Context-aware Cross-level Fusion Network for Camouflaged Object Detection

Yujia Sun[1*] , **Geng Chen**[2*] , **Tao Zhou**[3†] , **Yi Zhang**[4] , **Nian Liu**[5]

[1] School of Computer Science, Inner Mongolia University, China
[2] National Engineering Laboratory for Integrated Aero-Space-Ground-Ocean Big Data Application
Technology, School of Computer Science and Engineering, Northwestern Polytechnical University, China
[3]PCA Lab, the Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information
of Ministry of Education, School of Computer Science and Engineering, Nanjing University of Science
and Technology, Nanjing, China
[4]Institut National des Sciences Appliquées de Rennes, Rennes, France
[5]Inception Institute of Artificial Intelligence, Abu Dhabi, UAE
taozhou.ai@gmail.com

Abstract

Camouflaged object detection (COD) is a challenging task due to the low boundary contrast between the object and its surroundings. In addition, the appearance of camouflaged objects varies significantly, *e.g.*, object size and shape, aggravating the difficulties of accurate COD. In this paper, we propose a novel Context-aware Cross-level Fusion Network (C^2F-*Net*) to address the challenging COD task. Specifically, we propose an Attention-induced Cross-level Fusion Module (ACFM) to integrate the multi-level features with informative attention coefficients. The fused features are then fed to the proposed Dual-branch Global Context Module (DGCM), which yields multi-scale feature representations for exploiting rich global context information. In C^2F-*Net*, the two modules are conducted on high-level features using a cascaded manner. Extensive experiments on three widely used benchmark datasets demonstrate that our C^2F-*Net* is an effective COD model and outperforms state-of-the-art models remarkably. Our code is publicly available at: https://github.com/thograce/C2FNet.

1 Introduction

Due to the purpose of survive, wild animals have developed rich camouflage abilities. In practice, they try to change their appearance to "perfectly" integrate into the surrounding environment, in order to avoid the attention from other creatures. In recent years, camouflage has attracted increasing research interest from the computer vision community [Le *et al.*, 2019; Fan *et al.*, 2021a; Lyu *et al.*, 2021; Mei *et al.*, 2021; Zhai *et al.*, 2021]. Among various topics, camouflaged object detection (COD), which aims to identify and segment the camouflaged objects from images, is particularly popular. However, COD is a challenging task due to the nature of camouflage.

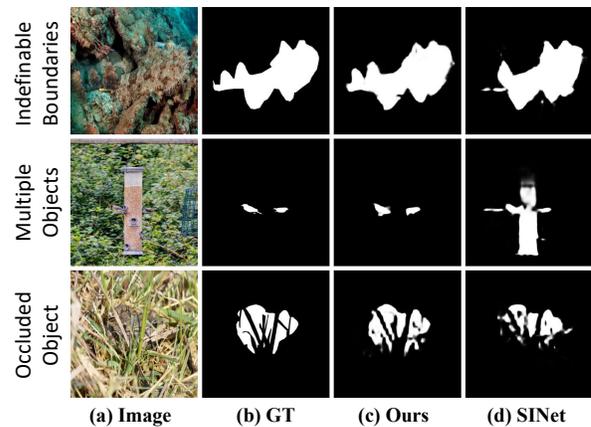*Co-first Authors. †Corresponding Author.

Figure 1: From top to bottom, three challenging camouflage scenarios with indefinable boundaries, multiple objects and occluded object are listed. Our model outperforms SINet [Fan *et al.*, 2021a], in these challenging scenarios.

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 Figure 1, it is very challenging to discover the fish from the underwater image. In addition, the camouflaged objects, most wild animals, are usually with varied appearance, *e.g.*, size and shape, which further aggravates the difficulties of accurate COD [Le *et al.*, 2019; Fan *et al.*, 2021a].

To address these challenges, deep learning techniques have been adopted and shown great potentials [Le *et al.*, 2019; Fan *et al.*, 2021a]. In addition, a number of COD datasets were constructed for training the deep learning models. For instance, [Le *et al.*, 2019] created the first COD dataset, called CAMO, consisting of 2500 images. However, due to the limited sample size, CAMO is unable to take full advantage of deep learning models. Recently, [Fan *et al.*, 2021a] constructed COD10K, the first large-scale COD dataset containing 10,000 images with the consideration of various challenging camouflage attributes in the real natural environments.

Apart from datasets, these two works also contribute to the COD from the aspect of deep learning models. [Le *et al.*, 2019] proposed an anabranch network, which performs classification to predict whether an image contains camouflaged objects, and then integrates this information into the camouflaged object segmentation task. Based on the observation that predators discover preys by first searching and then recognizing, [Fan *et al.*, 2021a] designed a searching module and an identification module to locate rough areas and to accurately segment camouflaged objects, respectively.

