# MFFN: Multi-view Feature Fusion Network for Camouflaged Object Detection

Dehua Zheng[1] Xiaochen Zheng[2] Laurence T. Yang[1*] Yuan Gao[3] Chenlu Zhu[4] Yiheng Ruan[4]
[1]Huazhong University of Science and Technology, China [2]ETH Zürich, Switzerland
[3]Hainan University, China [4]Hubei Chutian Expressway Digital Technology, China
dwardzheng@hust.edu.cn xzheng@student.ethz.ch ltyang@gmail.com

## Abstract

*Recent research about camouflaged object detection (COD) aims to segment highly concealed objects hidden in complex surroundings. The tiny, fuzzy camouflaged objects result in visually indistinguishable properties. However, current single-view COD detectors are sensitive to background distractors. Therefore, blurred boundaries and variable shapes of the camouflaged objects are challenging to be fully captured with a single-view detector. To overcome these obstacles, we propose a behavior-inspired framework, called **M**ulti-view **F**eature **F**usion **N**etwork (MFFN), which mimics the human behaviors of finding indistinct objects in images, i.e., observing from multiple angles, distances, perspectives. Specifically, the key idea behind it is to generate multiple ways of observation (**multi-view**) by data augmentation and apply them as inputs. MFFN captures critical boundary and semantic information by comparing and fusing extracted multi-view features. In addition, our MFFN exploits the dependence and interaction between views and channels. Specifically, our methods leverage the complementary information between different views through a two-stage attention module called **C**o-**a**ttention of **M**ulti-**v**iew (CAMV). And we design a local-overall module called **C**hannel **F**usion **U**nit (CFU) to explore the channel-wise contextual clues of diverse feature maps in an iterative manner. The experiment results show that our method performs favorably against existing state-of-the-art methods via training with the same data. The code will be available at* `https://github.com/dwardzheng/MFFN_COD`.

## 1. Introduction

Camouflage is a mechanism [3] by which organisms protect themselves in nature. Camouflaged object detection (COD) is a countermeasure against the camouflage mechanism, aiming to capture the slight differences between the object and the background to obtain accurate detection results. Unlike general object detection and salient object detection, in which the objects and background can be easily distinguished by human eyes or advanced deep learning models, COD is more challenging because it requires a sufficient amount of visual input and prior knowledge [41] to address the complicated problem caused by the highly intrinsic similarity between the target object and the background. Thus, COD has a wide range of valuable applications in promoting the search and detection of biological species [38], assisting the medical diagnosis with medical images [36, 12], and improving the detection of pests and diseases in agriculture [9].

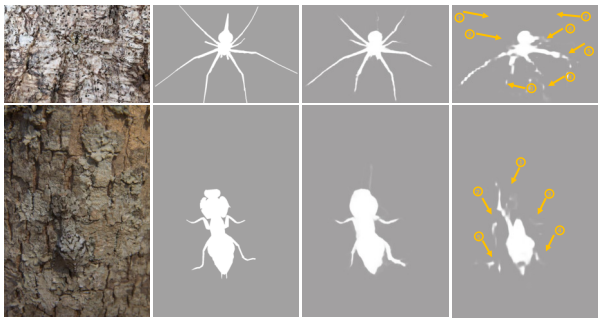



Figure 1: Visualization of camouflaged animal detection. The state-of-the-art and classic single-view COD model SINet [5] is confused by the background sharing highly similarities with target objects and missed a lot of boundary and region shape information (indicated by orange arrows). Our multi-view scheme will eliminate these distractors and perform more efficiently and effectively.

Recently, many researches put emphasis on learning from a fixed single view with either auxiliary tasks [17, 27, 29, 48, 55, 14], uncertainty discovery [19, 24], or vision transformers [47, 33] and their proposed methods achieved significant progress. Nevertheless, due to visual insignificance of camouflaged objects and contextual insufficiency from single-view input, they are still striving to precisely

*Corresponding author.

