network which encoded the camouflage foreground and background separately, enabling the two tasks to interact.

Song et al. [40] employed SwinTransformer [31] as the backbone and devised a coarse-to-fine two-stage focus scanning network. Hu et al. [41] adopted PVT [32] as the backbone and used a cyclic iterative refinement method to fuse features. In [20], Zhang et al. built a combined encoder on top of the CNN and Transformer for estimating uncertainty. To progressively aggregate neighboring Transformer features, Huang et al. [16] designed a layer-by-layer shrinking pyramid decoder.

C. Distraction Mining Strategy

The distraction mining strategy [2] aims to separate the foreground and background of an image and separately encode the features to strengthen the ability of the model for the sake of distinguishing between the foreground and background. As a result, this strategy is able to alleviate the heterogeneous interference in the foreground or background more accurately. Due to the high similarity between the camouflaged object and the background, the distraction mining strategy has achieved promising results in the COD task.

Mei et al. [2] distracted the high-level prediction map by multiplying the inverse with the current layer features and subsequently utilized simple addition and subtraction to suppress both the false positive interference in the foreground and the false negative interference in the background. To address the issue of feature confusion which might arise from the simple multiplication during the distraction process, the reverse feature map was inserted into the feature group of the current layer for feature fusion [18]. Zhu et al. [42] used the distraction mining strategy to highlight the camouflaged object boundary. The channel attention was used to fuse distraction features on top of the boundary information. In [21], three multi-head self-attention, together with different masking strategies, were used to model the global, foreground and background characteristics of a camouflaged image, respectively.

III. Boundary-Aware Distracted Attention Network

In this section, we first present the overall architecture of the proposed Boundary-Aware Distracted Attention Network (BADANet). Then we introduce each module of the BADANet and the loss function in detail.

A. Overview of the BADANet

The architecture of the proposed BADANet is illustrated in Fig. 2, which consists of an encoder, a boundary decoder, a feature refinement module and a distracted attention decoder. In particular, the pre-trained PVTv2 network [19] is used as the backbone of the encoder. Given an image, $I \in \mathbb{R}^{3\times H\times W}$, which contains at least a camouflaged object, it is first fed into the backbone network. The result is five sets of features, denoted as $\{f_i^t\}_{i=1}^5$. Then these features are sent to the boundary decoder and the feature refinement module simultaneously. The former aims to acquire the intermediate boundary and the boundary-related feature representation. On the other hand, the latter produces the richer multi-level feature representation. The features produced together with the output of the boundary decoder are further fed into the distracted attention decoder. Finally, camouflaged objects are segmented from the background boosted by the boundary information.

B. Encoder

We use the pre-trained PVTv2 network [19] as the backbone of the encoder. Given an image $I \in \mathbb{R}^{3\times H\times W}$, it is passed through the four blocks of the backbone, in which the resolutions of feature maps are $\frac{H}{4}\times\frac{W}{4}$, $\frac{H}{8}\times\frac{W}{8}$, $\frac{H}{16}\times\frac{W}{16}$ and $\frac{H}{32}\times\frac{W}{32}$, respectively, and four sets of features are extracted at these blocks. The numbers of channels of the four sets of features are 64, 128, 320 and 512, respectively.

To reduce the computational cost of following operations, the four sets of features are sent to four CBR blocks, respectively, each of which comprises a 3×3 convolution, a batch normalization layer and an ReLU activation function. In particular, the feature maps extracted at the first PVTv2 block are also upsampled to the size of $\frac{H}{2}\times\frac{W}{2}$ and are fed into an additional CBR block. As a result, the number of channels of the feature maps produced by the five CBR blocks is reduced to 64. In total, five sets of features are obtained, denoted as $\{f_i^t\}_{i=1}^5$.

C. Boundary Decoder

The boundary decoder receives the five sets of features generated by the encoder. It consists of a Boundary Shrinking Module (BSM), a Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP) block and a CBR block. The output of the boundary decoder contains the boundary map of the camouflaged image and the boundary-related features.

1) Boundary Shrinking Module: In the existing studies [25], [10], [20], [42], it has been demonstrated that the boundary cue is useful for boosting the performance of the COD approaches. We are motivated to design a Boundary Shrinking Module (BSM), as shown in Fig. 3, to obtain the boundary of objects and generate the corresponding boundary features.

Since the boundary detection task usually encounters the class imbalance issue, it is unavoidably influenced by the heterogeneous non-boundary information during the decoding process. Inspired by the SINetV2 [18], we use the element-wise multiplication in the hierarchical connections to alleviate the heterogeneous effect. Considering that high-resolution low-level features encode the shape characteristics of objects, which complements high-level features, the BSM not only fuses high-level features but also fuses low-level features. In contrast, only high-level features were used in the SINetV2. In terms of the five sets of features extracted at the encoder,

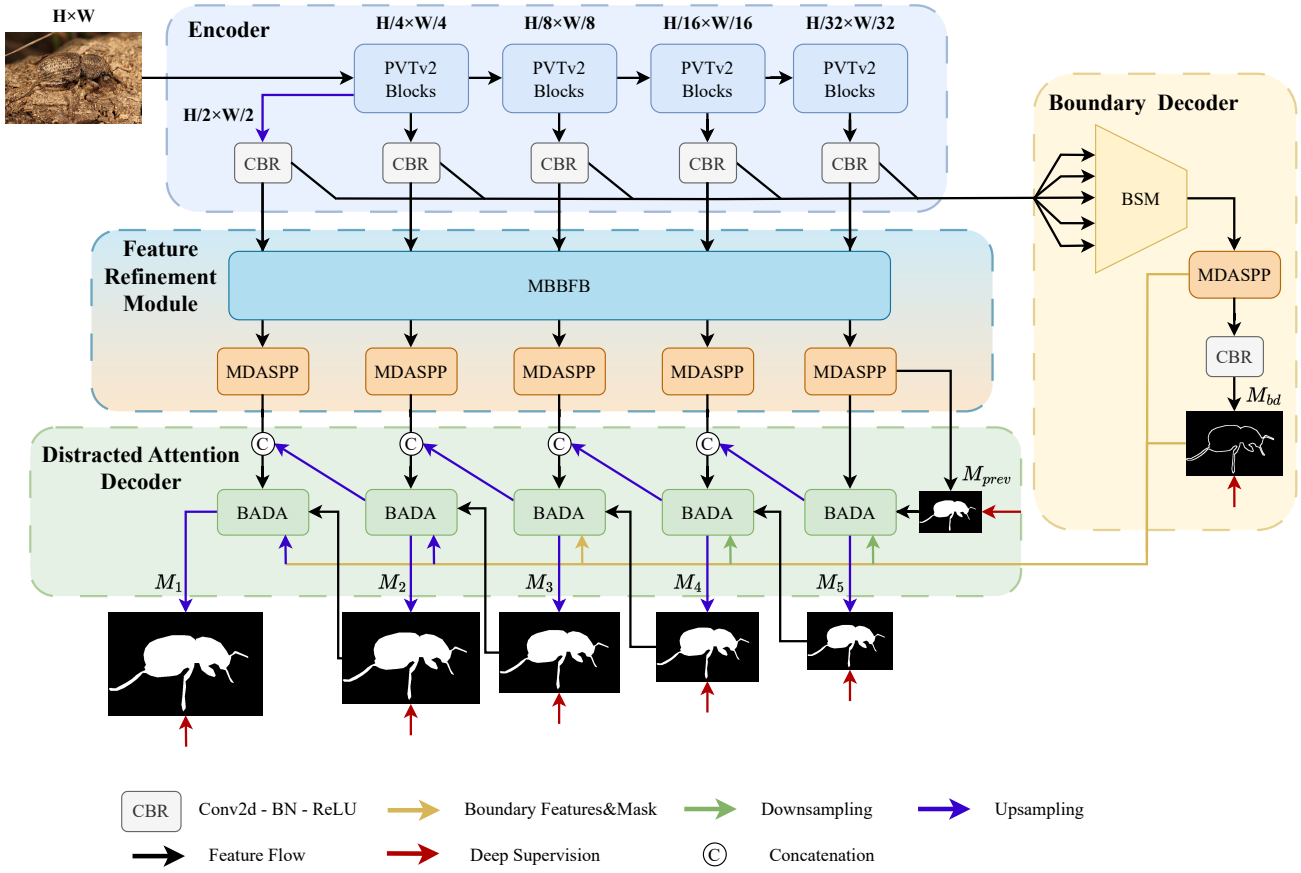

Fig. 2. Overview of the proposed Boundary-Aware Distracted Attention Network (BADANet).

