Camouflaged Object Detection via Context-Aware Cross-Level Fusion

Geng Chen, Si-Jie Liu, Yu-Jia Sun, Ge-Peng Ji, Ya-Feng Wu, and Tao Zhou, Member, IEEE

Abstract—Camouflaged object detection (COD) aims to identify the objects that conceal themselves in natural scenes. Accurate COD suffers from a number of challenges associated with low boundary contrast and the large variation of object appearances, e.g., object size and shape. To address these challenges, we propose a novel Context-aware Cross-level Fusion Network (C2F-Net), which fuses context-aware cross-level features for accurately identifying camouflaged objects. Specifically, we compute informative attention coefficients from multi-level features with our Attention-induced Cross-level Fusion Module (ACFM), which further integrates the features under the guidance of attention coefficients. We then propose a Dual-branch Global Context Module (DGCM) to refine the fused features for informative feature representations by exploiting rich global context information. Multiple ACFMs and DGCMs are integrated in a cascaded manner for generating a coarse prediction from high-level features. The coarse prediction acts as an attention map to refine the low-level features before passing them to our Camouflage Inference Module (CIM) to generate the final prediction. We perform extensive experiments on three widely used benchmark datasets and compare C2F-Net with state-of-the-art (SOTA) models. The results show that C2F-Net is an effective COD model and outperforms SOTA models remarkably. Further, an evaluation on polysegmentation datasets demonstrates the promising potentials of our C2F-Net in COD downstream applications. Our code is publicly available at: https://github.com/Ben57882/C2FNet-TSCVT

Index Terms—Camouflaged object detection, context-aware deep learning, feature fusion, polysegmentation.

I. INTRODUCTION

Over the course of evolution, wild animals have developed extensive camouflage abilities in order to survive. In practice, they try to alter their appearance to “perfectly” blend in with their surroundings in order to avoid attracting the notice of other creatures. Recently, camouflage has attracted increasing research interest from the computer vision community [3], [15], [34]. Among various topics, camouflaged object detection (COD), which aims to identify and segment camouflaged objects from images, is particularly popular. However, accurate COD is with considerable difficulties due to the special characteristics of camouflage. More specifically, due to the camouflage, the boundary contrast between an object and its surroundings is extremely low, leading to significant difficulties to identify/segment this object. As shown in the top row of Fig. 1, it is very challenging to discover the butterfly from the leaves. Also, in the middle row of Fig. 1, it is a tremendous task to find out the tortoise occluded by leaves on the ground. In addition, the camouflaged objects, most wild animals, are usually with varied appearances, e.g., size and shape, which further aggravates the difficulties in accurate COD.

To tackle these challenges, deep learning techniques [69] have been adopted and shown great potential. In addition, a number of COD datasets have been constructed for training the deep learning models. For instance, Le et al. [34] created the first COD dataset, called CAMO, consisting of 2,500 images. However, CAMO, which suffers from a limited sample size, is insufficient for taking full advantage of deep learning models. Recently, Fan et al. [15] constructed COD10K, the first large-scale COD dataset consisting of 10,000 images with the consideration of various challenging camouflage attributes in real natural environments. In addition to datasets, these two works also contribute to the COD from the aspect of deep learning models. Le et al. [34] proposed an anabranch network, which classifies whether an image contains camouflaged objects and then integrates this information into the COD task. Motivated by the fact that predators discover preys by searching first and then recognizing, Fan et al. [15] designed searching modules to identify the rough areas of camouflaged objects and then accurately segmented camouflaged objects with identification modules.

Although existing methods have shown promising performance in the detection of a single camouflaged object from relatively simple scenes, their performance degrades for a number of challenging cases, e.g., occlusion, multiple objects, etc. As shown in the top two rows of Fig. 1, state-of-the-
We propose C2F-Net, a novel COD model that integrates the cross-level features by considering rich global context information, effective cross-level fusion, and low-level feature refinement.

- We propose DGCM, a context-aware module that can capture valuable context information for improving the accuracy of COD. Furthermore, we integrate the cross-level features with ACFM, a novel fusion module that integrates the features with the valuable attention cues provided by MSCA.

- We propose CIM, an effective module capable of exploiting informative features from the low-level features refined by a coarse prediction map.

- Extensive experiments on three benchmark datasets demonstrate that our C2F-Net outperforms 19 state-of-the-art models in the terms of five evaluation metrics for the COD task. Further poly segmentation experiments indicate the superior performance of our C2F-Net in COD downstream applications.

This paper significantly extends our previous work published in the IJCAI 2021 [52], with multi-fold improvements as follows. (i) We improve the model by refining the low-level features with the coarse COD map and then predicting the final result with our CIM. The new improvement is detailed in Sec. III-D and has shown promising performance in both quantitative and qualitative evaluations (see Sec. IV). (ii) We provide more details to the conference version. Specifically, we add a subsection to review the downstream applications of COD and present recent advances in poly segmentation (see Sec. II-C). We also provide the details of evaluation metrics to better understand their characteristics (see Sec. IV-C). (iii) We compare the proposed model with more state-of-the-art methods to validate the effectiveness of our model (see Sec. IV-D), and provide a sub-class comparison and a more insightful discussion (see Sec. IV-D.2). (iv) We extend the proposed C2F-Net to a downstream application of COD, i.e., polyp segmentation. Quantitative and qualitative evaluations conducted on four benchmark datasets demonstrate the superiority of our model over other existing polyp segmentation methods (see Sec. IV-F).