recognize camouflaged objects and their performance needs to be improved. We found that the current COD methods are easily distracted by negative factors from deceptive background/surroundings, as illustrated in Fig. 1. As a result, it is hard to mine discriminative and fine-grained semantic cues of camouflaged objects, making accurately segment camouflaged objects from a confusing background and predict some uncertain regions incapable. Meanwhile, we learn that when people observe a concealed object in images, they usually adjust the viewing distance, change the viewing angle, and change the viewing position to find the target object more accurately. Inspired by it, we aim to design a simple yet efficient and effective strategy. The aforementioned considerations motivate us to consider the semantic and context exploration problem with **multi-view**. We argue that corresponding clues, correlations, and mutual constraints can be better obtained by utilizing information from different viewpoint of the scene (e.g., changing observation distances and angles) as complementary. Furthermore, we argue that carefully designing the encoded feature fusion modules can help the encoder learn accurate information corresponding to boundary and semantics. Taking these into mind, our research will focus on the following three aspects: *(1) how to exploit the effects of different types of views on COD task, and the combination of multi-view features to achieve the best detection effect; (2) how to better fuse the features from multiple views based on correlation awareness and how to enhance the semantic expression ability of multi-view feature maps without increasing model complexity; (3) how to incrementally explore the potential context relationships of a multi-channel feature map*.

To solve our concerned pain points of COD task, we propose a **M**ulti-view **F**eature **F**usion **N**etwork (MFFN) for the COD task to make up for the semantic deficiency of fixed view observation. First, we use the multi-view raw data, which are generated by different data augmentation, as the inputs of a backbone extractor with shared weights. We implement a ResNet model as the backbone extractor integrating the feature pyramid network (FPN) [22] to focus on object information of different scales. In addition, we design a **C**o-**a**ttention of **M**ulti-**v**iew (CAMV) module to integrate multi-view features and to explore the correlation between different view types. CAMV consists of two stages of attention operation. In the first stage, the inherent correlation and complementary analysis are mainly conducted for multiple viewing distances and angles to obtain the view features with a unified scale. In the second stage, the external constraint relations between viewing angles and distances are further leveraged to enhance feature maps' semantic expression. For the enhanced multi-view feature tensor, we design a **C**hannel **F**usion **U**nit (CFU) to further exploit the correlation between contexts. In the CFU module, we first carry out up-down feature interaction between channel dimensions and then carry out progressive iteration on the overall features. CAMV is applied to observe the multi-view attention features of different size feature maps of FPN architecture. The CFU module contains the previous layer's information as each size's feature maps are eventually restored to their original size. Finally, the final prediction results are obtained by sigmoid operation. The prediction further benefits from UAL design.

Our contribution can be summarized as follows: 1) We propose MFFN model to solve the challenging problems faced by single-view COD models. MFFN can capture complementary information acquired by different viewing angles and distances and discover the progressive connection between contexts.

2) We design the CAMV module to mine the complementary relationships within and between different types of view features and enhance the semantic expression ability of multi-view feature tensors, and use the CFU module to conduct progressive context cue mining.

3) Our model is tested on three datasets of CHAMELEON [37], COK10K [5] and NC4K [27], and quantitative analysis is conducted on five general evaluation indicators of $S_m$ [6], $F_\beta^w$ [28], $MAE$, $F_\beta$ [1] and $E_m$ [7], all of which achieved superior results.

## 2. Related work

**Salient Object Detection (SOD).** SOD is a kind of segmentation task in essence. It calculates saliency map first and then merges and segmented saliency object. In previous studies, traditional methods based on manual features pay more attention to color [2, 21], texture [46, 21], contrast [34, 15] and so on, but lack advantages in complex scenes and structured description. With the development of CNN, SOD algorithm has achieved leapfrog development. Li *et al*. [20] combines local information with global information to overcome the problem of highlighting object boundary but not the overall object in the model based on local. The model structure design idea of multi-level features, has been more widely applied in [23, 54, 13, 18]. Similar to COD, clear boundary information is crucial for SOD task [35, 52, 39]. The development of attention mechanism provides more schemes for exploring the correlation between channel dimension and spatial dimension [32, 8, 42]. The application of attention mechanism improves the performance of SOD model [26, 51, 44]. SOD faces simpler background surroundings. Although excellent performance can be obtained by applying relevant models to COD task, specific design is still needed to remove the interference from the background surroundings.