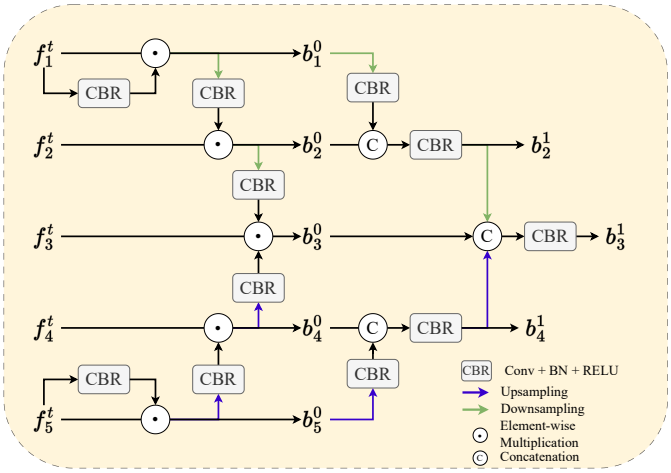

Fig. 3. The detailed structure of the Boundary Shrinking Module (BSM).

five different boundary maps, denoted as $\{b_i^1\}_{i=1}^5$, are produced, respectively. The boundary extraction operation can be expressed as

$$b_1^0 = CBR(f_1^t) \odot f_1^t,\ b_1^0, f_1^t \in \mathbb{R}^{64\times256\times256}, \tag{1}$$

$$b_5^0 = CBR(f_5^t) \odot f_5^t,\ \{b_5^0, f_5^t\} \in \mathbb{R}^{64\times16\times16}, \tag{2}$$

$$b_2^0 = CBR(Down(b_1^0)) \odot f_2^t,\ \{b_2^0, f_2^t\} \in \mathbb{R}^{64\times128\times128}, \tag{3}$$

$$b_4^0 = CBR(Up(b_5^0)) \odot f_4^t,\ \{b_4^0, f_4^t\} \in \mathbb{R}^{64\times32\times32}, \tag{4}$$

$$b_3^0 = CBR(Up(b_4^0)) \odot CBR(Down(b_2^0)) \odot f_3^t,\ \{b_3^0, f_3^t\} \in \mathbb{R}^{64\times64\times64}. \tag{5}$$

where $Down(\cdot)$ and $Up(\cdot)$ stand for the $2\times$ downsampling and upsampling operations, respectively, $\odot$ indicates the element-wise multiplication, and $CBR(\cdot)$ denotes the CBR block which consists of a 3×3 convolution, a batch normalization layer and an ReLU activation function.

Then we aggregate two low-resolution boundary maps, b_1^0 and b_2^0, and two high-resolution boundary maps, b_4^0 and b_5^0, into two medium-resolution boundary maps, b_2^1 and b_4^1, separately. This process can be expressed as

$$b_2^1 = CBR(Cat(CBR(Down(b_1^0)),\ b_2^0)),\ b_2^1 \in \mathbb{R}^{64\times128\times128}, \tag{6}$$

$$b_4^1 = CBR(Cat(CBR(Up(b_5^0)),\ b_4^0)),\ b_4^1 \in \mathbb{R}^{64\times32\times32}, \tag{7}$$

where $Cat(\cdot)$ represents the concatenation operation along the channel dimension. Finally, b_2^1 and b_4^1 are aggregate into a single boundary map, b_3^1, which can be formulated as

$$b_3^1 = CBR(Cat(CBR(Down(b_2^1)),\ b_3^0,\ CBR(Up(b_4^1)))),\ b_3^1 \in \mathbb{R}^{64\times64\times64}. \tag{8}$$

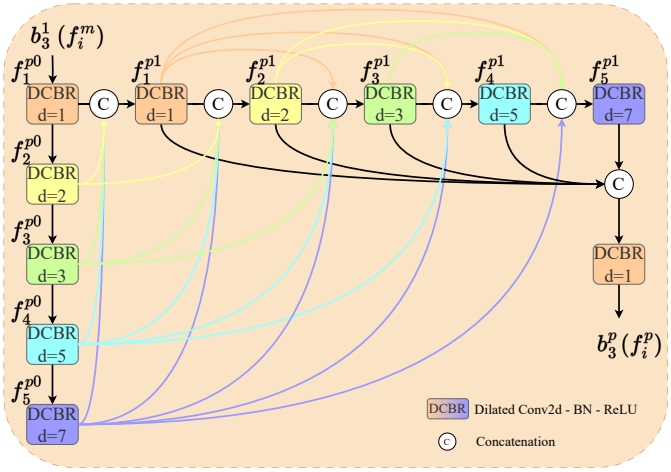

Fig. 4. The detailed structure of the Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP). The dilated convolutions in the vertical direction are used to extract the features at different spatial scales, while those in the horizontal direction are used to densely connect these features and build multiple pyramids.

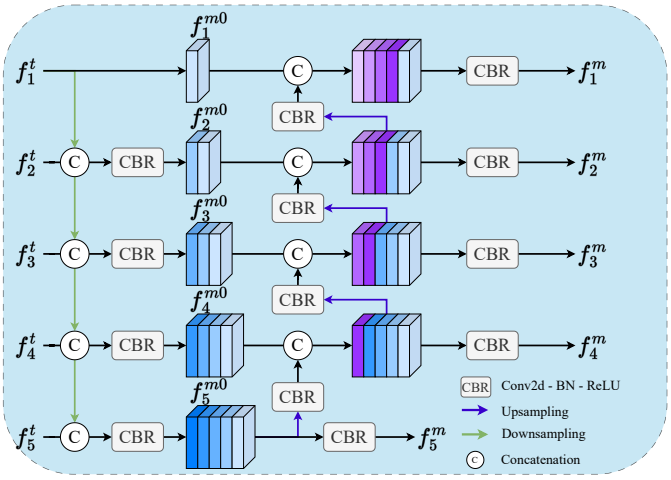

Fig. 5. The detailed structure of the Multi-Branch Bidirectional Fusion Block (MBBFB).

2) Multiple Dense Atrous Spatial Pyramid Pooling: Due to the high similarity between camouflaged objects and the background, detection of the scale change of these objects is challenging. In this situation, it is important to encode the characteristics of different spatial scales for the purpose of capturing the scale change of the camouflaged objects. Although Atrous Spatial Pyramid Pooling (ASPP) [43] employs a series of convolutions with varying dilation rates in order to acquire the image characteristics at disparate spatial scales, it lacks connections among the features extracted. In contrast, Dense Atrous Spatial Pyramid Pooling (DASPP) [44] further refines the receptive field by establishing dense connections in order that the features of different scales extracted using ASPP are linked. However, it put the less emphasis on the features extracted using low-dilation convolutions.

To address the above-mentioned issues, we propose a Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP) block. As depicted in Fig. 4, this block uses an additional set of interconnected dilated convolutions to build multiple pyramids. Specifically, we first take b_3^1 as input and use a set of interconnected dilated convolutions to extract five sets of features at five different spatial scales, which are denoted as $\{f_j^{p0}\}_{j=1}^5$, respectively. This process can be expressed as

$$(C,H,W) = \text{shape}(x),\ x \in \{b_3^1, f_i^m\}, \tag{9}$$

$$f_1^{p0} = DCBR_{d=1}(b_3^1),\ f_1^{p0} \in \mathbb{R}^{C\times H\times W}, \tag{10}$$

$$f_{j+1}^{p0} = DCBR_{d=di}(f_j^{p0}),\ f_{j+1}^{p0} \in \mathbb{R}^{C\times H\times W}, \tag{11}$$

where $DCBR(\cdot)$ contains a 3×3 dilated convolution, a batch normalization layer and an ReLU activation function and $di \in \{2,3,5,7\}$ within $DCBR(\cdot)$ denote different dilation rates.