The rest of this paper is arranged as follows. In Section II, we review a number of works that are closely related to ours. In Section III, we describe our C2F-Net in detail. In Section IV, we provide extensive experimental results, ablation studies, and the further application of our C2F-Net in polyp segmentation. Finally, we conclude this work in Section V.

II. RELATED WORK

In this section, we discuss three types of works that are closely related to our method, i.e., camouflaged object detection, context-aware deep learning, and downstream applications of COD.

A. Camouflaged Object Detection

Recently, significant efforts have been put into COD, which is an emerging field in the computer vision community. Early COD approaches identify camouflaged objects with visual features, including texture, color, gradient, and motion [44]. In practice, a single visual feature cannot provide comprehensive characteristics for camouflaged objects.
Therefore, multiple features are integrated for improving performance [32]. Moreover, the Bayesian framework has been employed for detecting moving camouflaged objects from videos [64]. Despite their advantages, existing approaches relying on hand-crafted features tend to fail in real-world applications since they can only work with relatively simple scenarios. To this end, deep learning models, which automatically learn features and are trained in an end-to-end manner, have been adopted for accurate COD, which acts as an effective solution to address the challenges associated with traditional hand-crafted features. Yan et al. [60] proposed a two-stream network, called MirrorNet, for COD with the original and flipped images. The underlying motivation lies in the fact that the flipped image can provide valuable cues for COD. Lamdouar et al. [33] identified camouflaged objects from videos by exploiting the motion information with a deep learning framework, which consists of two modules, i.e., a differentiable registration module and a motion segmentation module. Ren et al. [47] proposed a two-step texture-aware refinement module to amplify the differences between the camouflaged object and its surroundings for accurate COD. Ji et al. [31] designed a multivariate calibration strategy to contrast the initial edge prior obtained from a selective edge aggregation strategy that mimics human visual perception. Li et al. [35] presented an adversarial learning network with a similarity measure module to model the contradictory information, enhancing the ability to detect salient and camouflaged objects. Mei et al. [43] developed a bio-inspired framework named Positioning and Focus Network, which contains a positioning module for mimicking the detection process of predator and a focus module to perform the identification process. Lyu et al. [40] proposed the first ranking-based network based on Mask-RCNN [24] to learn the camouflage degree with joint fixation and camouflage decoder. They also collect the large-scale testing dataset NC4K. SINetV2 [16] designed a neighbor connection decoder and group-reversal attention for detecting camouflaged objects. Interested readers can refer to [3] for a comprehensive review of COD. Furthermore, it is worth noting that COD shares a number of similarities with the popular salient object detection (SOD) tasks, such as RGB SOD [66], RGB-Depth/RGB-Thermal SOD [7], [25], [62], remote sensing SOD [10], [63], Co-SOD [9], [18], etc. Different from existing models, our C2F-Net advances the COD accuracy by jointly considering the rich global context information, effective cross-level fusion, and low-level feature refinement.

B. Context-Aware Deep Learning

Due to its superior capability of enhancing the feature representation, contextual information acts as a vital role in segmenting objects. For this reason, efforts have been made to enrich the contextual information. For instance, Zhao et al. [65] proposed PSPNet to establish multi-scale representations around each pixel for obtaining rich context information. Chen et al. [6] constructed ASPP with different dilated convolutions in order to enrich context information. The self-attention mechanism has also been proposed and employed for capturing rich context information. Typical instances include DANet [22], which extracts contextual information with non-local modules [56], and CCNet [27], which employs multiple cascaded cross-attention modules to obtain dense contextual information. Context information also gains significant attention from the field of object segmentation. For instance, Zhang et al. [61] enriched context information with a multi-scale context-aware feature extraction module, which provides rich context information. Liu et al. [38] proposed PoolNet, which is equipped with a pyramid pool module for rich context features highly related to the accurate salient object detection. Chen et al. [8] proposed a global context flow module to transfer features containing global semantic information to multi-level features at different stages to improve the integrity of SOD. Tan et al. [54] integrated the local and global context features in an adaptive coarse-to-fine manner for improved feature representations. Besides, the motion context and information have been effectively exploited and used in motion prediction [50] and activity recognition [49].

C. Downstream Applications of COD

COD has a large number of downstream applications in practice. Typical instances include medical image segmentation (e.g., polyp segmentation [29], [30], COVID-19 lung infection segmentation [20], etc.), surface defect detection, recreational art, concealed-salient object transition, transparent objects detection, search engines, etc [16].

Among various applications, polyp segmentation is one of the most attractive tasks due to its vital application value and close relationship with COD. Accurate polyp segmentation can provide vital information from colonoscopy images for identifying polyps, which have a high risk to develop into serious colorectal cancer that threatens human life. Early polyp segmentation methods rely on hand-crafted features [41], which are unable to fully capture the useful information for segmenting polyps. To address these challenges, deep learning has been employed for polyp segmentation and has shown great success. For instance, Brandao et al. [4] utilized the fully convolutional networks (FCNs) to segment polyps in an end-to-end manner. Akbari et al. [1] employed modified FCNs for accurate polyp segmentation. The variants of U-Net, i.e., U-Net++ [68] and ResUNet++ [28], were also utilized for polyp segmentation for improved performance. Boundary information was also employed for improving the accuracy of polyp segmentation [21], [45]. Ji et al. [29] proposed a normalized self-attention module to progressively segment the polyp region in the video clips. More recently, Fan et al. [17] proposed Pranet to segment polyps with paralleled partial decoder and reverse attention mechanism, significantly improving the performance. In this study, we further demonstrate the effectiveness of our C2F-Net by applying it to the polyp segmentation task.