**Camouflaged Object Detection (COD).** In recent years, some researches applied multi-task learning to detect the camouflaged objects. Le *et al*. [17] introduced the binary

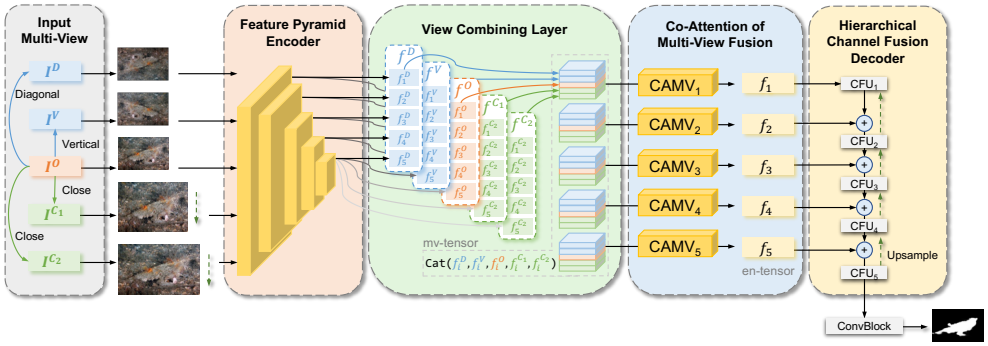



Figure 2: Overview of our model structure. We generate multiple views (**D**iagonally and **V**ertically flipped views, **C**lose looking views) of the data by different transformation methods. The shared pyramid feature encoder is applied to extract hierarchical features of different scales corresponding to different view choices. The view combining layer concatenates features of same level from different views ($f_i^D, f_i^V, f_i^O, f_i^{C_1}, f_i^{C_2}$) channel-wisely and output multi-view feature tensors (*mv-tensor*s). The model feeds *mv-tensor*s into CAMVs and obtain multi-view enhanced feature tensor (*en-tensor*) $f_i$. CAMV is adopted to fuse features and aggregate vital clues between different views by a two-stage co-attention mechanism. The *en-tensor*s are further decoded and the contextual correlation are exploited by hierarchical channel fusion unit simultaneously. In the end, a probability map of camouflaged object in the input image is computed by several convolutional blocks.

classification task as the second branch and auxiliary task of camouflaged object segmentation. Zhu *et al*. [55] proposed a new boundary-guided separated attention network (BSA-NET), which uses two streams of separated attention modules to highlight the boundaries of camouflaged objects. Lv *et al*. [27] proposed a multi-task learning framework to jointly localize and segment the camouflaged objects while inferring their ranks. Zhai *et al*. [48] designed a mutual graph learning model to detect the edge and region of the objects simultaneously. There are some uncertainty-aware methods. Li *et al*. [19] proposed an uncertainty-aware framework containing a joint network for both salient and camouflaged object detection. Yang *et al*. [47] introduced Bayesian learning into the uncertainty-guided transformer reasoning model. Liu *et al*. [24] designed an aleatoric uncertainty estimation network to indicate the prediction awareness. Sun *et al*. [40] placed emphasis on rich global context information with the integration of cross-level features. Pei *et al*. [33] applied a one-stage location-sensing transformer and further fused the features from transformer and CNN. Some bio-inspired methods are proposed. For example, [30, 29, 5] use multi-scale information but from one single view. Meanwhile, [30] shows single-view information is not sufficient for accurately detecting camouflaged objects. We hereby argue that view generation and selection might play an important role and we aim to develop our model by mimicking the behavior of humans when understanding complicated concealed objects by altering the way they observing an image. Our proposed method exploits the visual perception knowledge and semantic cues by aggregating complementary information from multi-view. Accordingly, our model is simple yet efficient and effective to comprehensively understand scene and to accurately segment the camouflaged objects.

## 3. Method

**Motivation.** Motivated by the challenges of single-view COD models, we attempt to capture boundary and regional semantic information with rich viewing angles and flexible viewing distances. In order to merge diverse context information from features of multi-view inputs and FPN multi-level outputs, we design a feature fusion module based on two-stage attention mechanism to obtain enhanced feature tensors. It also avoids redundant structural design. To leverage the rich information contained in channel dimensions, we design a local-overall context/cues mining structure based on channel-wise integration. Meanwhile, it also enhances the information expression of the feature tensors.