Then we use a second set of interconnected dilated convolutions to densely connect these feature sets and build a series of pyramids with regard to all the convolutions with different expansion rates. Furthermore, a set of features are obtained by shrinking the pyramid features, denoted as $\{f_j^{p1}\}_{j=1}^5$. This process can be formulated as

$$f_1^{p1} = DCBR_{d=1}(Cat(\{f_j^{p0}\}_{j=1}^5)),\quad f_1^{p1} \in \mathbb{R}^{C\times H\times W}. \tag{12}$$

$$f_{j+1}^{p1} = DCBR_{d=di}(Cat(\{f_k^{p1}\}_{k=1}^{j},\ \{f_k^{p0}\}_{k=j+1}^5)),\quad f_{j+1}^{p1} \in \mathbb{R}^{C\times H\times W}. \tag{13}$$

To obtain the multi-scale features, these features are finally fused as

$$b_3^p = DCBR_{d=1}(Cat(\{f_j^{p1}\}_{j=1}^5)),\quad b_3^p \in \mathbb{R}^{C\times H\times W}. \tag{14}$$

As a result, the proposed MDASPP produces the more discriminant feature representation.

D. Feature Refinement Module

To derive the more detailed features for the subsequent distraction mining process, we further refine the preliminary features extracted using the encoder. The feature refinement module is fulfilled in two stages. First, the refinement operation along the channel dimension is implemented using a Multi-Branch Bidirectional Fusion Block (MBBFB) that we design. Second, the refinement operation along the spatial dimension is implemented using five parallel MDASPP in terms of the five sets of features extracted using the encoder, respectively.

1) Multi-branch Bidirectional Fusion Block: To jointly exploit the multi-scale features extracted using the encoder, we design a Multi-branch Bidirectional Fusion Block (MBBFB), as shown in Fig. 5. Each branch fuses the features in two neighboring branches at the channel dimension. In essence, the MBBFB plays the role of coarse decoding. In contrast to the unidirectional fusion strategy, the bidirectional fusion strategy can strengthen the low-resolution features which are ignored in the unidirectional strategy.

Regarding the multi-scale features $\{f_i^t\}_{i=1}^5$, starting from the branch with the highest resolution features, the MBBFB

continuously fuses the features at the current branch with the low-resolution features at the next branch through the down-sampling and convolution operations. The number of channels of low-resolution features are enlarged in this process. As a result, a set of features $\{f_i^{m0}\}_{i=1}^{5}$ are produced. The above process can be expressed as

$$f_1^{m0} = f_1^t,\ f_1^{m0} \in \mathbb{R}^{64\times 256\times 256}, \tag{15}$$

$$\begin{aligned} f_{i+1}^{m0} &= CBR(Cat(Down(f_i^{m0}),\ f_{i+1}^t)), \\ f_{i+1}^{m0} &\in \mathbb{R}^{C_{i+1}\times S_i\times S_i},\ C_{i+1} = 64(i+1), S_i = \frac{256}{2^i}. \end{aligned} \tag{16}$$

Beginning with the features of the lowest resolution, MBBFB then fuses them with the high-resolution features at the next branch via the upsampling and convolution operations. During this process, the number of channels of the high-resolution features is expanded. In terms of each branch, a CBR block is further used to reduce the number of channels. Finally, five sets of features are derived, which are denoted as $\{f_i^m\}_{i=1}^{5}$. The above process can be formulated as

$$f_5^m = CBR(f_5^{m0}),\ f_5^m \in \mathbb{R}^{64\times 16\times 16}, \tag{17}$$

$$\begin{aligned} f_{5-i}^m &= CBR(Cat(CBR(Up(f_{5-i+1}^{m0})),\ f_{5-i}^{m0})), \\ f_{5-i}^m &\in \mathbb{R}^{64\times S_i\times S_i},\ S_i = 16\cdot 2^i. \end{aligned} \tag{18}$$

The proposed MBBFB aggregates high-resolution features and low-resolution features along the channel dimension. In particular, it is focused on strengthening the low-resolution features. Therefore, the richer multi-scale features can be obtained using the MBBFB. Furthermore, as shown in Fig. 4, similar to b_3^p obtained in MDASPP, $\{f_i^m\}_{i=1}^{5}$ also mines the feature information of different spatial scales through MDASPP. Finally, five sets of feature $\{f_i^p\}_{i=1}^{5}$ are derived.

E. Distracted Attention Decoder

The distracted attention decoder comprises five Boundary-Aware Distracted Attention (BADA) blocks. Each BADA block integrates the boundary features and mask extracted using the boundary decoder and the features produced by the feature refinement module. Particularly, the attention mechanism is used to strengthen the extraction of distraction features for distraction mining. The decoding process starts with low-resolution features and progressively performs the distraction mining and upsampling operations. The final high-resolution camouflaged object mask is produced by the last BADA block.

1) Boundary-Aware Distracted Attention: The feature fusion processes used for the existing distraction mining approaches were normally developed based on the simplistic addition, subtraction and group connection operations. In view of the powerful feature extraction and fusion ability that the Multi-head Self-attention (MHSA) mechanism [30] have manifested, we propose a Boundary-Aware Distracted Attention (BADA) mechanism, as shown in Fig. 6. In essence, the BADA mechanism integrates the distraction mining strategy with the self-attention mechanism. It employs the attention mechanism in order to extract the distraction features. In addition, we utilize the boundary mask and features to boost the distraction process. As a result, the camouflaged object mask presents

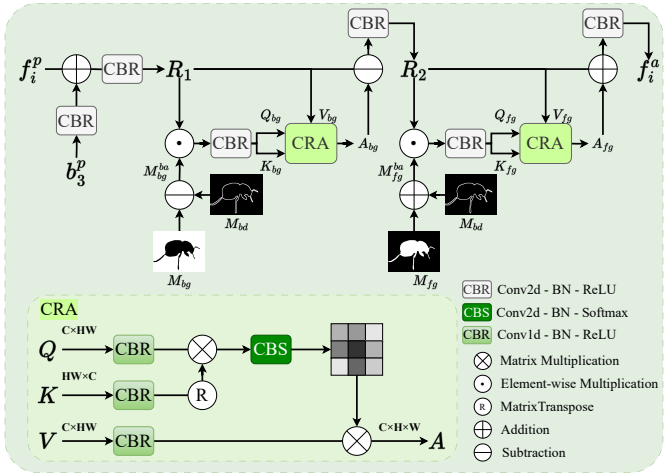

Fig. 6. The detailed structure of the BADA block.

the finer boundaries after the distraction mining process is complete.

To be specific, the input features $\{f_i^p\}_{i=1}^{5}$ are first fused with the boundary features b_3^p extracted using the boundary decoder. The resultant features is denoted as R_1. Then, the boundary mask M_{bd} extracted using the boundary decoder is subtracted from the background mask M_{bg} obtained from the prior prediction layer, to generate the boundary-aware background mask M_{bg}^{ba}. R_1 are further multiplied with the mask M_{bg}^{ba} in the element-wise manner. Given that the query and key, denoted as Q_{bg} and K_{bg}, are derived from the product while R_1 is used as the value V_{bg}, the self-attention computation is performed. The result is the refined features A_{bg}. The above process can be formulated as

$$(C,H,W) = \text{shape}(f_i^p), \tag{19}$$

$$R_1 = CBR(f_i^p + CBR(b_3^p)),\ R_1 \in \mathbb{R}^{C\times H\times W}, \tag{20}$$

$$M_{bg}^{ba} = M_{bg} - M_{bd},\ M_{bg}^{ba} \in \mathbb{R}^{H\times W}, \tag{21}$$

$$Q_{bg}, K_{bg} = CBR(R_1 \odot M_{bg}^{ba}),\ Q_{bg}, K_{bg} \in \mathbb{R}^{C\times H\times W}, \tag{22}$$

$$V_{bg} = R_1,\ V_{bg} \in \mathbb{R}^{C\times H\times W}, \tag{23}$$

$$A_{bg} = CRA(Q_{bg},\ K_{bg},\ V_{bg}),\ A_{bg} \in \mathbb{R}^{C\times H\times W}, \tag{24}$$

where $\odot$ denotes the element-wise multiplication operation and $CRA(\cdot)$ stands for the Convolution Refinement Attention (CRA) block which refines the attention weights using additional convolutions based on the MHSA mechanism.