Existing methods have shown promising performance in the detection of a single camouflaged object from relatively simple scenes. However, their performance degrades significantly for a number of challenging cases, such as multiple objects, occlusion, etc. As shown in Figure 1, stat-of-the-art models fail to correctly identify all camouflaged objects and show unsatisfactory performance when occlusion exits. A key solution to address these challenges is a significant large receptive field, which provides rich context information for accurate COD. In addition, the cross-level feature fusion also plays a vital role in the success of COD. However, existing works usually overlook the importance of these two key factors. It is, therefore, greatly desired for a unified COD framework that jointly considers both rich context information and effective cross-level feature fusion.

To this end, we propose a novel COD model, called Context-aware Cross-level Fusion Network (*C^{2}F-Net*). We first fuse the cross-level features extracted from the backbone using an Attention-induced Cross-level Fusion Module (ACFM), which integrates the features with the guidance from a Multi-Scale Channel Attention (MSCA) component. The ACFM computes the attention coefficients from multi-level features, employs the resulting attention coefficients to refine the features, and then integrates the resulting features for the fused one. Subsequently, we propose a Dual-branch Global Context Module (DGCM) to exploit the rich context information from the fused features. The DGCM transforms the input feature into multi-scaled ones with two parallel branches, employs the MSCA components to compute the attention coefficients, and integrates the features by considering the attention coefficients. The ACFM and DGCM are organized in a cascaded manner at two stages, from high-level to low-level. The final DGCM predicts the segmentation map of the camouflaged object(s) from an image. In summary, the main contributions of this paper are as follows:

- We propose a novel COD model, *C^{2}F-Net*, which integrates the cross-level features with the consideration of rich global context information.
- We propose a context-aware module, DGCM, which exploits global context information from the fused features. Our DGCM is capable of capturing valuable context information, which is the key factor for improving COD accuracy.
- We integrate the cross-level features with an effective fusion module, ACFM, which integrates the features with the valuable attention cues provided by MSCA.
- Extensive experiments on three benchmark datasets demonstrate that our *C^{2}F-Net* outperforms 14 state-of-the-art models in the terms of four evaluation metrics.

2 Related Work

In this section, we discuss the two types of works that are closely related to our model, *i.e.*, camouflaged object detection and context-aware deep learning.

2.1 Camouflaged Object Detection

As an emerging field, COD has gained increasing attention in recent years. Early works rely on visual features, such as color, texture, motion, and gradient [Mondal, 2020]. In practice, the techniques relying on a single feature are unable to provide promising performance, therefore [Jiang, 2011] integrated different features for improved performance. However, limited by hand-crafted features, these methods are only suitable for relatively simple scenarios and tend to fail in real-world applications. To address this, deep learning techniques have been employed for COD by taking advantage of deep features automatically learned by the network from extensive training images, which are more generic and effective than hand-crafted features. For instance, [Li *et al.*, 2017] identified the camouflaged object with a region proposal network. [Yan *et al.*, 2020] observed that flipping the image provides vital cues for identifying the camouflaged object, therefore a two-stream MirrorNet was proposed with the original and flipped images as inputs for COD. [Lamdouar *et al.*, 2020] exploited the motion information to identify camouflaged objects in videos with a deep learning framework, consisting of two components, *i.e.*, a differentiable registration module to highlight object boundaries and a motion segmentation module with memory that discovers moving regions. This method overcomes the limitations of traditional motion information based methods that tend to fail when the target is stationary. For COD application, polyp segmentation has also received widespread attention. [Fan *et al.*, 2020] proposed a parallel reverse attention network (PraNet), which first predicts rough regions and then refines the boundaries.