### 3.1. Multi-view Generation

As shown in Fig. 1, the single-view model misses necessary boundary, region, and shape information. Inspired by human behavior, taking complementary views of observation into account will overcome this defect and we design three different views: *distance, angle, and perspective view*. We obtain different *distance views* through the resize operation with the proportional interval of the resize operation larger than 0.5 to increase the distinction. We get different *angle views* by mirror transformation, including horizontal, vertical and diagonal mirror transformation. We obtain different *perspective views* through affine transformation. Specifically, three corresponding points on the

original and the target image are selected as references to calculate the transformation matrix. The above operations are based on OpenCV and the implementation in OpenCV is in Appendix B. The ablation study proves that the combination of two angle views obtained by mirror transformation and two close distance views obtained by resize operation is an effective selection scheme. As shown in the Appendix A, our multi-view strategy can be easily transferred to the SOD task and achieve excellent performance in salient object detection (SOD) task.

### 3.2. Architecture Overview

The overview of our proposed MFFN is illustrated in Fig. 2. MFFN adopts ResNet [11] as the backbone network for feature extraction, and adopts the FPN [22] to capture feature information of different levels from different views. We design the CAMV module to merge diverse context information and to capture complementary information from encoded multi-view features. Furthermore, we applied CFU module to fuse the channel-wise context information and clues in an iterative manner. As shown in Fig. 2, given an input original image $I^O \in \mathbb{R}^{H\times W\times 3}$, we create *flipped* and *close* views by applying mirror and resize transformation. The multi-view inputs are defined as $\{I^D \in \mathbb{R}^{H\times W\times 3}, I^V \in \mathbb{R}^{H\times W\times 3}, I^O \in \mathbb{R}^{H\times W\times 3}, I^{C_1} \in \mathbb{R}^{H_1\times W_1\times 3}, I^{C_2} \in \mathbb{R}^{H_2\times W_2\times 3}\}$, where $D, V$ indicate diagonally and vertically flipped views, $O$ indicates original view, and $C_1, C_2$ represent two different scale close views. We input each observed view into a backbone network with shared weights, and obtain feature maps of different levels through FPN [22]. Then, we apply CAMV module to fuse the multi-view feature tensors from a specific FPN stage by a two-stage attention mechanism. Furthermore, we design the CFU module to mine the contextual correlation and critical clues between the multi-view enhanced feature maps $f_1, ..., f_5$. Finally, MFFN restores the feature maps to its original size by gradual upsampling structure, so as to obtain the final output results.

### 3.3. Co-attention of Multi-view

The COD methods proposed in recent years pay more attention to feature mining from a fixed view and thus ignore information complemented from different views. Inspired by the biological mechanism, visual information from different ways of observing and watching can be correlated and complemented. Based on the above discoveries, we implement CAMV, consisting of a two-stage attention mechanism to complement boundary information with features from different viewing angules and enhance semantic information with different viewing distance. CAMV reduces redundant network design through multi-view interaction and fusion.

The framework of CAMV is shown in Fig. 3. Since the scales of multiple viewing distances features $f_i^{C_1} \in \mathbb{R}^{h_1\times w_1\times c}$, $f_i^{C_2} \in \mathbb{R}^{h_2\times w_2\times c}$ differs, we first align its scale to be consistent resolution (dimension) with $f_i^O \in \mathbb{R}^{h\times w\times c}$ through downsampling. Then we carry out post-processing to $f_i^V, f_i^D, f_i^O, f_i^{C_1}, f_i^{C_2}$ and we only need to post-process the features from different angles $f_i^V, f_i^D, f_i^O$ while keeping the resolution unchanged. After post-processing, we cross-concatenate encoded multi-view feature tensors $f_i^V, f_i^D, f_i^O, f_i^{C_1}, f_i^{C_2}$ from same level $i$ (the *mv-tensor* in Fig. 2) to compose one multi-view enhanced feature tensor (the *en-tensor* in Fig. 2). We design a two-stage attention module to enhance feature interaction and correlational clues mining from different views.