III. PROPOSED METHOD

In this section, we first present the overall framework of the proposed C2F-Net. Then, we detail three key components in
The last DGCM provides an initial prediction \( f_D \). Two modules are organized in a cascaded manner.

The details of our \( \text{C}^2\text{F}-\text{Net} \). Finally, we detail the overall loss function for the proposed COD method.

\subsection{Overall Architecture}

Fig. 2 shows the overall framework of our \( \text{C}^2\text{F}-\text{Net} \), which can effectively integrate context-aware cross-level features to boost the accuracy of camouflaged object detection. Specifically, Res2Net-50 [23] is adopted as the backbone to extract multi-scale features, denoted as \( f_i \) with \( i = 1, 2, \ldots, 5 \). In our model, the extracted features are divided into two groups, i.e., low-level features \( Q_l = \{ f_1, f_2, f_3 \} \) and high-level features \( Q_h = \{ f_3, f_4, f_5 \} \). Then, three receptive field blocks (RFB) are utilized to enrich the features from \( Q_h \) by expanding the receptive fields. We utilize the same setting in [23], and the RFB includes five branches \( b_k \) with \( k = 1, 2, \ldots, 5 \). In each branch, a \( 1 \times 1 \) convolutional layer is first adopted to reduce the original channel size to 64. Following that, a \( (2k-1) \times (2k-1) \) convolutional layer and a \( 3 \times 3 \) convolutional layer with a specific dilation rate \( (2k-1) \) (when \( k > 2 \)) are adopted. The first four branches are concatenated to obtain the fused feature, which is fed into a \( 1 \times 1 \) convolutional layer for reducing the channel size to 64. Besides, the output of the fifth branch is added by the feature from the first four branches, which is further fed to a ReLU activation to obtain the final feature. Moreover, an attention-induced cross-level fusion module (ACFM) is proposed to fuse high-level features, and a dual-branch global context module (DGCM) is proposed to fully exploit multi-scale context information from the fused features. Two modules are organized in a cascaded manner. The last DGCM provides an initial prediction \( f_D \). We then refine the low-level features \( Q_l \) with \( f_D \) and predict the final COD result with our camouflage inference module (CIM).

\subsection{Attention-Induced Cross-Level Fusion Module}

Natural differences often exist among different types of camouflaged objects. What’s more, due to the observation distance and relative locations to the surrounding background, the size of similar camouflaged objects could change dramatically. To address the above issues, we design a novel ACFM by introducing a multi-scale channel attention (MSCA) [11] strategy to integrate cross-level features, which can effectively mine multi-scale information to alleviate scale variations.

The detailed structure of MSCA is shown in Fig. 2. In MSCA, it is a self-attention based on a two-branch structure. Specifically, the first branch adopts global average pooling (GAP) to exploit global contexts, which can emphasize globally distributed for large objects. The second branch holds the size of the original feature to obtain local contexts, which can prevent small camouflaged objects that are often ignored. Different from several existing multi-scale attention mechanisms, MSCA adopts point convolution in two branches to restore and compress features along the channel dimension, resulting in aggregating multi-scale context information. Cross-level features have different contributions to the COD task, thus fusing multi-level features can complement each other to obtain more comprehensive feature representations. To achieve this, we conduct ACFM on high-level features, and the cross-level fusion process is implemented by

\[
F_{ab} = \mathcal{M}(F_a \oplus F_b) \odot F_a \oplus (1 - \mathcal{M}(F_a \oplus F_b)) \odot F_b,
\]
where $\mathcal{M}$ denotes the MSCA operation, $F_a$ and $F_b$ denote the two input features. Besides, $\odot$ denotes the initial fusion of $F_a$ and $F_b$, in which $F_b$ is up-sampled and then added with $F_a$. $\ominus$ denotes element-wise multiplication. $(1 - \mathcal{M}(F_a \oplus F_b))$ corresponds to the dotted line in Fig. 2. Moreover, $F_{ab}$ is then fed into a $3 \times 3$ convolutional layer, followed by batch normalization and a $ReLU$ function. Thus, we can obtain the fused cross-level feature, i.e., $F$.

C. Dual-Branch Global Context Module

The proposed ACFM is adopted to fuse multi-scale features, in which a multi-scale channel attention strategy is introduced to obtain informative attention-based features. In addition, global context information is important for boosting the COD performance. Thus, a dual-branch global context module (DGCM) is proposed to fully mine global context information from the fused features. In detail, the output $F \in \mathbb{R}^{W \times H \times C}$ of the ACFM is first fed to two branches with a convolutional layer and average pooling, respectively, thus we obtain the sub-features $F_c \in \mathbb{R}^{W \times H \times C}$ and $F_p \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times C}$. After that, in order to learn multi-scale attention-based features, $F_c$ and $F_p$ are also fed to the MSCA module. Then, an element-wise multiplication is utilized to integrate the feature outputs of MSCA and the corresponding features ($F_c$ or $F_p$), and then we obtain $F_{cm} \in \mathbb{R}^{W \times H \times C}$ and $F_{pm} \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times C}$. Further, we adopt an addition operation to integrate the features from the above two-branch structure, and then the fused feature is denoted as $F_{pm}$. Finally, the residual connection is adopted to combine $F$ with $F_{pm}$ to obtain $F'$.

The detailed processing steps are presented as follows:

$$
\begin{align*}
F_c &= C(F), \\
F_{cm} &= F_c \odot \mathcal{M}(F_c), \\
F_p &= \mathcal{P}(F), \\
F_{pm} &= F_p \odot \mathcal{M}(F_p), \\
F' &= C(F \odot C(F_{cm} \oplus \mathcal{U}(F_{pm}))),
\end{align*}
$$

where $C$, $\mathcal{M}$, $\mathcal{P}$, and $\mathcal{U}$ denote the convolution, MSCA, average pooling, and upsampling operation, respectively.