2.2 Context-aware Deep Learning

The contextual information plays a crucial role in object segmentation tasks due to its great capability of enhancing the feature representation for improved performance. Efforts have been dedicated to enriching the contextual information. For instance, PSPNet [Zhao *et al.*, 2017] obtains rich context information by establishing multi-scale representations around each pixel. [Chen *et al.*, 2018] employed different dilated convolutions to construct ASPP to achieve context information acquisition. DANet [Fu *et al.*, 2019] utilizes non-local modules [Wang *et al.*, 2018] to extract contextual information. CCNet [Huang *et al.*, 2019b] obtains dense contextual information by cyclically cross-attention modules. In the field of salient object detection, [Zhang *et al.*, 2018] utilized a multi-scale context-aware feature extraction module to allow multi-level feature maps to obtain rich context information. PoolNet [Liu *et al.*, 2019], which is equipped with a pyramid pool module, employs a global guidance module to obtain the location information of underlying salient targets for different feature layers. GCFANet [Chen *et al.*, 2020] utilizes a global context flow module to transfer features containing global semantics to feature mapping at different stages to improve the
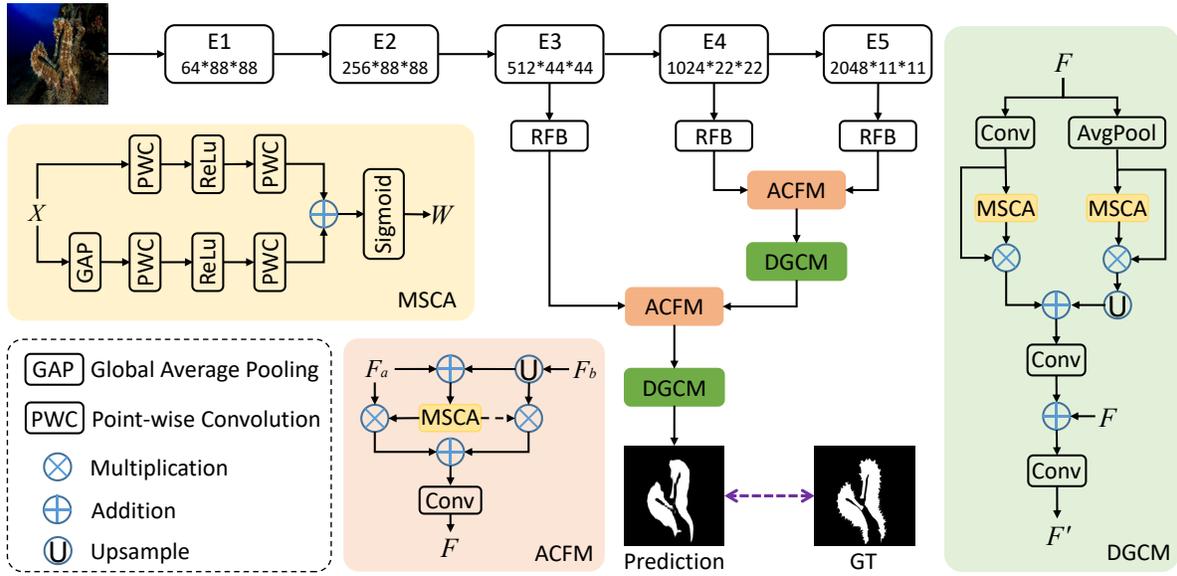
Figure 2: The overall architecture of the proposed model, which consists of two key components, *i.e.*, attention-induced cross-level fusion module and dual-branch global context module. See § 3 for details.

integrity of the detection of prominent objects. LCANet [Tan *et al.*, 2020] adaptively integrates the local area context and the global scene context in a coarse-to-fine structure to enhance local context features.

3 Proposed Method

In this section, we first outline the overall architecture of the proposed C^2F-*Net*. Then, we provide the details of the two proposed modules. Finally, we give the overall loss function for the proposed model.

3.1 Overall Architecture

Figure 2 shows the overall architecture of the proposed C^2F-*Net*, which fuses context-aware cross-level features to improve the camouflaged object detection performance. Specifically, we adopt Res2Net-50 [Gao *et al.*, 2019] to extract its features at five different layers, denoted as f_i ($i = 1, 2, \ldots, 5$). Then, a receptive field block (RFB) is adopted to expand the receptive field for capturing richer features in a specific layer. We utilize the same setting in [Gao *et al.*, 2019], and the RFB component includes five branches b_k ($k = 1, 2, \ldots, 5$). In each branch, the first convolutional layer has dimensions 1×1 to reduce the channel size to 64. This is followed by two layers, *i.e.*, a $(2k-1) \times (2k-1)$ convolutional layer and a 3×3 convolutional layer with a specific dilation rate $(2k-1)$ when $k > 2$. The first four branches are concatenated and then their channel size is reduced to 64 using a 1×1 convolutional operation. Then, the fifth branch is added and the whole module is fed to a *ReLU* activation function to obtain the final feature. After that, we propose an Attention-induced Cross-level Fusion Module (ACFM) to integrate multi-scale features, and a Dual-branch Global Context Module (DGCM) to mine multi-scale context information within the fused features. Finally, we obtain the pre-

diction result for camouflaged object detection. We give the details of each key component below.

3.2 Attention-induced Cross-level Fusion Module

There are natural differences among different types of camouflaged objects. Besides, the size of similar camouflaged objects may also vary greatly, due to the observation distance and their relative location to the surrounding environment. In other words, for a captured image with camouflaged object(s), the proportion of the camouflaged object(s) often varies. To address these challenges, we propose an ACFM via introducing a Multi-Scale Channel Attention (MSCA) [Dai *et al.*, 2020] mechanism to effectively fuse cross-level features, which can exploit multi-scale information to alleviate scale changes. The MSCA has strong adaptability to different scale targets, and its structure is shown in Figure 2. MSCA is based on a double-branch structure, in which one branch uses global average pooling to obtain global contexts to emphasize globally distributed large objects, and the other branch maintains the original feature size to obtain local contexts to avoid small objects being ignored. Different from other multi-scale attention mechanisms, MSCA uses point convolution in two branches to compress and restore features along the channel dimension, thereby aggregating multi-scale channel context information. Features at different levels have different contributions to the task, and fusion of features at multiple levels can complement each other to obtain a comprehensive feature expression.