The attention of the first stage aims to aggregate the correlative clues of viewing distance and viewing angle respectively. Taking the feature tensor $\{f_i^D, f_i^V, f_i^O\}$ from three viewing angles as an example, we first compress channel feature through the convolution layer to obtain $f_i^{Ang} \in \mathbb{R}^{h\times w\times c}$, and then input $f_i^{Ang}$ into three parallel tensor multiple modulus multiplication modules to calculate attention. The process is formulated as

$$
\begin{aligned}
f_i^{Ang} &= \texttt{ReLU}(\texttt{Conv}(\texttt{Cat}(f_i^D, f_i^V, f_i^O))) \\
u_A &= \sigma(f_i^{Ang} \times_1 U_{A_1} \times_2 U_{A_2} \times_3 U_{A_3}) \\
u_B &= \sigma(f_i^{Ang} \times_1 U_{B_1} \times_2 U_{B_2} \times_3 U_{B_3}) \\
u_C &= \sigma(f_i^{Ang} \times_1 U_{C_1} \times_2 U_{C_2} \times_3 U_{C_3}) \\
F_i^{Ang} &= f_i^D \odot u_A + f_i^V \odot u_B + f_i^O \odot u_C
\end{aligned}
\tag{1}
$$

where $\{u_A, u_B, u_C\}$ are attention factors, $\{f_i^D, f_i^V, f_i^O\}$ indicate feature tensors from three different viewing angles, $F_i^{Ang}$ represents *en-tensor* of the first stage attention, $\sigma$ denotes sigmoid function scaling the weight value into (0, 1). $\texttt{Cat}()$ is the concatenation operation along channel and $\texttt{ReLU}()$ represents the activation function. $\{U_{A_i}, U_{B_i}, U_{C_i}\}$ represent the parameter matrix of attention factor calculation modules based on tensor multiple modulus multiplication operation, $\times_i$ represents modular multiplication [16], $\odot$ means element-by-element multiplication. Similarly, we can process the feature tensor $F_i^{Dist}$ of distance-based views after fusion by the same operation. Through such two parallel internal-attention (*In-att*) feature fusion modules, we can enhance the semantic information of the feature maps from different angles and distance.

In the second stage of the attention mechanism, we further interact $F_i^{Ang}$ and $F_i^{Dist}$. As shown in Fig. 3, the features of discriminative viewing angles $F_i^{'Ang}$ is obtained by boundary separation based on self-attention, and $F_i^{'Ang}$ will be used as a complement to $F_i^{Dist}$. Furthermore, we concatenate $F_i^{'Ang}$ and $F_i^{Dist}$ together to obtain the multi-view intermediate feature tensor $F_{MV}$. Finally, we fuse $F_{MV}$ to obtain the final output of CAMV module. The specific

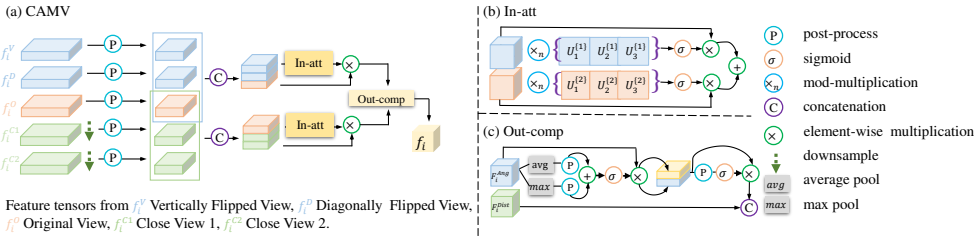



Figure 3: The architecture of our CAMV module. First, the two view types are processed by intra-class attention (*In-att*) mechanism, and then the two view types are fused by the complementation of external classes (*out-comp*). In a nutshell, CAMV consists of a two-stage attention mechanism.

process is defined by the following formula:

$$
\begin{aligned}
F_i^{A1} &= \texttt{Conv}(\texttt{ReLU}(\texttt{Conv}(\texttt{AvgPool}(F_i^{Ang})))) \\
F_i^{A2} &= \texttt{Conv}(\texttt{ReLU}(\texttt{Conv}(\texttt{MaxPool}(F_i^{Ang})))) \\
F_i^{Ang} &= F_i^{Ang} \odot \sigma(F_i^{A1} + F_i^{A2}) \\
F_i^{'A1} &= \text{avg}(F_i^{Ang}) \\
F_i^{'A2} &= \max(F_i^{Ang}) \\
F_i^{'Ang} &= F_i^{Ang} \odot \sigma(\texttt{Conv}(\texttt{Cat}(F_i^{'A1}, F_i^{'A2})))
\end{aligned}
\tag{2}
$$

where $\texttt{MaxPool}()$ and and $\texttt{AvgPool}()$ mean maximum and average pooling respectively, $mean()$ indicates taking the mean of the elements and $max()$ indicates taking the maximum of the elements along the channel dimension. Generally speaking, $\texttt{AvgPool}()$ can preserve more background information, and $\texttt{MaxPool}()$ can preserve more texture information. Thus, abundant boundary information will help to better capture the blurred differences in shape, color, scale and so on between the object and the background.