Subsequently, the subtraction operation is conducted on R_1 and A_{bg} and the result is R_2. The foreground mask M_{fg} derived from the prior prediction layer is added with the boundary mask M_{bd}. The result is a boundary-aware foreground mask M_{fg}^{ba}. The same computation is performed on both R_2 and M_{fg}^{ba}. To derive the final features f_i^a, the resultant features A_{fg} are added to R_2.

Algorithm 1 Illustration of the forward propagation process of our BADANet.

Require: Input image I, model parameters θ
Ensure: Camouflaged object mask P_i^s, boundary map P_b

// Stage 1: Multi-scale feature extraction
$\{f_i^t\}_{i=1}^5 \leftarrow \text{PVTv2\_Encoder}(I)$
// Stage 2: Boundary decoder
$b_3^1 \leftarrow \text{BoundaryShrinkingModule}(\{f_i^t\})$
$b_3^p \leftarrow \text{MDASPP}(b_3^1)$
$M_{bd} \leftarrow \text{Sigmoid}(\text{Conv}(b_3^p))$
// Stage 3: Feature refinement
$\{f_i^p\}_{i=1}^5 \leftarrow \text{FeatureRefinementModule}(\{f_i^t\})$
// Stage 4: Distracted Attention Decoder
for $i \leftarrow 5$ **downto** 1 **do**
  $f_i^a \leftarrow \text{BADA}(f_i^p, b_3^p, M_{prev}, M_{bd})$
  $M_i \leftarrow \text{Sigmoid}(\text{Conv}(f_i^a))$
end for
// Stage 5: Output
$\{P_i^s\}_{i=1}^6 \leftarrow (M_i, M_{prev})$
$P_b \leftarrow M_{bd}$
return (P_i^s, P_b)

F. Loss Function

The loss function used for the proposed BADANet consists of a segmentation loss function and a boundary detection loss function. Following the existing studies [18], [17], [45], both the weighted BCE loss L_{BCE}^w and the weighted IOU loss L_{IOU}^w [46] are utilized as the segmentation loss function. In terms of each BADA block, the two weighted loss functions are added for the sake of enabling the model trained to pay more attention to the global structure of the image and difficult pixels [46]. In [10], [25], the dice loss function L_{dice} [47] was used as the boundary loss function. Inspired by L_{BCE}^w and L_{IOU}^w, we adapt a weighted dice loss L_{dice}^w and use it as the boundary loss function. The combined loss function L can be written as

$$L = \sum_{i=1}^{6} (L_{BCE}^w(P_i^s,\ G_i^s) + L_{IOU}^w(P_i^s,\ G_i^s)) + L_{dice}^w(P^b,\ G^b), \tag{25}$$

where $\{G_i^s\}_{i=1}^6$ and G^b represent the sixsegmentation ground-truth masks and the boundary ground-truth mask, respectively, and $\{P_i^s\}_{i=1}^6$ and P^b denote the corresponding six segmentation masks and the boundary masks produced by the model, respectively.

To facilitate a clearer understanding of our BADANet, the forward propagation process is illustrated in Algorithm 1.

IV. Experiments

In this section, we first elaborate the experimental setup, including data sets, evaluation metrics and implementation details. Then we compare the proposed method with 18 baseline approaches quantitatively and qualitatively and perform a series of analysis. Finally, we conduct an ablation study to validate the effectiveness of the components of the BADANet.

A. Experimental Setup

1) Data Sets: We used four publicly available COD data sets, including CAMO [52], CHAMELEON [53], COD10K [1] and NC4K [24]. The CAMO data set comprises 1,250 camouflaged and 1,250 non-camouflaged images. Only 76 manually-annotated camouflaged images are contained in the CHAMELEON data set. The COD10K data set consists of 5,066 camouflaged images, 3,000 background images and 1,934 non-camouflaged images. The NC4K data set comprises 4,121 camouflaged images. In accordance with the data splitting scheme used in the previous studies [1], [17], [3], a training set was built on top of the 1,000 camouflaged images in the CAMO [52] data set and the 3,040 camouflaged images in the COD10K data set, while the remaining images in the four data sets were utilized as the testing set.

2) Evaluation Metrics: Following the existing studies [40], [16], [20], we used four commonly-used COD evaluation metrics, including the structure measure S_α [54], mean F-measure F_β [55], mean E-measure E_ϕ [56] and mean absolute error M.

3) Implementation Details: The proposed BADANet was implemented using Pytorch. The pre-trained PVTv2 [19] was used as the backbone network. We utilized the AdamW optimizer [57] with the initial learning rate of 2.5e-5. The learning rate followed a linear warm-up and linear decay strategy. During the training process, the images were resized to the resolution of 512×512 pixels. The random flipping and rotation operations were employed for data augmentation. The network was end-to-end trained for 80 epochs on a single NVIDIA 3090 GPU. The batch size was set to 6.

B. Experimental Results

In this section, we report the results obtained using our BADANet, along with 18 baselines, within the quantitative and qualitative comparisons, robustness analysis, computational complexity analysis and failure case analysis.

1) Quantitative Comparison: In Table I, the S_α, F_β, E_ϕ and M values obtained using the proposed BADANet, together with 18 baselines, on the four COD data sets are presented. As can be seen, our method normally outperformed its counterparts across the four data sets no matter which data set was used. Compared with PUENet [20], the average improvement of our method on the four data sets in terms of the S_α, F_β, E_ϕ and M metrics were 1.8%, 2.3%, 0.6% and 9.4%, respectively. In contrast to FSNet [40] and HitNet [41], the four values were 2.2%, 2.6%, 0.5% and 9.4%, and 3.0%, 2.6%, 1.6% and 20.9%, respectively. It should be noted that PUENet [20] utilized the distraction mining strategy but did not utilize the boundary cue. Although both FSNet [40] and HitNet [41] employed multi-stage refinement approaches, they did not exploit the distraction mining strategy. In contrast, the proposed method jointly utilized both the boundary cue and the distraction mining strategy.

2) Qualitative Comparison: The detection masks obtained using the proposed BADANet and six state-of-the-art COD methods are shown in Fig. 7. It can be observed that our method produced the more accurate COD results on the large

TABLE I
QUANTITATIVE COMPARISON BETWEEN OUR METHOD AND 18 BASELINE METHODS. THE TOP THREE RESULTS ARE INDICATED IN THE **RED**, GREEN AND *Blue* FONTS, RESPECTIVELY.