Remarks: It is worth noting that our DGCM aims to enhance the fused features of ACFM, so that the proposed model adaptively extracts multi-scale information from a specific level. For clarity, the overall structure is presented in Fig. 2.

D. Camouflage Inference Module

To effectively make use of multi-scale features, the low-level features (i.e., $Q_l = \{f_1, f_2, f_3\}$) are refined by using the output (denoted as $f_D$) of the second DGCM. The refinement flow is presented to effectively fuse all the features from multiple layers. Followed the setting in the BBS-Net [19], we utilize a modified RFB module (denoted as $\mathcal{R}$) to fully expand the receptive fields for obtaining richer features and reducing the computation. As shown in Fig. 2, $f_D$ is adopted to refine three low-level features by using an element-wise multiplication operation, i.e., $f_1 \otimes f_D$, $f_2 \otimes f_D$, and $f_3 \otimes f_D$. Each refined feature is fed into a RFB module, and we can obtain $\mathcal{R}(f_1 \otimes f_D)$, $\mathcal{R}(f_2 \otimes f_D)$, and $\mathcal{R}(f_3 \otimes f_D)$. Then, $\mathcal{R}(f_1 \otimes f_D)$ and $\mathcal{R}(f_2 \otimes f_D)$ are cascaded and fed into a Conv block. After that, the output and $\mathcal{R}(f_3 \otimes f_D)$ are further cascaded and fed into a Conv block with a channel size of 16.

Further, we propose a Camouflage Inference Module (CIM) to generate the final detection results, which can take advantage of the low-level features via a multi-scale strategy. In the CIM, to fully exploit multi-scale information, we further propose a Multi-scale Residual Block (MRB) to detect local and multi-scale features. Specifically, we build a two-stream network and each stream utilizes a different convolutional kernel. As shown in Fig. 2, the current feature representation $X$ is fed to the first stream network with two sequential operations, in which the former sequential operation consists of a $3 \times 3$ convolution followed by batch normalization and a $ReLU$ function (denoted “$B_{conv3}$”), and the latter consists of a $3 \times 3$ transpose convolution followed by batch normalization and a $ReLU$ function (denoted “$B_{Deconv3}$”). Besides, $X$ is also fed to the second stream network two other sequential operations, in which the convolutional kernel is $5 \times 5$. The above processing steps can be presented as follows:

$$
\begin{align*}
F_{conv3} &= B_{Deconv3} X, \\
F_{Deconv5} &= B_{Deconv5} X, \\
F_{cm} &= C_{Deconv5}(3), \\
F_{pm} &= C_{Deconv5}(4), \\
F' &= C_{Deconv5}(5),
\end{align*}
$$

Then, the stream features are cascaded and then fed into a $3 \times 3$ sequential layer to obtain the fused multi-scale feature representation, i.e., $F_f = B_{conv3} \odot 3(C_{at}(F_{conv3}, F_{conv5}))$. To preserve the original information from the input $X$, we adopt residual learning in the MRB. Therefore, we obtain the output of each MRB by adding $X$ and the fused multi-scale feature. The whole CIM consists of three $1 \times 1$ convolution and two MRB, which is used to generate the final prediction $P$.

E. Loss Function

The binary-cross entropy loss ($L_{BCE}$) is widely used to calculate the loss of each pixel to form a pixel restriction on the network. To effectively exploit the global structure, the work [46] introduced IoU loss ($L_{IoU}$) to form a global restriction on the network. However, the above losses are often treated equally to all pixels, ignoring the differences among pixels. To overcome this issue, this work [57] improves $L_{BCE}$ and $L_{IoU}$ into the weighted binary cross-entropy loss ($L_{p_{BCE}}$) and IoU loss ($L_{p_{IoU}}$). Specifically, by computing the difference between the center pixel and its surroundings, each pixel can be assigned a different weight, so that the hard pixels are paid more attention. In this case, the basic loss function of the proposed model is defined as:

$$
L = L_{p_{IoU}} + L_{p_{BCE}}
$$

where $L(f_D, G)$ denotes the loss between the initial prediction map and ground truth, and $L(P, G)$ denotes the loss between the final prediction map and ground truth.

IV. EXPERIMENTS

In this section, we first present the implementation details, datasets, and evaluation metrics used in our experiments. Then, we show quantitative and qualitative results by comparing...
our model with other existing methods. Further, we conduct ablation studies to validate the effectiveness of each key component. Finally, we extend the application of our model to polyp segmentation.

A. Implementation Details

The proposed model is implemented using PyTorch, and the Res2Net-50 [23] pretrained on ImageNet is adopted as the backbone. We adopt the AdaX [36] algorithm to optimize the overall framework. To enhance the generalizability of the proposed model, we adopt a multi-scale training strategy [0.75, 1, 1.25] in the training stage. Besides, the input images are resized to 352 × 352. The initial learning rate is set to 1e−4, which will decay 10 times after 30 epochs. The network is trained on an NVIDIA Tesla V100 GPU. The batch size is set to be 30, and the training process takes about 6 hours over 100 epochs.

B. Datasets

To validate the effectiveness of the proposed model, we carry out comparison experiments on three benchmark datasets. The details of each COD dataset are provided below.