Through the two-stage attention blocks in CAMV, we carry out implicit interaction and semantic correlation mining for features from different views. The viewing angle and distance well complement the differences between them. The boosted feature expression makes the camouflaged object more clearly separate from the background surroundings. To sum up, CAMV aggregates feature maps from different views, and integrates the auxiliary boundary information into the main branch that incorporates the distance views. Thus, we will transmit a semantically enhanced and more compact feature map into the next processing unit.

### 3.4. Channel Fusion Unit

The input of CFU module is the integrated feature map $f_i$ from CAMV, which is an embedding obtained by fusing features from different views. CFU splits the feature map $f_i$ from CAMV module into $j$ chunks $\{f_i^1, f_i^2, ..., f_i^k, ..., f_i^j\}$ along the channel dimension, where $k$ indicates the index of different chunks. All chunks $\{f_i^1, f_i^2, ..., f_i^k, ..., f_i^j\}$ have a consistent number of channels. CFU executes channel-wise local interaction process (CLIP) between adjacent chunks $f_i^{k-1}$ and $f_i^k$ to connect all channels of $f_i^{k-1}$ and $f_i^k$. The output of CLIP is further interacted with the next chunk $f_i^{k+1}$. In this way, all channels of $f_i$ interact with each other. Then, the outputs of all CLIP will be reassembled into one feature map, which will be used as the input of the overall iteration, giving full consideration to the idea of consistency between the overall and the local. The CLIP is described as follows:

$$
\textbf{CLIP}(f_i^{k+1}, f_i^k) = \texttt{Tucker}(\texttt{Cat}(f_i^{k+1}, \texttt{Conv}(f_i^k)))
\tag{3}
$$

where $\texttt{Tucker}()$ represents the interaction fusion operation based on tensor multiple modulus multiplication, which can filter out the redundant features by splicing and make its semantic expression more compact.

The overall progressive iteration (OPI), which aims to explore the potential semantic relevance of context, conducts progressive iterations from the overall. This iterative hybrid strategy helps to obtain a more powerful feature representation. The output $z_i$ of the final CLIP is the input of OPI. We define the initial value of $z_i$ as $z_i^0$. For each OPI,

$$
\begin{aligned}
z_i^0 &= \texttt{CBR}(z_i) \\
z_i^{s+1} &= \texttt{CBR}(z_i^s + z_i^0)
\end{aligned}
\tag{4}
$$

where $\texttt{CBR}()$ represents a block unit mainly based on convolution layer, including the combination of multiple convolutional layers, batch normalization, and activation layers. We adopt FPN [22] architecture as the feature extractor, which results in multi-level feature maps of different scales. We adopt a progressive upsampling method to gradually restore the feature maps of different levels to be consistent resolution. Finally, a fusion unit and sigmoid function are used to obtain the predicted results.

Table 1: Comparison of evaluation results of different models on CHAMELEON,COD10K and NC4K. The best model results will be highlighted in **green**.