As discussed in [Wu *et al.*, 2019], low-level features with greater spatial resolution demand more computing resources than high-level features, but contribute less to the performance of deep integration models. Motivated by this observation, we only carry out ACFM in high-level features. Specifically, we call f_i ($i = 3, 4, 5$) as high-level features, and the

cross-level fusion process can be described as follows:

$$F_{ab} = \mathcal{M}(F_a \uplus F_b) \otimes F_a \oplus (1 - \mathcal{M}(F_a \uplus F_b)) \otimes F_b, \quad (1)$$

where $\mathcal{M}$ represents the MSCA, F_a and F_b represent the two input features. $\oplus$ represents the initial fusion of F_a and F_b, that is, F_b is up-sampled twice and then added element-wisely with F_a. $\otimes$ represents element-wise multiplication. And $(1 - \mathcal{M}(F_a \uplus F_b))$ corresponds to the dotted line in Figure 2. After that, the feature F_{ab} is fed into a 3×3 convolutional layer, followed by batch normalization and a *ReLU* activation function. Finally, we obtain the cross-level fused feature F.

3.3 Dual-branch Global Context Module

We have adopted the proposed ACFM to fuse multi-scale features at different levels, which introduces the multi-scale channel attention mechanism to obtain informative attention-based fused features. Besides, global context information is critical for improving the camouflaged object detection performance. Thus, we propose a Dual-branch Global Context Module (DGCM) to exploit rich global context information within the fused cross-level features. Specifically, the output $F \in \mathbb{R}^{W \times H \times C}$ of the ACFM is fed to two branches though a convolutional operation and average pooling, respectively. We can obtain the sub features $F_c \in \mathbb{R}^{W \times H \times C}$ and $F_p \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times C}$. To learn multi-scale attention-based feature representations, F_c and F_p are first fed to the MSCA component. Element-wise multiplication is adopted to fuse the 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 $F_{pm} \in \mathbb{R}^{\frac{W}{2} \times \frac{H}{2} \times C}$. After that, we directly fuse the features from the two branches using an addition operation to obtain F_{cpm}. Finally, a residual structure is used to fuse F and F_{cpm} to get F'. The above process can be described as follows:

$$\begin{cases} F_c = \mathcal{C}(F), F_{cm} = F_c \otimes \mathcal{M}(F_c), \\ F_p = \mathcal{C}(\mathcal{P}(F)), F_{pm} = F_p \otimes \mathcal{M}(F_p), \\ F' = \mathcal{C}(F \oplus \mathcal{C}(F_{cm} \oplus \mathcal{U}(F_{pm}))), \end{cases} \quad (2)$$

where $\mathcal{C}$, $\mathcal{P}$, $\mathcal{M}$, and $\mathcal{U}$ represent the convolution, average pooling, MSCA, and upsampling operations, respectively. It is worth noting that the proposed DGCM is used to enhance the fusion features of ACFM, which makes our model can adaptively extract multi-scale information from a specific level during the training phase. The detailed structure is shown in Figure 2.

3.4 Loss Function

The binary cross-entropy loss ($\mathcal{L}_{BCE}$) is used to independently calculate the loss of each pixel to form a pixel restriction on the network. In order to make up for its deficiency of ignoring the global structure, [Qin *et al.*, 2019] introduced IoU loss($\mathcal{L}_{IoU}$) to focus on the global structure to form a global restriction on the network. These losses are treated equally to all pixels, ignoring the differences between pixels. Based on this, [Wei *et al.*, 2020] improved the above two losses into the weighted binary cross-entropy loss ($\mathcal{L}_{BCE}^w$) and IoU loss ($\mathcal{L}_{IoU}^w$). By calculating the difference between the center pixel and its surrounding environment, each pixel is assigned a different weight, so that the hard pixels can be obtained more

attention. In summary, the loss function of our model is defined as: $\mathcal{L} = \mathcal{L}_{IoU}^w + \mathcal{L}_{BCE}^w$. The network prediction result P is up-sampled to have the same size with ground-truth map G. Thus, the total loss for the proposed model can be formulated as: $\mathcal{L}_{total} = \mathcal{L}(G, P)$.

4 Experiments

In this section, we first provide the implementation details, datasets and evaluation metrics. Then, we provide quantitative and qualitative comparison results, and present results of ablation studies to validate the effectiveness of key modules.

4.1 Implementation Details

We implement our model using PyTorch and adopt Res2Net-50 [Gao *et al.*, 2019] pretrained on ImageNet as our network backbone. We carry out a multi-scale training strategy $\{0.75, 1, 1.25\}$ instead of data augmentation and resize all input images to 352×352. We utilize the AdaX [Li *et al.*, 2020] optimization algorithm to optimize the overall parameters by setting the initial learning rate as $1e - 4$. Accelerated by an NVIDIA Tesla P40 GPU, the whole network takes about 4 hours to converge over 40 epochs for training with a batch size of 32. During training, the learning rate decays 10 times after 30 epochs.

4.2 Datasets

We conduct experiments on three public benchmark datasets for camouflaged object detection. The details of each dataset are as follows:

- CHAMELEON [Fan *et al.*, 2021a] contains 76 camouflaged images, which are collected through the Google search engine with the keyword "camouflage animals".