- CHAMELEON [15] is collected using the Google search engine with the keyword “camouflage animals,” which consists of 76 camouflaged images using as a testing dataset.
- CAMO [34] consists of 1,250 camouflaged images (1,000 for training, 250 for testing), which includes eight categories.
- COD10K [15] is currently the largest COD dataset with pixel-level annotations. It consists of 5,066 images, where 3,040 for training and 2,026 for testing. Besides, this dataset is split into 5 super-classes and 69 sub-classes.

Training and testing settings. Following the setting in [15], for the CAMO dataset, we utilize the default training set. For the COD10K dataset, we utilize the default training camouflaged images. In the test stage, we test our C2F-Net and other compared methods on the test sets of CAMO and COD10K, and the whole CHAMELEON dataset.

C. Evaluation Metrics

We adopt five widely used metrics, namely MAE (\(M\)), S-measure (\(S_a\)), F-measure (\(F_\beta\)), weighted F-measure (\(F_\beta^w\)), and E-measure (\(E_\phi\)), to evaluate the effectiveness of our model and all compared methods. The detailed definitions for these metrics are as follows:

1) Mean absolute error (MAE, \(M\)) computes the average difference between the normalized prediction and ground truth, which is defined by

\[
M = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x, y) - G(x, y)|,
\]

where \(P(x, y)\) and \(G(x, y)\) correspond to the prediction map and ground truth value at the pixel location \((x, y)\).

2) S-measure \((S_a)\) [12] computes the object-aware similarity \(S_o\) and region-aware structure similarity \(S_r\) between the prediction map and the ground truth by

\[
S_a = \alpha * S_o + (1 - \alpha) * S_r,
\]

where \(\alpha\) is a trade-off parameter to balance \(S_o\) and \(S_r\), and it is set to 0.5 as in [12].

3) F-measure \((F_\beta)\) is employed to evaluate binary map results between the prediction and the ground truth. It is defined as:

\[
F_\beta = \frac{(1 + \beta^2) Precision \cdot Recall}{\beta^2 Precision + Recall},
\]

where \(\beta\) is the weight between the precision and recall. \(\beta^2 = 0.3\) is the default setting since the precision is often weighted more than the recall. Besides, the improved version of \(F_\beta\), i.e., weighted F-measure \((F_\beta^w)\), is proposed to overcome the interpolation, dependency, and equal-importance flaws of \(F_\beta\). This metric synthetically considers both weighted precision and weighted recall [42].

4) E-measure \((E_\phi)\) [13], [14] evaluates the overall and local accuracy of a binary foreground map, which can be defined by

\[
E_\phi = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \phi(x, y).
\]

In summary, for these five metrics above, higher \(S_a, F_\beta^w, F_\beta, E_\phi\), and lower \(M\) indicate better performance. Note that we adopt adaptive F/E-measure in our evaluation.

D. Comparison With State-of-the-Art COD Methods

We compare the proposed method with 19 state-of-the-art (SOTA) COD methods, including FPN [37], MaskRCNN [24], PSPNet [65], UNet++ [68], PiCANet [39], MSRCNN [26], BASNet [46], PFANet [67], CDP [59], HTC [5], E2GNet [66], ANet-SRM [34], SINet [15], MirrorNet [60], UCNet [58], PranNet [17], LSR [40], PFNet [43], and ERRNet [31]. For a fair comparison, the results of MirrorNet are collected from link,\(^1\) the results of ERRNet are taken from link,\(^2\) and the results of remaining 18 methods are collected from link.\(^3\)

1) Quantitative Evaluation: Table I shows the quantitative results of the proposed model, all compared COD methods on the three benchmark datasets. Meanwhile, the backbones of all the methods are provided in Table I.

a) Performance on CAMO: From Table I, compared with 19 SOTA models, it can be observed that our C2F-Net performs better than other models in terms of all metrics. Specifically, compared with LSR, the second-best model, \(S_a, F_\beta^w, F_\beta, E_\phi\) increased by 1.6%, 4.9%, and 1.6%,

\(^1\)https://sites.google.com/view/ltmhla/research/camo
\(^2\)https://github.com/GeelsaJJ/ERRNet
\(^3\)http://dpfan.net/socbenchmark
respectively. This is mainly because that the proposed ACFM can mine multi-scale information to alleviate scale variations, and DCCM aims to enhance the fused features of ACFM for providing global context information. More importantly, the proposed CIM can make use of the low-level features and further exploit multi-scale information to boost the COD performance.

b) Performance on CHAMELEON: Our C2F-Net is also tested on CHAMELEON [15] dataset. As reported in Table I, our C2F-Net obtains better performances across all the metrics. Specifically, compared with the LSR, $E_\phi$ and $F_\beta$ increase by 1.2% and 2.7%, respectively. Overall, the results demonstrate the good generalization ability of our model.

c) Performance on COD10K: The COD10K dataset is a more challenging sequence, which contains 2,026 testing images distributed in 5 super-classes. It can be observed that our C2F-Net achieves the new SOTA performances across all five metrics. The main reason for this robustness results is that the mature designed modules can combine multi-level features and exploit rich global context information.

2) Per-Subclass Performance: In addition to the overall quantitative comparison on the COD10K dataset,
super-class results are also presented in Table II. In this experiment, four main super-classes, including “Aquatic,” “Terrestrial,” “Flying,” and “Amphibian,” are taken into consideration. From the results in Table II, it can be seen that C²F-Net outperforms other SOTA methods on the super-class “Aquatic” and “Flying” in terms of all four metrics. On the super-class “Terrestrial,” as shown in Table II, our model achieves the best performance in terms of $F_β$, $E_ϕ$, and $S_α$ metrics. Except for $E_ϕ$ and $MAE$, the proposed method outperforms other SOTA models on the super-class “Amphibian.”