Method	Publication	Backbone	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
			$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
SINet [1]	CVPR2020	ResNet50	0.745	0.702	0.804	0.092	0.872	0.827	0.936	0.034	0.776	0.679	0.864	0.043	0.810	0.772	0.873	0.057
PFNet [2]	CVPR2021	ResNet50	0.782	0.744	0.840	0.085	0.882	0.826	0.922	0.033	0.800	0.700	0.875	0.040	0.829	0.782	0.886	0.053
C^2FNet [48]	IJCAI2021	Res2Net50	0.796	0.762	0.854	0.080	0.888	0.844	0.935	0.032	0.813	0.723	0.890	0.036	0.838	0.795	0.897	0.049
UGTR [36]	ICCV2021	ResNet50	0.784	0.686	0.859	0.086	0.888	0.796	0.918	0.031	0.817	0.667	0.853	0.036	0.839	0.786	0.875	0.052
LSR [24]	CVPR2021	ResNet50	0.787	0.725	0.838	0.080	0.893	0.839	0.938	0.033	0.804	0.685	0.880	0.037	0.840	0.779	0.895	0.048
MGL-R [49]	CVPR2021	ResNet50	0.775	0.726	0.812	0.088	0.893	0.834	0.918	0.031	0.814	0.711	0.852	0.035	0.833	0.782	0.867	0.053
SINetV2 [18]	TPAMI2022	Res2Net50	0.820	0.782	0.882	0.070	0.888	0.835	0.942	0.030	0.815	0.718	0.887	0.037	0.847	0.805	0.903	0.048
PreyNet [50]	ACM MM2022	Res2Net50	0.813	0.793	0.876	0.071	0.902	0.866	0.951	0.027	0.830	0.741	0.895	0.032	0.838	0.798	0.887	0.047
SegMaR [51]	CVPR2022	ResNet50	0.815	0.795	0.874	0.071	0.906	0.872	0.951	0.025	0.833	0.757	0.899	0.033	0.841	0.821	0.896	0.046
BGNet [25]	IJCAI2022	Res2Net50	0.812	0.789	0.870	0.073	0.901	0.860	0.943	0.027	0.831	0.753	0.901	0.033	0.851	0.820	0.907	0.044
BSANet [42]	AAAI2022	Res2Net50	0.796	0.763	0.851	0.079	0.895	0.858	0.946	0.027	0.818	0.738	0.891	0.034	0.841	0.808	0.897	0.048
ZoomNet [3]	CVPR2022	ResNet50	0.820	0.794	0.877	0.066	0.902	0.864	0.943	0.023	0.838	0.766	0.888	0.029	0.853	0.818	0.896	0.043
FEDER [10]	CVPR2023	ResNet50	0.807	0.781	0.873	0.069	0.894	0.851	0.947	0.028	0.823	0.751	0.900	0.032	0.846	0.824	0.905	0.045
DGNet [17]	MIR2023	EfficientNet	0.839	0.806	0.901	0.057	0.890	0.834	0.938	0.029	0.822	0.728	0.896	0.033	0.857	0.814	0.911	0.042
FSPNet [16]	CVPR2023	ViT	0.856	0.830	0.899	*0.050*	*0.914*	*0.879*	*0.960*	*0.022*	0.851	0.769	0.895	0.026	0.879	0.843	0.915	*0.035*
PUENet [20]	TIP2023	Res2Net50+ViT	*0.877*	*0.860*	*0.930*	0.045	0.910	0.869	0.957	*0.022*	0.873	*0.812*	0.938	0.022	0.898	0.874	**0.945**	**0.028**
FSNet [40]	TIP2023	SwinT	0.880	0.861	0.933	**0.041**	0.905	0.868	0.963	*0.022*	0.870	0.810	0.938	*0.023*	*0.891*	*0.866*	*0.940*	0.031
HitNet [41]	AAAI2023	PVT	0.849	0.831	0.906	0.055	0.921	0.900	**0.967**	0.019	*0.871*	0.823	*0.935*	*0.023*	0.875	0.853	0.926	0.037
BADANet	Ours	PVTv2	**0.891**	**0.870**	**0.938**	**0.041**	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	0.943	**0.028**

object (see Row 1), medium object (see Row 2) and small object (see Row 3). This finding should be due to the multi-scale feature fusion achieved by the multi-branch fusion and the spatial scale feature mining used in our method. On the other hand, our method also exhibited the finer boundary localization and foreground-background discrimination ability in the scenes with blurred boundaries (see Rows 4 and 5) and the scenes with the highly similar colors and textures between the foreground and background (see Rows 6 and 7). These advantages should be attributed to the injection of the boundary characteristics and the ability to separate the foreground from the background due to the proposed distracted attention mechanism.

In addition, the proposed method manifested advantages on occluded objects (see Rows 8 and 9). This observation should result from the injected boundary information which provided more references for the demarcation between the object and the occluder. Although BGNet [25] and FEDER [10] which also exploited the boundary information performed properly on occlusion objects, some details were still missed, compared with the results that our method produced. This finding suggests that the proposed distracted attention plays an important role on capturing the detailed characteristics.

3) Robustness Analysis: To rigorously assess the robustness of the proposed method, we trained and tested the network with 10 different random seeds (0–9) individually. Regarding the four performance metrics, we calculated the mean and standard deviation with the 95% confidence interval (95% CI). As shown in Table II, the variations across the 10 runs are negligible, reflected by the minor standard deviations and confidence intervals. It is indicated that our method is robust to the random initialization of the model. In this context, the proposed method was not only highly stable under varying random initializations, but also achieved consistent and reliable performance gains over the baseline methods.

4) Computational Complexity Analysis: We compared the proposed BADANet with four state-of-the-art COD methods in computational complexity, including FLOPs, number of parameters and inference speed. The results are shown in Table III. Although our model incurred the second highest FLOPs value (635.6G), it contained a moderate number of parameters (78.8M) and showed a real-time detection speed (12.5 FPS). More importantly, our BADANet normally surpassed its counterparts across the four data sets in terms of four different performance metrics (see Table I). Compared with FSNet [40], our model required less than half the parameters (78.8M vs. 170.2M) but achieved the higher detection accuracy with a comparable inference speed. In addition, our BADANet showed a notable accuracy advantage over HitNet [41], which used the same backbone, even if the computational complexity of our method was heavier than that of HitNet. These results demonstrate that our BADANet strikes a proper balance between detection accuracy and computational complexity.

5) Failure Cases: Although our BADANet has achieved state-of-the-art performance on the four data sets, it still encountered challenges, as illustrated in Fig. 8. The first and second rows show two examples of *under-segmentation*, where the camouflaged objects were not completely detected. The third row presents a case of *mis-segmentation*, in which the three methods failed to detect all the camouflaged objects