| Accepted by | Model | CHAMELEON | | | | | COD10K | | | | | NC4K | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ |
| Salient Object Detection / Medical Image Segmentation | | | | | | | | | | | | | | | | |
| CVPR2018 | PiCANet [26] | 0.765 | 0.552 | 0.085 | 0.618 | 0.846 | 0.696 | 0.415 | 0.081 | 0.489 | 0.788 | 0.758 | 0.57 | 0.088 | 0.64 | 0.835 |
| CVPR2019 | BASNet [35] | 0.847 | 0.771 | 0.044 | 0.795 | 0.894 | 0.661 | 0.432 | 0.071 | 0.486 | 0.749 | 0.695 | 0.546 | 0.095 | 0.61 | 0.785 |
| CVPR2019 | CPD [44] | 0.857 | 0.731 | 0.048 | 0.771 | 0.923 | 0.75 | 0.531 | 0.053 | 0.595 | 0.853 | 0.787 | 0.645 | 0.072 | 0.705 | 0.866 |
| CVPR2019 | PoolNet [25] | 0.845 | 0.69 | 0.054 | 0.749 | 0.933 | 0.74 | 0.506 | 0.056 | 0.575 | 0.844 | 0.785 | 0.635 | 0.073 | 0.699 | 0.865 |
| ICCV2019 | EGNet [52] | 0.797 | 0.649 | 0.065 | 0.702 | 0.884 | 0.736 | 0.517 | 0.061 | 0.582 | 0.854 | 0.777 | 0.639 | 0.075 | 0.696 | 0.864 |
| AAAI2020 | F3Net [43] | 0.848 | 0.744 | 0.047 | 0.77 | 0.917 | 0.739 | 0.544 | 0.051 | 0.593 | 0.819 | 0.78 | 0.656 | 0.07 | 0.705 | 0.848 |
| ICCV2019 | SCRN [45] | 0.876 | 0.741 | 0.042 | 0.787 | 0.939 | 0.789 | 0.575 | 0.047 | 0.651 | 0.88 | 0.83 | 0.698 | 0.059 | 0.757 | 0.897 |
| CVPR2020 | CSNet [10] | 0.856 | 0.718 | 0.047 | 0.766 | 0.928 | 0.778 | 0.569 | 0.047 | 0.634 | 0.871 | 0.75 | 0.603 | 0.088 | 0.655 | 0.793 |
| CVPR2020 | SSAL [50] | 0.757 | 0.639 | 0.071 | 0.702 | 0.856 | 0.668 | 0.454 | 0.066 | 0.527 | 0.7789 | 0.699 | 0.561 | 0.093 | 0.644 | 0.812 |
| CVPR2020 | UCNet [49] | 0.88 | 0.817 | 0.036 | 0.836 | 0.941 | 0.776 | 0.633 | 0.042 | 0.681 | 0.867 | 0.811 | 0.729 | 0.055 | 0.775 | 0.886 |
| CVPR2020 | MINet [31] | 0.855 | 0.771 | 0.036 | 0.802 | 0.937 | 0.77 | 0.608 | 0.042 | 0.657 | 0.859 | 0.812 | 0.72 | 0.056 | 0.764 | 0.887 |
| CVPR2020 | ITSD [53] | 0.814 | 0.662 | 0.057 | 0.705 | 0.901 | 0.767 | 0.557 | 0.051 | 0.615 | 0.861 | 0.811 | 0.679 | 0.064 | 0.729 | 0.883 |
| MICCAI2020 | PraNet [4] | 0.86 | 0.763 | 0.044 | 0.789 | 0.935 | 0.789 | 0.629 | 0.045 | 0.671 | 0.879 | 0.822 | 0.724 | 0.059 | 0.763 | 0.888 |
| Camouflaged Object Detection | | | | | | | | | | | | | | | | |
| CVPR2020 | SINet [5] | 0.872 | 0.806 | 0.034 | 0.827 | 0.946 | 0.776 | 0.631 | 0.043 | 0.679 | 0.874 | 0.808 | 0.723 | 0.058 | 0.769 | 0.883 |
| CVPR2021 | SLSR [27] | 0.89 | 0.822 | 0.03 | 0.841 | 0.948 | 0.804 | 0.673 | 0.037 | 0.715 | 0.892 | 0.84 | 0.766 | 0.048 | 0.804 | 0.907 |
| CVPR2021 | MGL-R [48] | 0.893 | 0.812 | 0.031 | 0.833 | 0.941 | 0.814 | 0.666 | 0.035 | 0.71 | 0.89 | 0.833 | 0.739 | 0.053 | 0.782 | 0.893 |
| CVPR2021 | PFNet [29] | 0.882 | 0.81 | 0.033 | 0.828 | 0.945 | 0.8 | 0.66 | 0.04 | 0.701 | 0.89 | 0.829 | 0.745 | 0.053 | 0.784 | 0.898 |
| CVPR2021 | UJSC* [19] | 0.891 | 0.833 | .0.030 | 0.847 | 0.955 | 0.809 | 0.684 | 0.035 | 0.721 | 0.891 | 0.842 | 0.771 | 0.047 | 0.806 | 0.907 |
| IJCAI2021 | C2FNet [40] | 0.888 | 0.828 | 0.032 | 0.844 | 0.946 | 0.813 | 0.686 | 0.036 | 0.723 | 0.9 | 0.838 | 0.762 | 0.049 | 0.795 | 0.904 |
| ICCV2021 | UGTR [47] | 0.888 | 0.794 | 0.031 | 0.819 | 0.94 | 0.817 | 0.666 | 0.036 | 0.711 | 0.89 | 0.839 | 0.746 | 0.052 | 0.787 | 0.899 |
| CVPR2022 | ZoomNet [30] | 0.902 | 0.845 | 0.023 | 0.864 | 0.958 | 0.838 | 0.729 | 0.029 | 0.766 | 0.911 | 0.853 | 0.784 | 0.043 | 0.818 | 0.912 |
| **OURS** | **MFFN** | **0.905** | **0.852** | **0.021** | **0.871** | **0.963** | **0.846** | **0.745** | **0.028** | **0.782** | **0.917** | **0.856** | **0.791** | **0.042** | **0.827** | **0.915** |