3) Qualitative Evaluation: Fig. 3 shows the qualitative comparison results. These examples are collected from five super-classes in the COD10K dataset, including “Aquatic,” “Terrestrial,” “Flying,” and “Amphibian,” and “Others” (from $1^{st}$ row to $5^{th}$ row). The challenge factors are also presented, i.e., multiple objects (see the $6^{th}$ and $7^{th}$ rows) and occlusion (see the $8^{th}$ and $9^{th}$ rows). As shown in Fig. 3, C²F-Net improves the visual results compared to other COD methods in the detailed branches of the ghost pipefish and mantis (see the $1^{st}$ and $3^{rd}$ rows). In the context under
different lightning, compared with PFNet [43], LSR [40], and SINet [15], our model can accurately identify camouflaged objects and their boundary (see the 5th row). For multiple objects, compared with PFNet [43], LSR [40], C2F-Net effectively detect multiple concealed objects (see the 6th and 7th rows). To detect occluded objects, compared with other methods, our model can recognize the certain boundaries of the camouflaged objects. Overall, compared with other SOTA models, our C2F-Net can achieve better performance by detecting more accurate and complete camouflaged objects with rich details.

4) Model Size and Inference Time: The results, shown in Tables IV, are the model sizes and inference times of our model and representative models. # Param is measured in million (M) and memory is measured in giga (G). As can be observed, our model is with the minimal parameters in comparison with the representative models. Its memory cost is also controlled well, i.e., much less than ERRNet and PFNet. However, we realize that the inference speed of our C2F-Net is not promising enough, which is mainly caused by the new improvement applied to our model. Specifically, the IJCAI version of our model can achieve an Fps of 39.07, which is reduced to 27.41 after using the CIM. We will seek to improve the CIM with more efficient design in future work. Note that we perform experiments to test the inference time on a single Tesla P100 GPU.

5) Failure Cases: We present a number of typical failure cases in Fig. 4. As can be observed, our model may fail in some extremely challenging cases. However, it is worth noting that, in these cases, existing SOTA models also fail and our model still outperforms the SOTA models.

E. Ablation Study

To evaluate the effectiveness of each component in the proposed model, we designed seven ablation experiments (as presented in Table III). In the No.1 (Basic) experiment, we removed all ACFMs, DGCMs, and CIM. Specifically, the IJCAI version of our model can achieve an Fps of 39.07, which is reduced to 27.41 after using the CIM. We will seek to improve the CIM with more efficient design in future work. Note that we perform experiments to test the inference time on a single Tesla P100 GPU.
TABLE V
QUANTITATIVE EVALUATION FOR POLYP SEGMENTATION ON THE FOUR DATASETS. THE BEST RESULTS ARE HIGHLIGHTED IN BOLD FONTS

| Method | CVC-300 | CVC-ClinicDB | CVC-ColonDB | ETIS-LaribPolypDB |
|--------|---------|--------------|-------------|------------------|
|        | $S_a \uparrow$ | $E_\phi \uparrow$ | $F_p^{\uparrow}$ | $M \downarrow$ | $S_a \uparrow$ | $E_\phi \uparrow$ | $F_p^{\uparrow}$ | $M \downarrow$ | $S_a \uparrow$ | $E_\phi \uparrow$ | $F_p^{\uparrow}$ | $M \downarrow$ |
| U-Net [48] | 0.843 | 0.867 | 0.684 | 0.703 | 0.022 | 0.889 | 0.917 | 0.811 | 0.804 | 0.019 | 0.712 | 0.763 | 0.498 | 0.569 | 0.061 | 0.684 | 0.645 | 0.366 | 0.398 | 0.036 |
| U-Net++ [68] | 0.839 | 0.884 | 0.687 | 0.706 | 0.018 | 0.873 | 0.898 | 0.785 | 0.784 | 0.022 | 0.691 | 0.762 | 0.467 | 0.560 | 0.064 | 0.683 | 0.704 | 0.390 | 0.465 | 0.035 |
| SFA [21] | 0.640 | 0.604 | 0.341 | 0.353 | 0.065 | 0.793 | 0.816 | 0.647 | 0.655 | 0.042 | 0.634 | 0.648 | 0.379 | 0.407 | 0.094 | 0.557 | 0.515 | 0.231 | 0.255 | 0.109 |
| PraNet [17] | 0.925 | 0.938 | 0.843 | 0.824 | 0.010 | 0.936 | 0.957 | 0.896 | 0.885 | 0.009 | 0.820 | 0.847 | 0.699 | 0.718 | 0.043 | 0.794 | 0.792 | 0.600 | 0.602 | 0.031 |
| C2F-Net | 0.928 | 0.970 | 0.864 | 0.868 | 0.008 | 0.942 | 0.978 | 0.923 | 0.922 | 0.009 | 0.821 | 0.883 | 0.713 | 0.735 | 0.044 | 0.798 | 0.853 | 0.629 | 0.649 | 0.032 |

Fig. 6. Visual comparisons of different methods on the polyp segmentation task.

then predicted the COD result with our CIM. In the No.5 (Basic+ACFM+DGCM) experiment, ACFMs and DGCNs are kept, while the CIM is taken out. In the No.6 (MSCA → Conv) experiment, we replaced MSCA with a convolutional block. Finally, No.7 (Basic+ACFM+DGCM+CIM) is the full version of our CIM.