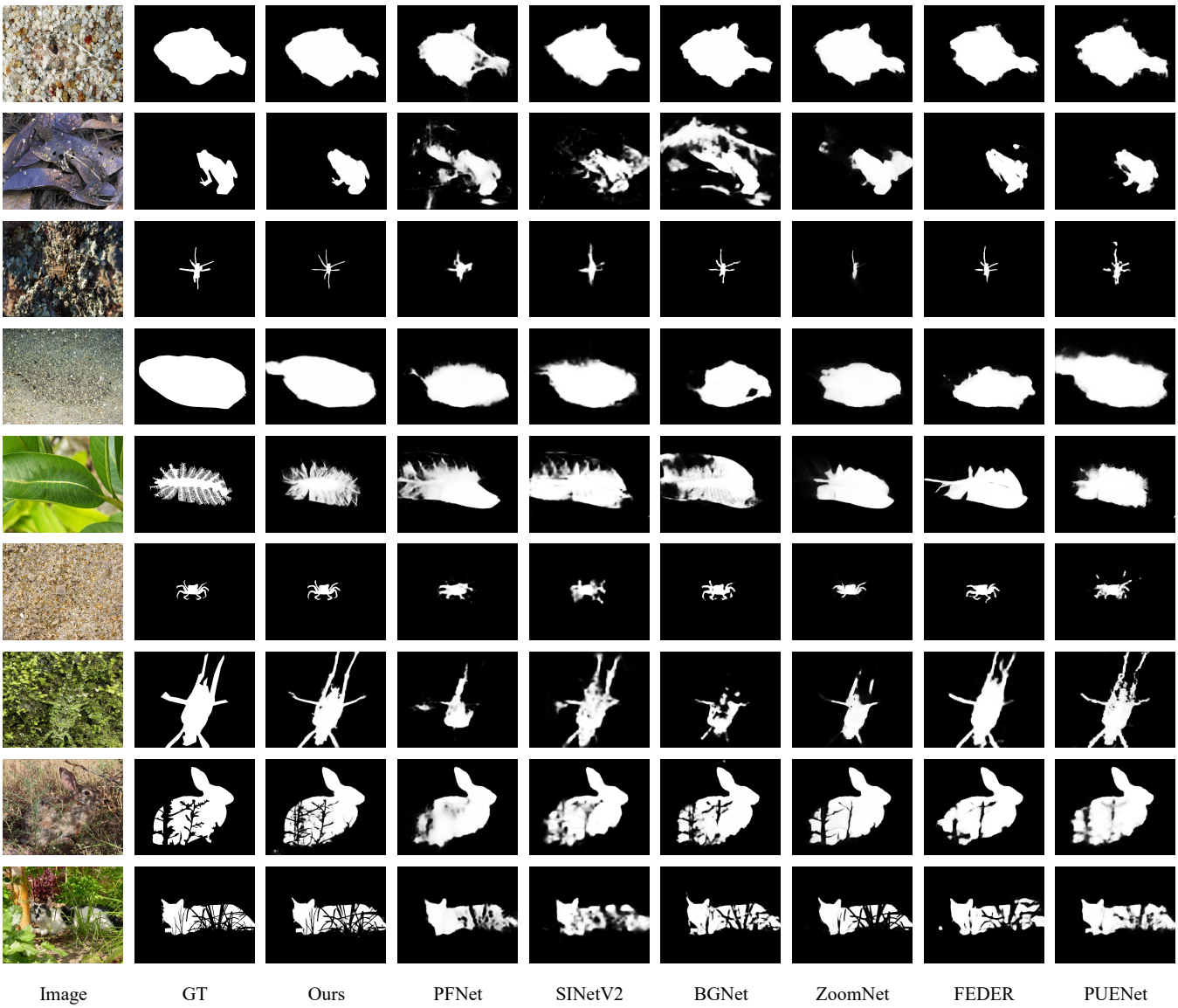

Fig. 7. Visual comparison of the masks produced by our method and six state-of-the-art COD methods. (For the best view, please zoom in).

TABLE II
THE MEAN AND STANDARD DEVIATION WITH THE 95% CONFIDENCE INTERVAL IN TERMS OF EACH PERFORMANCE METRIC OBTAINED USING THE PROPOSED BADANET ALONG WITH 10 RANDOM SEEDS ON THE FOUR COD DATA SETS.

Metric	CAMO-Test (250)	CHAMELEON (76)	COD10K-Test (2026)	NC4K (4121)
$S_\alpha \uparrow$	0.891 ± 0.0019 [0.889, 0.892]	0.930 ± 0.0015 [0.929, 0.931]	0.896 ± 0.0007 [0.895, 0.896]	0.905 ± 0.0010 [0.904, 0.906]
$F_\beta \uparrow$	0.870 ± 0.0022 [0.868, 0.871]	0.901 ± 0.0031 [0.898, 0.903]	0.843 ± 0.0026 [0.841, 0.845]	0.881 ± 0.0013 [0.880, 0.882]
$E_\phi \uparrow$	0.938 ± 0.0016 [0.937, 0.939]	0.967 ± 0.0026 [0.964, 0.968]	0.946 ± 0.0016 [0.945, 0.947]	0.943 ± 0.0010 [0.943, 0.944]
$M \downarrow$	0.041 ± 0.0007 [0.041, 0.042]	0.018 ± 0.0012 [0.017, 0.019]	0.019 ± 0.0003 [0.019, 0.019]	0.028 ± 0.0006 [0.028, 0.029]

TABLE III
COMPARISON BETWEEN OUR BADANET AND FOUR STATE-OF-THE-ART COD METHODS IN COMPUTATIONAL COMPLEXITY (FLOPS), NUMBER OF PARAMETERS AND INFERENCE SPEED.

Method	Backbone	FLOPs (G)	Params (M)	Speed (FPS)
FEDER	ResNet50	128.2	44.1	12.6
BGNet	Res2Net50	117.0	79.9	16.5
FSNet	SwinT	796.1	170.2	15.7
HitNet	PVTv2	112.5	25.7	19.4
BADANet (Ours)	PVTv2	635.6	78.8	12.5

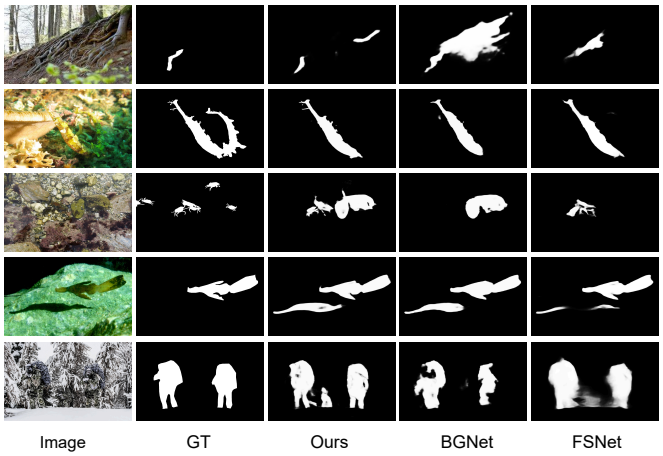

Fig. 8. Five camouflaged images on which not only our BADANet but also both the BGNet [25] and the FSNet [40] failed to some extent.

while they wrongly recognized the background as camouflaged objects. The fourth row further demonstrates a case of *over-segmentation*, where the shadow of the camouflaged object was misidentified as one or two camouflaged objects. The fifth row highlights an issue of *annotation error*, in which only two camouflaged armed men were annotated in the ground-truth data even though three armed men were visible. In this case, our method correctly detected the unannotated armed man. However, this detection was treated as a false positive under the current benchmark. In contrast, FSNet [40] detected the third armed man with less precision, while BGNet [25] completely missed the armed man.

We attribute these failure cases to four factors. First, the extremely cluttered background and the subtle similarity between objects and surroundings in texture posed a great challenge to the discriminative ability of the methods. Second, limited training samples for rare or highly camouflaged patterns restricted the generalization of these methods. Third, those methods struggled with the ambiguity in object boundaries, in particular, when reflections or shadows were present. Fourth, occasional annotation inconsistencies occurred in the data sets, which not only introduced noise to the training process but also penalized correct detections during the inference process. In this situation, however, the performance of our method was similar or even superior to that of BGNet [25] and FSNet [40].

6) Visualization of Feature and Boundary Maps: To provide an intuitive understanding of the inherent mechanism of our model, we visualized the feature maps extracted using the MBBFB at five different scales and the boundary map M_{bd} produced by the boundary decoder in Fig. 9. As can be seen, our model tends to detect semantically meaningful regions across scales while preserving detailed boundary information.

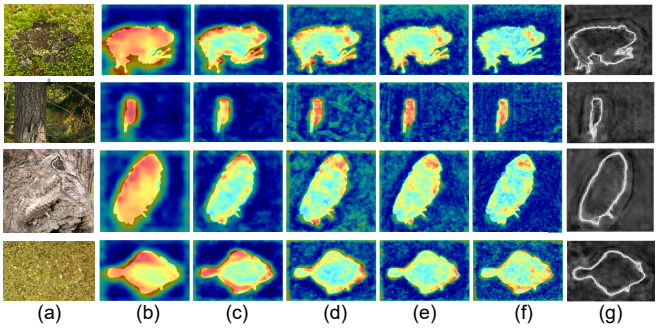

Fig. 9. Visualization of the feature maps produced by the MBBFB of our model at five different scales and the boundary map M_{bd} obtained using the boundary decoder. Within each row, (a) shows an original camouflaged images, (b-f) present the five feature maps extracted from this image and (g) shows the associated boundary map.

C. Ablation Study

To investigate the effectiveness of different components of the proposed BADANet, we performed a series of ablation experiments on the four COD data sets.

1) Effect of the MBBFB: We examined the impact of the MBBFB by removing it from the proposed BADANet. As depicted in Table IV, the performance of the network dropped without using the MBBFB, compared with the the BADANet which used the MBBFB. It was suggested that the abundant multi-scale features produced by the MBBFB were useful.

2) Effect of the MDASPP: We conducted an ablation experiment to evaluate the effectiveness of the proposed MDASPP module by comparing it with the standard ASPP module [43] and the deformable ASPP (DASPP) module [44]. As shown in Table V, the MDASPP consistently achieved the best performance across all the four data sets in terms of different metrics. It is indicated that our MDASPP better captured multi-scale contextual information than its counterparts.

To investigate the impact of the dilation rate combination on the MDASPP module, we further performed an additional ablation experiment. Specifically, we tested three combinations of dilation rates, including [1,2,3,4,5], [1,3,6,12,18] and [1,2,3,5,7]. As reported in Table VI, the performance differences between these combinations are relatively minor. It is suggested that the MDASPP is robust to the choice of dilation rates. This finding should be due to the fact that the multi-dilation design can effectively capture contextual information across different scales while minor variations in the dilation configuration do not affect detection performance obviously.