- CAMO [Le *et al.*, 2019] has 1,25k images (1k for training, 0.25k for testing), including 8 categories.

- COD10K [Fan *et al.*, 2021a] is currently the largest camouflaged object detection dataset. It consists of 5,066 camouflaged images (3040 for training, 2026 for testing), which is divided into 5 super-classes and 69 sub-classes. This dataset also provides high-quality fine annotation, reaching the level of matting.

Training/Testing Sets. Following the training setting in [Fan *et al.*, 2021a], for both CAMO and COD10K datasets, we use the default training sets. Then, we evaluate our model and all compared methods on the whole CHAMELEON dataset and the test sets of CAMO and COD10K.

4.3 Evaluation Metrics

To comprehensively compare our proposed model with other state-of-the-art methods, we utilize four popular metrics to evaluate the COD performance. ● Mean absolute error (MAE, M) [Perazzi *et al.*, 2012] evaluates the average pixel-level relative error between the normalized prediction and the ground truth. ● Weighted F-measure (F_β^w) [Margolin *et al.*, 2014] is an overall performance measurement that synthetically considers both weighted precision and weighted recall. ● S-measure (S_α) [Fan *et al.*, 2017] computes the object-aware and region-aware structure similarities between

Method	Year	CHAMELEON				CAMO-Test				COD10K-Test			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
FPN	2017	0.794	0.783	0.590	0.075	0.684	0.677	0.483	0.131	0.697	0.691	0.411	0.075
MaskRCNN	2017	0.643	0.778	0.518	0.099	0.574	0.715	0.430	0.151	0.613	0.748	0.402	0.080
PSPNet	2017	0.773	0.758	0.555	0.085	0.663	0.659	0.455	0.139	0.678	0.680	0.377	0.080
UNet++	2018	0.695	0.762	0.501	0.094	0.599	0.653	0.392	0.149	0.623	0.672	0.350	0.086
PiCANet	2018	0.769	0.749	0.536	0.085	0.609	0.584	0.356	0.156	0.649	0.643	0.322	0.090
MSRCNN	2019	0.637	0.686	0.443	0.091	0.617	0.669	0.454	0.133	0.641	0.706	0.419	0.073
BASNet	2019	0.687	0.721	0.474	0.118	0.618	0.661	0.413	0.159	0.634	0.678	0.365	0.105
PFANet	2019	0.679	0.648	0.378	0.144	0.659	0.622	0.391	0.172	0.636	0.618	0.286	0.128
CPD	2019	0.853	0.866	0.706	0.052	0.726	0.729	0.550	0.115	0.747	0.770	0.508	0.059
HTC	2019	0.517	0.489	0.204	0.129	0.476	0.442	0.174	0.172	0.548	0.520	0.221	0.088
EGNet	2019	0.848	0.870	0.702	0.050	0.732	0.768	0.583	0.104	0.737	0.779	0.509	0.056
ANet-SRM	2019	‡	‡	‡	‡	0.682	0.685	0.484	0.126	‡	‡	‡	‡
SINet	2020	0.869	0.891	0.740	0.044	0.751	0.771	0.606	0.100	0.771	0.806	0.551	0.051
MirrorNet	2020	‡	‡	‡	‡	0.741	0.804	0.652	0.100	‡	‡	‡	‡
C^2F-Net (Ours)	2021	**0.888**	**0.935**	**0.828**	**0.032**	**0.796**	**0.854**	**0.719**	**0.080**	**0.813**	**0.890**	**0.686**	**0.036**

Table 1: Quantitative comparison with state-of-the-art methods for COD on three benchmark datasets using four widely used evaluation metrics (*i.e.*, S_α, E_ϕ, F_β^w, and M). "↑" / "↓" indicates that larger or smaller is better. The best results are highlighted in **Bold** fonts.

Ver.	Method	CHAMELEON				CAMO-Test				COD10K-Test			
		$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$M \downarrow$
No.1	Basic	0.882	0.928	0.813	0.033	0.774	0.829	0.684	0.090	0.805	0.880	0.668	0.038
No.2	Basic+ACFM	0.886	**0.944**	0.823	**0.028**	0.784	0.831	0.694	0.086	0.808	0.881	0.675	0.037
No.3	Basic+DGCM	0.883	0.936	0.820	0.030	0.786	0.840	0.696	0.081	0.808	0.881	0.676	**0.036**
No.4	Basic+ACFM+DGCM	**0.888**	0.935	**0.828**	0.032	**0.796**	**0.854**	**0.719**	**0.080**	**0.813**	**0.890**	**0.686**	**0.036**

Table 2: Quantitative evaluation for ablation studies on the three datasets. The best results are highlighted in **Bold** fonts. Ver. = Version.

the prediction and the ground truth. ● E-measure (E_ϕ) [Fan *et al.*, 2021b] is based on the human visual perception mechanism to evaluate the overall and local accuracy of COD.