1) Effectiveness of ACFM: We first study the effectiveness of ACFM. As shown in Table III, it can be observed that No.2 (Basic+ACFM) outperforms No.1 (Basic), clearly indicating that the ACFM is helpful and critical to boost the COD performance, increasing the $F_w^{\beta}$ by 2.15% on CAMO. Compared to the basic network, the ACFM could improve the ability to detect the main part of the camouflaged object (4th col).

2) Effectiveness of DGCM: From the results in Table III, it can be seen that No.3 (Basic + DGCM) outperforms better than No.1 (Basic) on all used datasets. Specifically, the adaptive $E_\phi$ increased 1.81% on CHAMELEON. The results indicate that introducing the DGCM can enable our model to detect objects accurately.

3) Effectiveness of CIM: As shown in Table III, No.4 (Basic + CIM) outperforms No.1 (Basic) on three benchmarks datasets. This owes to CIM’s ability of exploiting multi-scale information from low-level features, which plays a crucial role in boosting the COD performance.

4) Effectiveness of ACFM & DGCM: To evaluate the combination of ACFM and DGCM, we carry out the ablation study of No.5. As shown in Table III, the performance of the combination is generally better than the first four settings.

5) Effectiveness of MSCA: In this ablation study, we replace it with a convolutional layer (denoted “MSCA → Conv”). The comparison results of the No.6 are shown in Table III, indicating that the use of MSCA significantly improves the results, especially on the CAMO-Test dataset.

6) Effectiveness of CIM & ACFM & DGCM: To evaluate the full version of our model, we test the performance of No.7. From the results in Table III, it can be seen that the full version of C2F-Net is better than all other ablated ones. Additionally, the visual comparisons in Fig. 5 further show that our full model is more conducive to identifying and locating camouflaged objects.

It is worth noting that No.5 (Basic+ACFM+DGCM) is the conference version of our model, which was published in IJCAI 2021 [52]. The results, shown in Table I, indicate that the full version of our C2F-Net (i.e., No.7) outperforms No. 5 on all three datasets. This sufficiently demonstrates the
effectiveness of our new improvement applied to the model. The underlying reason for effectiveness lies in the ability of CIM in exploiting useful information from low-level features with the guidance of coarse prediction map. In addition, the MRBs in CIM promote making full use of the features for improved performance. Due to these novel designs, CIM effectively improves the performance and provides results much better than the previous version [52].

F. Application to Polyp Segmentation

As aforementioned in Sec. II-C, COD has rich downstream applications in practice. Therefore, we further evaluate our C2F-Net by applying it to a typical COD downstream application, polyp segmentation.

Experiments are conducted on four public datasets, including ETIS [51], CVC-ClinicCB [2], CVC-ColonDB [53], and CVC-300 [55]. We compare C2F-Net with four cutting-edge polyp segmentation models, i.e., UNet [48], UNet++ [68], SFA [21], and PraNet [17]. The results of these four methods are taken from https://github.com/DengPingFan/PraNet. Similar to the COD experiments, we evaluate the results using five metrics, including MAE, \( S_\alpha \), \( F^w_\beta \), \( F_\beta \), and \( E_\phi \).

1) Qualitative Evaluation: Table V summarizes the quantitative results of different methods on four polyp datasets.

a) Performance on CVC-300: As shown in Table V, C2F-Net outperforms the other four cutting-edge methods on CVC-300. Compared with the second-best method, ParNet [17], the performance of \( E_\phi \) improves by 3.30%.

b) Performance on CVC-ClinicDB: We further test our model on CVC-ClinicDB, which includes 612 open-access images from 31 colonoscopy clips. The performance of \( F_\beta \) increases by 4.01%.

c) Performance on CVC-ColonDB: CVC-ColonDB is a small dataset, which contains 380 images from 15 short colonoscopy copy sequences. Except for the MAE, C2F-Net obtains better performance. Especially, compared to ParNet [17], the performance of \( E_\phi \) increases by 4.01%.

d) Performance on ETIS: ETIS consists of 196 polyp images for early diagnosis of colorectal cancer. Compared to ParNet [17], the results of \( E_\phi \) and \( F_\beta \) are increased by 7.15% and 7.24%, respectively.

2) Qualitative Evaluation: Fig. 6 shows the visual results of different models on four polyp datasets, including small polyps (1st row), different lightning (2nd row), and low-contrast scenes (3rd row). For the small polyps, C2F-Net can accurately detect boundaries, while UNet [48] and UNet++ [68] fail. Furthermore, our model outperforms competing methods under different lightning (2nd row), demonstrating its ability to identify polyps under poor visual conditions. In the condition of low-contrast scenes, the results of C2F-Net make the minimum error areas (3rd row). We can see that C2F-Net performs better than PraNet [17], the SOTA poly segmentation model that is designed to handle challenging conditions, such as varied size, complicated context, and so on.

V. Conclusion

In this paper, we have proposed a novel C2F-Net for the COD task. Our C2F-Net can effectively integrate the cross-level features that contain rich global context information. In C2F-Net, DGCMs are proposed to exploit global context information from the fused features. Furthermore, the high-level features are fused with ACFMs, which integrate the features under the guidance of valuable attention cues provided by MSCAs. Finally, we propose CIM to predict the final result from the low-level features enhanced by the coarse prediction map. Experiments on three public datasets show that our model outperforms other state-of-the-art COD methods. Moreover, the ablation studies validate the effectiveness of each key component. Further evaluation on the four polyp segmentation datasets demonstrates the promising potentials of our C2F-Net in COD downstream applications.

REFERENCES
A. V. Mamonov, I. N. Figueiredo, P. N. Figueiredo, and Y.-H. R. Tsai, “Automated polyp detection in colon capsule endoscopy,” IEEE Trans. Med. Imag., vol. 33, no. 7, pp. 1488–1502, Jul. 2014.