3) Effect of the BADA: To examine the impact of the BADA, we first removed the boundary data fed into the BADANet. Then we obtained three variants of the BADANet by removing the BADA, replacing the BADA by the Focus Module (FM) of the PFNet [2] and retaining the BADA. The results derived using these variants and our BADANet are reported in Table

TABLE IV
THE EFFECT OF THE MBBFB ON THE PROPOSED BADANET ON FOUR COD DATA SETS.

Method	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
w/o MBBFB	0.885	0.867	0.934	0.043	0.921	0.886	0.953	0.021	0.893	0.840	0.942	**0.019**	0.901	0.878	**0.943**	0.029
w/ MBBFB	**0.891**	**0.870**	**0.938**	**0.041**	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	**0.943**	**0.028**

TABLE V
THE COMPARISON BETWEEN THE PROPOSED MDASPP MODULE AND THE STANDARD ASPP MODULE [43] AND THE DEFORMABLE ASPP (DASPP) MODULE [44] ACROSS FOUR COD DATA SETS WHEN EACH MODULE WAS USED WITH OUR BADANET.

Method	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
ASPP [43]	0.879	0.858	0.928	0.047	0.925	0.892	0.960	0.021	0.892	0.837	0.941	0.020	0.897	0.873	0.937	0.031
DASPP [44]	0.884	0.859	0.925	0.046	0.927	0.898	0.963	0.020	0.890	0.833	0.939	0.021	0.899	0.872	0.937	0.031
MDASPP	**0.891**	**0.870**	**0.938**	**0.041**	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	**0.943**	**0.028**

TABLE VI
THE COMPARISON OF THREE COMBINATIONS OF DILATION RATES USED IN THE MDASPP.

Dilation Rates	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
[1,2,3,4,5]	0.890	**0.872**	0.939	0.041	0.929	0.898	0.959	0.019	0.895	0.843	0.945	**0.019**	0.904	**0.882**	**0.944**	**0.028**
[1,3,6,12,18]	**0.892**	**0.872**	**0.940**	**0.040**	0.925	0.891	0.959	0.020	0.895	**0.844**	**0.946**	**0.019**	0.903	0.880	0.942	0.029
[1,2,3,5,7]	0.891	0.870	0.938	0.041	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	0.843	**0.946**	**0.019**	**0.905**	0.881	0.943	**0.028**

TABLE VII
THE EFFECT OF THE BADA AND THE BOUNDARY DATA ON THE PROPOSED BADANET.

Method	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
w/o Boundary w/o BADA	0.882	0.855	0.925	0.048	0.926	0.892	0.958	0.021	0.889	0.827	0.937	0.021	0.898	0.869	0.936	0.031
w/o Boundary & BADA $\rightarrow$ FM	0.879	0.855	0.922	0.048	0.923	0.886	0.952	0.022	0.887	0.828	0.936	0.021	0.897	0.872	0.934	0.032
w/o Boundary w/ BADA	0.875	0.852	0.920	0.049	0.926	0.898	0.963	0.020	0.888	0.832	0.939	0.020	0.896	0.871	0.935	0.031
w/ Boundary & w/ BADA	**0.891**	**0.870**	**0.938**	**0.041**	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	**0.943**	**0.028**

VII. It can be observed that the application of either the BADA or the boundary data boosted the performance of the network. It has been demonstrated that the proposed BADA was effective, in particular, the boundary data was available.

We further investigated the attention mechanism utilized in the BADA. Specifically, we compared the vanilla MHSA [28] with our Convolution Refinement Attention (CRA) block. As shown in Table VIII, the CRA block consistently outperformed the Vanilla MHSA mechanism across the four data sets. These results verified the effectiveness of the incorporation of convolutional refinement into the attention mechanism.

To assess the impact of boundary thickness on the BADA module, we also varied the thickness of M_{bd} in different pixels. The results are reported in Table IX. It can be seen that different thickness values only led to marginal variations across the four data sets. This observation demonstrates that the BADA module is robust to the setting of boundary thickness.

4) Effect of the Backbone Network: We evaluated the performance of our BADANet with five different backbone networks. As reported in Table X, it is shown that the proposed BADANet normally achieved the best result when PVTv2 [19] was employed as the backbone across the four data sets. This finding highlights the advantage of integrating our BADANet with a powerful backbone.

V. CONCLUSION

In this study, we addressed the challenges that the existing distraction mining methods normally encounter. That is to say, insufficient feature fusion and the absence of the boundary cue during the distraction mining process. To this end, we introduced a Boundary-Aware Distracted Attention Network (BADANet). This network utilizes both the boundary cue and the distraction mining strategy. Our BADANet commences with a pre-trained encoder. The features extracted using this encoder are then sent to a Boundary Shrinking Module (BSM) that we designed, which is followed by a Multiple Dense Atrous Spatial Pyramid Pooling (MDASPP) module. The output is a boundary map of the input image. In terms of an encoder block, the features extracted are also fed into a Multi-branch Bidirectional Fusion Block (MBBFB). This block can be used to achieve the comprehensive bidirectional fusion of multi-scale features across the channel dimension. The features fused are further passed through a series of MDASPP modules with regard to different encoder blocks, to excavate

TABLE VIII
THE COMPARISON BETWEEN THE CRA BLOCK AND THE VANILLA MHSA MECHANISM.

Method	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
Vanilla MHSA [28]	0.885	0.855	0.925	0.047	0.925	0.889	0.957	0.021	0.890	0.831	0.938	0.021	0.899	0.869	0.934	0.032
CRA	**0.891**	**0.870**	**0.938**	**0.041**	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	**0.943**	**0.028**

TABLE IX
THE EFFECT OF BOUNDARY THICKNESS ON THE BADA MODULE.

Thickness	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
1	**0.891**	**0.870**	**0.938**	**0.041**	0.930	0.901	0.967	**0.018**	0.896	0.843	0.946	**0.019**	**0.905**	0.881	**0.943**	**0.028**
3	0.886	0.868	0.937	0.043	**0.931**	**0.905**	**0.971**	**0.018**	0.895	0.843	0.947	**0.019**	0.903	0.881	**0.943**	**0.028**
5	0.889	0.868	0.937	0.042	0.928	0.900	0.965	0.020	**0.897**	**0.848**	**0.948**	**0.019**	0.904	**0.882**	**0.943**	0.029
7	0.885	0.865	0.932	0.045	0.929	0.900	0.961	0.019	**0.897**	0.846	0.947	**0.019**	**0.905**	**0.882**	**0.943**	0.029

TABLE X
THE COMPARISON OF FIVE BACKBONE NETWORKS WHEN EACH OF THEM WAS USED TOGETHER WITH THE PROPOSED BADANET.

Backbone	CAMO-Test (250)				CHAMELEON (76)				COD10K-Test (2026)				NC4K (4121)			
	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta\uparrow$	$E_\phi\uparrow$	$M\downarrow$
ResNet50 [58]	0.814	0.766	0.853	0.078	0.898	0.842	0.938	0.031	0.843	0.762	0.901	0.033	0.857	0.812	0.897	0.047
Res2Net50 [38]	0.805	0.761	0.848	0.080	0.894	0.840	0.923	0.034	0.839	0.755	0.896	0.034	0.861	0.818	0.902	0.047
SwinT [31]	0.890	**0.870**	0.934	**0.040**	0.922	*0.889*	0.958	*0.020*	0.888	0.831	0.943	0.020	0.903	0.878	**0.943**	0.029
PVT [32]	0.862	0.846	0.912	0.052	0.914	0.878	0.952	0.023	*0.875*	0.816	0.933	0.023	0.884	0.856	0.927	0.036
PVTv2 [19]	**0.891**	**0.870**	**0.938**	0.041	**0.930**	**0.901**	**0.967**	**0.018**	**0.896**	**0.843**	**0.946**	**0.019**	**0.905**	**0.881**	**0.943**	**0.028**

the features at diverse spatial scales. With regard to each MDASPP module, a Boundary-Aware Distracted Attention (BADA) block that we introduced receives both the features that the module produced and the boundary map. The BADA block integrates the distraction mining mechanism and the boundary information in order to reinforce the boundary of an object. Finally, the last BADA block generates the detection mask. Extensive experiments have been performed on four popular COD data sets. The results showed that the proposed BADANet normally outperformed the 18 baselines that we tested. We believe that the promising results should benefit from the boundary-aware distracted attention mechanism that we deliberately introduced.