4.4 Performance Comparison

We compare the proposed method with 14 state-of-the-art COD methods, including FPN [Lin *et al.*, 2017], MaskR-CNN [He *et al.*, 2017], PSPNet [Zhao *et al.*, 2017], UNet++ [Zhou *et al.*, 2018], PiCANet [Liu *et al.*, 2018], MSRCNN [Huang *et al.*, 2019a], BASNet [Qin *et al.*, 2019], PFANet [Zhao and Wu, 2019], CPD [Wu *et al.*, 2019], HTC [Chen *et al.*, 2019], EGNet [Zhao *et al.*, 2019], ANet-SRM [Le *et al.*, 2019], SINet [Fan *et al.*, 2021a] and Mirror-Net [Yan *et al.*, 2020]. For a fair comparison, the results of the first 13 methods are taken from [Fan *et al.*, 2021a]. Besides, the results of MirrorNet based on ResNet50 are taken from https://sites.google.com/view/ltnghia/research/camo.

Quantitative Evaluation. Table 1 summarizes the quantitative results of different COD methods on the three benchmark datasets. On the three datasets, the four evaluation indicators all have better performance than previous methods. Specifically, compared with the ResNet50-based SINet specifically designed for COD tasks, S_α increased by 4.54% on average, E_ϕ increased by 8.71% on average, and F_β^w increased by 18.35% on average.

Qualitative Evaluation. Figure 3 shows the qualitative results of different COD methods. Examples are from the five super-classes of COD10K from top to bottom: aquatic, ter-

restrial, flying, amphibious, and others. Compared with other cutting-edge models, our model can achieve better visual effects by detecting more accurate and complete camouflaged objects with rich details.

4.5 Ablation Study

In order to verify the effectiveness of each key module, we designed three ablation experiments, and the results are shown in Table 2. In No.1 (Basic) experiment, we removed all ACFMs and DGCMs while keeping RFB modules, and then simply fused the features from the last three layers via a summation operation. In No.2 (Basic+ACFM) experiment, we removed the DGCMs and directly connected the two ACFMs. In No.3 (Basic+DGCM) experiment, we replaced the ACFMs with the combined operation of upsampling and then adding, and the DGCMs remained unchanged. No.4 (Basic+ACFM+DGCM) is the complete model, which is consistent with the structure of Figure 2. Besides, we replaced all the MSCAs in the whole network with the convolution layers to verify the effectiveness of MSCA, *i.e.*, "MSCA→Conv" in Table 3.

Effectiveness of ACFM. We investigate the importance of the ACFM. From Table 2, we observe that No.2 (Basic + ACFM) outperforms No.1 (Basic), clearly showing that the attention-induced cross-level fusion module is necessary for improving performance.

Effectiveness of DGCM. We further investigate the contribution of the DGCM. We observe that No.3 (Basic + DGCM) improves No.1 (Basic) performance on the three benchmark

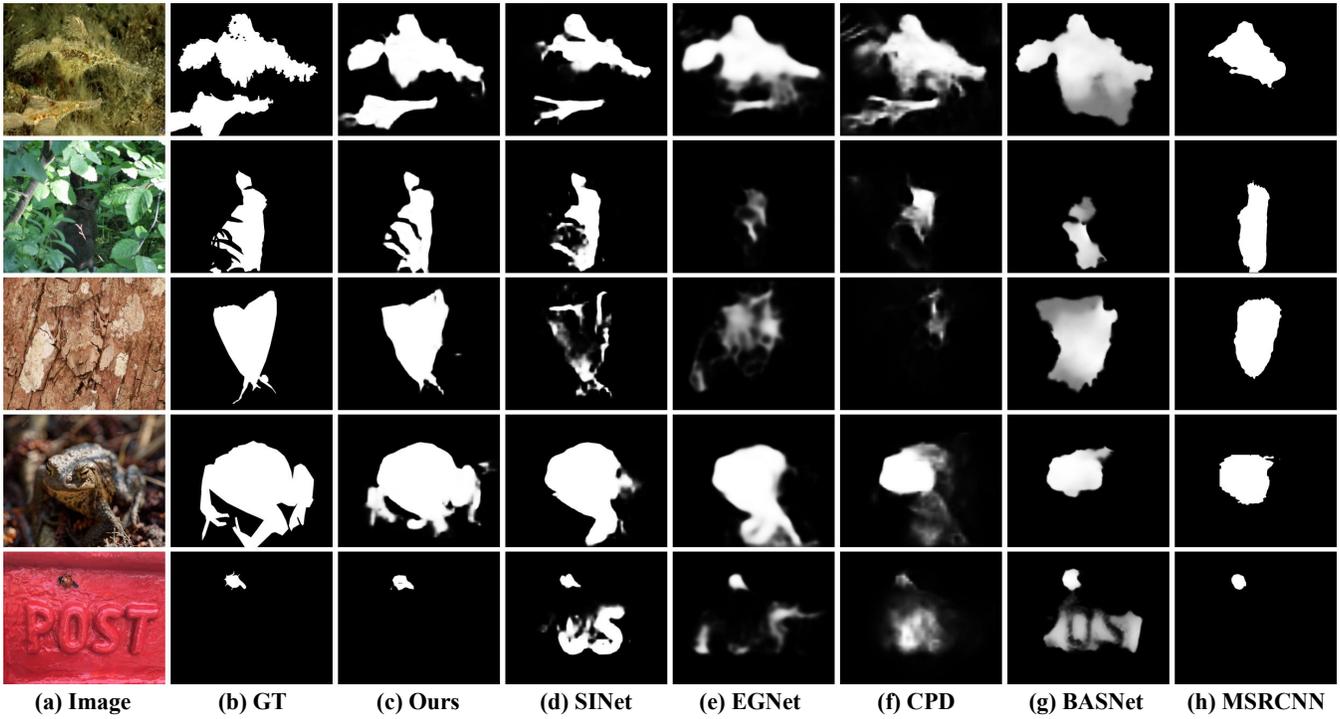| **(a) Image** | **(b) GT** | **(c) Ours** | **(d) SINet** | **(e) EGNet** | **(f) CPD** | **(g) BASNet** | **(h) MSRCNN** |