J. Ren et al., “Deep texture-aware features for camouflaged object detection,” IEEE Trans. Circuits Syst. Video Technol., early access, Nov. 13, 2021, doi: 10.1109/TCSVT.2021.3126591.

R. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., Jun. 2014, pp. 248–255.

H. Mei, G.-P. Ji, Z. Wei, X. Yang, X. Wei, and D.-P. Fan, “Camouflaged object segmentation with distraction mining,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2021, pp. 8772–8781.

M. Mondal, “Camouflaged object detection and tracking: A survey,” Int. J. Image Graph., vol. 20, no. 4, Oct. 2020, Art. no. 2050028.

B. Sivaprasakasam, K. Ram, J. Joseph, and M. Sivaprasakasam, “PoiNet: Shape and boundary aware joint multi-task deep network for medical image segmentation,” in Proc. 41st Annu. Int. Conf. IEEE Eng. Med. Biol. Soc. (EMBC), Jul. 2019, pp. 7223–7226.

Q. Zhang, C. Zhu, S. Wang, and M. Ye, “A Bayesian approach to colo-rectal cancer detection,” IEEE Trans. Med. Imag., vol. 30, pp. 1305–1317, 2021.

J. Tan, P. Xiong, Z. Lv, K. Xiao, and Y. He, “Local context attention for salient object segmentation,” in Proc. Asian Conf. Comput. Vis., 2020, pp. 706–722.

D. Vázquez et al., “A benchmark for endoluminal scene segmentation of colonoscopy images,” J. Healthcare Eng., vol. 2017, pp. 1–9, Jul. 2017.

X. Wang, R. Girshick, A. Gupta, and K. He, “Non-local neural networks,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 7974–7983.

J. Wei, S. Wang, and Q. Huang, “P3Net: Fusion, feedback and focus for salient object detection,” in Proc. AAAI Conf. Artif. Intell., vol. 34, no. 7, 2020, pp. 12321–12328.

Y.-H. Wu, Y. Liu, Z. Zhang, W. Gao, and M.-M. Cheng, “Regularized density-connected pyramid network for salient instance segmentation,” IEEE Trans. Image Process., vol. 30, pp. 3897–3907, 2021.

Z. Wu, L. Su, and Q. Huang, “Casceded partial decoder for fast and accurate salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2019, pp. 3907–3916.

J. Yan, T. Le, K. Nguyen, M. Tran, T. Do, and T. V. Nguyen, “MirrorNet: Bio-inspired camouflaged object segmentation,” IEEE Access, vol. 9, pp. 43290–43300, 2021.

L. Zhang, J. Dai, H. Li, Y. He, and G. Wang, “A bi-directional message passing model for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., Jun. 2018, pp. 1741–1750.

Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han, “RGB-T salient object detection via fusing multi-level CNN features,” IEEE Trans. Image Process., vol. 29, pp. 3321–3335, 2020.
[65] H. Zhao, J. Shi, X. Qi, X. Wang, and J. Jia, “Pyramid scene parsing network,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2881–2890.

[66] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, “EGNet: Edge guidance network for salient object detection,” in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*., Oct. 2019, pp. 8779–8788.

[67] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[68] Z. Zhou et al., “UNet++: Redesigning skip connections to exploit multiscale features in image segmentation,” *IEEE Trans. Med. Imag.*, vol. 39, no. 6, pp. 1856–1867, Dec. 2019.

[69] M. Zhuge, X. Lu, Y. Guo, Z. Cai, and S. Chen, “CubeNet: X-shape connection for camouflaged object detection,” *Pattern Recognit.*, vol. 127, Jul. 2022, Art. no. 108644.

Geng Chen received the Ph.D. degree from Northwestern Polytechnical University, China, in 2016. He was a Research Scientist with the Inception Institute of Artificial Intelligence, United Arab Emirates, from 2019 to 2021, and a Post-Doctoral Research Associate with The University of North Carolina at Chapel Hill, USA, from 2016 to 2019. He is a Professor with Northwestern Polytechnical University. His research interests lie in medical image analysis and computer vision.

Si-Jie Liu is currently pursuing the Ph.D. degree with the School of Power and Energy, Northwestern Polytechnical University, China, under the supervision of Prof. Ya-Feng Wu. His research interests mainly focus on computer measurement and control, digital twin for industrial equipment maintenance, and computer vision.

Yu-Jia Sun is currently pursuing the master's degree with the School of Computer Science, Inner Mongolia University, China, under the supervision of Prof. Hong-Xi Wei. Her research interests mainly focus on computer vision and camouflaged object detection.

Ge-Peng Ji received the master’s degree in communication and information systems from the School of Computer Science, Wuhan University, in 2021. His research interests lie in computer vision, especially in a variety of dense prediction tasks, such as video segmentation, medical image segmentation, camouflaged object segmentation, and saliency detection. He received the MICCAI Student Travel Award in 2021.

Ya-Feng Wu received the Ph.D. degree from Northwestern Polytechnical University, China, in 2002. He is a Professor with the School of Power and Energy, Northwestern Polytechnical University. His main research interests include modern signal processing, active noise control, and computer measurement and control.

Tao Zhou (Member, IEEE) received the Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. From 2016 to 2018, he was a Post-Doctoral Fellow with the BRIC and IDEA Laboratory, The University of North Carolina at Chapel Hill. From 2018 to 2020, he was a Research Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. He is currently a Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning, computer vision, and medical image analysis.