REFERENCES

[1] 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.

[2] 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.

[3] Y. Pang, X. Zhao, T.-Z. Xiang, L. Zhang, and H. Lu, "Zoom in and out: A mixed-scale triplet network for camouflaged object detection," in *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, 2022, pp. 2160–2170.

[4] G. Yue, H. Xiao, H. Xie, T. Zhou, W. Zhou, W. Yan, B. Zhao, T. Wang, and Q. Jiang, "Dual-constraint coarse-to-fine network for camouflaged object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[5] F. Wang, Y. Su, R. Wang, J. Sun, F. Sun, and H. Li, "Cross-modal and cross-level attention interaction network for salient object detection," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 6, pp. 2907–2920, 2024.

[6] S. Bhuyan, A. Kar, D. Sen, and S. Deb, "Rgb-d fusion through zero-shot fuzzy membership learning for salient object detection," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 7, pp. 3638–3652, 2024.

[7] W. Zhou, B. Wang, X. Dong, C. Xu, and F. Qiang, "Location, neighborhood, and semantic guidance network for rgb-d co-salient object detection," *IEEE Transactions on Artificial Intelligence*, pp. 1–14, 2025.

[8] W. Weng, M. Wei, J. Ren, and F. Shen, "Enhancing aerial object detection with selective frequency interaction network," *IEEE Transactions on Artificial Intelligence*, vol. 5, no. 12, pp. 6109–6120, 2024.

[9] Y. Yin, Z. Yuan, Y. He, and X. Bao, "Vodacbd: Vehicle object detection based on adaptive convolution and bifurcation decoupling," *IEEE Transactions on Artificial Intelligence*, vol. 6, no. 5, pp. 1298–1308, 2025.

[10] C. He, K. Li, Y. Zhang, L. Tang, Y. Zhang, Z. Guo, and X. Li, "Camouflaged object detection with feature decomposition and edge reconstruction," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 22 046–22 055.

[11] G.-P. Ji, G. Xiao, Y.-C. Chou, D.-P. Fan, K. Zhao, G. Chen, and L. Van Gool, "Video polyp segmentation: A deep learning perspective," *Machine Intelligence Research*, vol. 19, no. 6, pp. 531–549, 2022.

[12] 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 transactions on medical imaging*, vol. 39, no. 8, pp. 2626–2637, 2020.

[13] N. U. Bhajantri and P. Nagabhushan, "Camouflage defect identification: a novel approach," in *9th International Conference on Information Technology (ICIT'06)*. IEEE, 2006, pp. 145–148.

[14] J. Xiao, L. Qiao, R. Stolkin, and A. Leonardis, "Distractor-supported single target tracking in extremely cluttered scenes," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 2016, pp. 121–136.

[15] D. J. A. Rustia, C. E. Lin, J.-Y. Chung, Y.-J. Zhuang, J.-C. Hsu, and T.-T. Lin, "Application of an image and environmental sensor network for automated greenhouse insect pest monitoring," *Journal of Asia-Pacific Entomology*, vol. 23, no. 1, pp. 17–28, 2020.

[16] Z. Huang, H. Dai, T.-Z. Xiang, S. Wang, H.-X. Chen, J. Qin, and H. Xiong, "Feature shrinkage pyramid for camouflaged object detection

with transformers," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2023, pp. 5557–5566.
[17] G.-P. Ji, D.-P. Fan, Y.-C. Chou, D. Dai, A. Liniger, and L. Van Gool, "Deep gradient learning for efficient camouflaged object detection," *Machine Intelligence Research*, vol. 20, no. 1, pp. 92–108, 2023.
[18] D. Fan, G. Ji, M. Cheng, and L. Shao, "Concealed object detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 10, pp. 6024–6042, 2022.
[19] 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.
[20] Y. Zhang, J. Zhang, W. Hamidouche, and O. Deforges, "Predictive uncertainty estimation for camouflaged object detection," *IEEE Transactions on Image Processing*, 2023.
[21] B. Yin, X. Zhang, Q. Hou, B.-Y. Sun, D.-P. Fan, and L. Van Gool, "Camoformer: Masked separable attention for camouflaged object detection," *arXiv preprint arXiv:2212.06570*, 2022.
[22] J. Y. Y. H. W. Hou and J. Li, "Detection of the mobile object with camouflage color under dynamic background based on optical flow," *Procedia Engineering*, vol. 15, pp. 2201–2205, 2011.
[23] Y. Pan, Y. Chen, Q. Fu, P. Zhang, X. Xu *et al.*, "Study on the camouflaged target detection method based on 3d convexity," *Modern Applied Science*, vol. 5, no. 4, p. 152, 2011.
[24] 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.
[25] Y. Sun, S. Wang, C. Chen, and T.-Z. Xiang, "Boundary-guided camouflaged object detection," *arXiv preprint arXiv:2207.00794*, 2022.
[26] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suarez, M. Jagersand, and L. Shao, "Boundary-aware segmentation network for mobile and web applications," *arXiv preprint arXiv:2101.04704*, 2021.
[27] Y. Zhong, B. Li, L. Tang, S. Kuang, S. Wu, and S. Ding, "Detecting camouflaged object in frequency domain," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 4504–4513.
[28] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.
[29] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European conference on computer vision*. Springer, 2020, pp. 213–229.
[30] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly *et al.*, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.
[31] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 10 012–10 022.
[32] W. Wang, E. Xie, X. Li, D.-P. Fan, K. Song, D. Liang, T. Lu, P. Luo, and L. Shao, "Pyramid vision transformer: A versatile backbone for dense prediction without convolutions," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 568–578.
[33] D. Zhang, H. Zhang, J. Tang, M. Wang, X. Hua, and Q. Sun, "Feature pyramid transformer," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXVIII 16*. Springer, 2020, pp. 323–339.
[34] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2021, pp. 7262–7272.
[35] X. Chen, B. Yan, J. Zhu, D. Wang, X. Yang, and H. Lu, "Transformer tracking," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 8126–8135.
[36] 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.
[37] Q. Zhang, Y. Ge, C. Zhang, and H. Bi, "Tprnet: camouflaged object detection via transformer-induced progressive refinement network," *The Visual Computer*, pp. 1–15, 2022.
[38] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE transactions on pattern analysis and machine intelligence*, vol. 43, no. 2, pp. 652–662, 2019.
[39] Z. Liu, Z. Zhang, Y. Tan, and W. Wu, "Boosting camouflaged object detection with dual-task interactive transformer," in *2022 26th International Conference on Pattern Recognition (ICPR)*. IEEE, 2022, pp. 140–146.
[40] Z. Song, X. Kang, X. Wei, H. Liu, R. Dian, and S. Li, "Fsnet: Focus scanning network for camouflaged object detection," *IEEE Transactions on Image Processing*, vol. 32, pp. 2267–2278, 2023.
[41] X. Hu, S. Wang, X. Qin, H. Dai, W. Ren, D. Luo, Y. Tai, and L. Shao, "High-resolution iterative feedback network for camouflaged object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, no. 1, 2023, pp. 881–889.
[42] 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.
[43] L.-C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," *arXiv preprint arXiv:1706.05587*, 2017.
[44] M. Yang, K. Yu, C. Zhang, Z. Li, and K. Yang, "Denseaspp for semantic segmentation in street scenes," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 3684–3692.
[45] 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.
[46] 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.
[47] 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.
[48] 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.
[49] 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.
[50] 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.
[51] 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.
[52] 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.
[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] 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.
[55] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 248–255.
[56] 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.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.
[58] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.