Figure 3: Qualitative results of our model and five state-of-the-art methods (*i.e.*, SINet [Fan *et al.*, 2021a], EGNet [Zhao *et al.*, 2019], CPD [Wu *et al.*, 2019], BASNet [Qin *et al.*, 2019], and MSRCNN [Huang *et al.*, 2019a]).

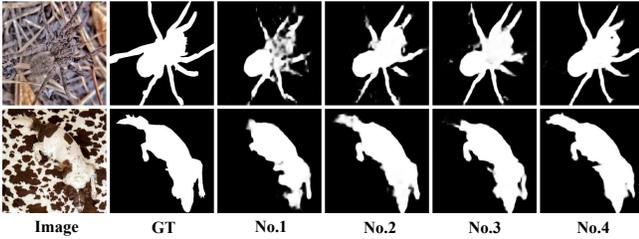| **Image** | **GT** | **No.1** | **No.2** | **No.3** | **No.4** |

Figure 4: Visual comparisons for showing the benefits of different modules. No.1, No.2, No.3 and No.4 as shown in the Table 2.

datasets. These improvements suggest that introducing dual-branch global context module can enable our model to accurately detect objects.

Effectiveness of ACFM & DGCM. To assess the combination of the ACFM and DGCM, we evaluate the performance of No.4. As shown in Table 2, our model is generally better than other settings. The visual comparison results in Figure 4 also show that the complete structure is more conducive to detecting camouflaged objects.

Effectiveness of MSCA. To validate the effectiveness of MSCA, we utilize a convolution layer instead of it (denoted "MSCA→Conv"). The comparison results are shown in Table 3. It can be observed that C²F-Net outperforms "MSCA→Conv" on three datasets in terms of two evaluation metrics. In particular, the use of MSCA significantly improves the F_β^w score by a large margin of 1.9% on the CAMO-Test dataset.

Method	CAMO-Test		COD10K-Test		CHAMELEON	
	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$	$E_\phi \uparrow$	$F_\beta^w \uparrow$
MSCA→Conv	0.843	0.700	0.885	0.679	0.934	0.826
C²F-Net	**0.854**	**0.719**	**0.890**	**0.686**	**0.935**	**0.828**

Table 3: Ablation study on MSCA.

5 Conclusion

In this paper, we propose a novel network C^2F-*Net* for camouflaged object detection, which integrates the cross-level features with the consideration of rich global context information. We propose a context-aware module, DGCM, which exploits global context information from the fused features. We integrates the cross-level features with an effective fusion module, ACFM, which integrates the features with the consideration of valuable attention cues provided by MSCA. Extensive experimental results on three benchmark datasets demonstrate that the proposed model outperforms other state-of-the-art methods.

References

[Chen *et al.*, 2018] L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE TPAMI*, 2018.

[Chen *et al.*, 2019] K. Chen, J.M. Pang, J.Q. Wang, Y. Xiong, X.X. Li, S.Y. Sun, , et al. Hybrid task cascade for instance segmentation. In *CVPR*, 2019.

[Chen *et al.*, 2020] Z.Y. Chen, Q.Q. Xu, R.M. Cong, and Q.M. Huang. Global context-aware progressive aggregation network for salient object detection. In *AAAI*, 2020.

[Dai *et al.*, 2020] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard. Attentional feature fusion. In *WACV*, 2020.

[Fan *et al.*, 2017] D.P. Fan, M.M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *ICCV*, 2017.

[Fan *et al.*, 2020] D.P. Fan, G.P. Ji, T. Zhou, G. Chen, H.Z. Fu, J.B. Shen, and L. Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *MICCAI*, 2020.

[Fan *et al.*, 2021a] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao. Concealed object detection. *IEEE TPAMI*, 2021.

[Fan *et al.*, 2021b] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng. Cognitive vision inspired object segmentation metric and loss function. *SCIENTIA SINICA Informationis*, 2021.

[Fu *et al.*, 2019] J. Fu, J. Liu, H.J. Tian, Z.W. Fang, and H.Q. Lu. Dual attention network for scene segmentation. In *CVPR*, 2019.

[Gao *et al.*, 2019] S.H. Gao, M.M. Cheng, K. Zhao, X.Y. Zhang, M.H. Yang, and P.H. Torr. Res2net: A new multi-scale backbone architecture. *IEEE TPAMI*, 2019.

[He *et al.*, 2017] K.M. He, G. Gkioxari, P. Dollár, and R. Girshick. Mask r-cnn. In *ICCV*, 2017.

[Huang *et al.*, 2019a] Z.J. Huang, L.C. Huang, Y.C. Gong, C. Huang, and X.G. Wang. Mask scoring r-cnn. In *CVPR*, 2019.

[Huang *et al.*, 2019b] Z.L. Huang, X.G. Wang, L.C. Huang, C. Huang, Y.C. Wei, and W.Y. Liu. Ccnet: Criss-cross attention for semantic segmentation. In *ICCV*, 2019.

[Jiang, 2011] Z.H. Jiang. Object modelling and tracking in videos via multidimensional features. *ISRN SP*, 2011.

[Lamdouar *et al.*, 2020] H. Lamdouar, C. Yang, W.D. Xie, and A. Zisserman. Betrayed by motion: Camouflaged object discovery via motion segmentation. In *ACCV*, 2020.

[Le *et al.*, 2019] T.N. Le, T.V. Nguyen, Z.L. Nie, M.T. Tran, and A. Sugimoto. Anabranch network for camouflaged object segmentation. *CVIU*, 2019.

[Li *et al.*, 2017] C. Li, X.Y. Zhao, and Y.Q. Wan. The camouflage color target detection with deep networks. In *ICONIP*, 2017.

[Li *et al.*, 2020] Wenjie Li, Zhaoyang Zhang, Xinjiang Wang, and Ping Luo. Adax: Adaptive gradient descent with exponential long term memory. *arXiv preprint arXiv:2004.09740*, 2020.

[Lin *et al.*, 2017] T.Y. Lin, P. Dollár, R. Girshick, K.M. He, B. Hariharan, and S. Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017.

[Liu *et al.*, 2018] N. Liu, J.W. Han, and M.H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *CVPR*, 2018.

[Liu *et al.*, 2019] J.J. Liu, Q.B. Hou, M.M. Cheng, J.S. Feng, and J.M. Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, 2019.

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

[Margolin *et al.*, 2014] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps? In *CVPR*, 2014.

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

[Mondal, 2020] A. Mondal. Camouflaged object detection and tracking: A survey. *IJIG*, 2020.

[Perazzi *et al.*, 2012] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *CVPR*, 2012.

[Qin *et al.*, 2019] X.B. Qin, Z.C. Zhang, C.Y. Huang, C. Gao, M. Dehghan, and M. Jagersand. Basnet: Boundary-aware salient object detection. In *CVPR*, 2019.

[Tan *et al.*, 2020] J. Tan, P.F. Xiong, Z.Y. Lv, K.T. Xiao, and Y.W. He. Local context attention for salient object segmentation. In *ACCV*, 2020.

[Wang *et al.*, 2018] X.L. Wang, R. Girshick, A. Gupta, and K.M. Gupta. Non-local neural networks. In *CVPR*, 2018.

[Wei *et al.*, 2020] J. Wei, S.H. Wang, and Q.M. Huang. F³net: Fusion, feedback and focus for salient object detection. In *AAAI*, 2020.

[Wu *et al.*, 2019] Z. Wu, L. Su, and Q.M. Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, 2019.

[Yan *et al.*, 2020] J.N. Yan, T.N. Le, K.D. Nguyen, M.T. Tran, T.T. Do, and T.V. Nguyen. Mirrornet: Bio-inspired adversarial attack for camouflaged object segmentation. *arXiv preprint arXiv:2007.12881*, 2020.

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

[Zhang *et al.*, 2018] L. Zhang, J. Dai, H.C. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In *CVPR*, 2018.

[Zhao and Wu, 2019] T. Zhao and X.Q. Wu. Pyramid feature attention network for saliency detection. In *CVPR*, 2019.

[Zhao *et al.*, 2017] H.S. Zhao, J.P. Shi, X.J. Qi, X.G. Wang, and J.Y. Jia. Pyramid scene parsing network. In *CVPR*, 2017.

[Zhao *et al.*, 2019] J.X. Zhao, J.J. Liu, D.P. Fan, Y. Cao, J.F. Yang, and M.M. Cheng. Egnet: Edge guidance network for salient object detection. In *CVPR*, 2019.

[Zhou *et al.*, 2018] Z.W. Zhou, M.M.R. Siddiquee, N. Tajbakhsh, and J.M. Liang. Unet++: A nested u-net architecture for medical image segmentation. In *DLMIA and ML-CDS*. Springer, 2018.