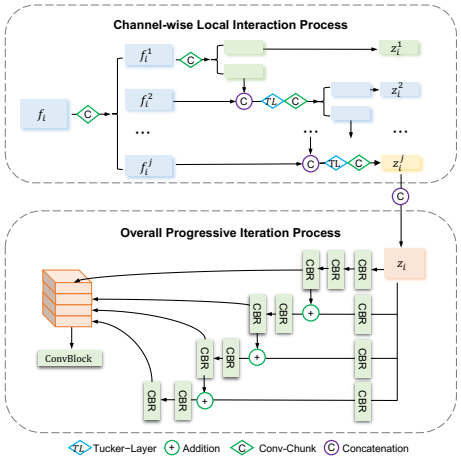



Figure 4: The architecture of the CFU module. CFU first performs feature interaction based on Tucker layer and convolution layer through channel expansion and split, and then obtains the final output through up-down correlation cue mining.

### 3.5. Loss Functions

Binary cross entropy loss (BECL) is often used in various image segmentation tasks, and its mathematical form is as follows:

$$l_{BCEL}^{i,j} = -g_{i,j} \log p_{i,j} - (1 - g_{i,j}) \log(1 - p_{i,j}) \quad (5)$$

where $g_{i,j} \in \{0, 1\}$ and $p_{i,j} \in [0, 1]$ denote the ground truth and the predicted value at position(i,j), respectively. Because the camouflage object is often seriously disturbed by the background surroundings. As a result, the model produces serious fuzziness and uncertainty in prediction. For this reason, we design uncertainty perceived loss (UAL) [30] as an auxiliary of BCEL to improve the prediction ability of the model for camouflaged objects. And its mathematical form is as follows:

$$l_{UAL}^{i,j} = 1 - |2p_{i,j} - 1|^2 \quad (6)$$

finally,the total loss function can be written as:

$$L = L_{DCEL} + \lambda L_{UAL} \quad (7)$$

We use the UAL form of the quadratic power because the quadratic curve has a gentle gradient around 0 and 1 while maintaining a reasonable penalty interval around 0.5. The cosine strategy is used to dynamically adjust the $\lambda$.

## 4. Experiments

### 4.1. Experiment Setup

**Datasets.** We use four COD datasets, CAMO [17], CHAMELEON [37], COD10K [5] and NC4K [27]. CAMO consists of 1,250 camouflaged and 1,250 non-camouflaged images. CHAMELEON contains 76 hand-annotated images. COD10K includes 5,066 camouflaged, 3,000 background. NC4K is another COD testing dataset including

4,121 images. In this work, we use CAMO and COD10K to construct a training set containing 4,040 camouflage images. To fully verify the generalization ability of the model, we conducted tests on CHAMELEON and NC4K that did not participate in the training, as well as the rest of COD10K. The train, validation, and test sets have been split by default in their original corresponding datasets.

**Evaluation Metrics.** To facilitate comparison with previous methods, we adopt the following evaluation indicators: Structure-measure ($S_m$) which is used as an assessment of structural similarity, F-measure ($F_\beta$) which is used to balance estimates of accuracy and recall rates, weighted F-measure ($F_\beta^w$), mean absolute error (MAE), Enhanced-alignment measure ($E_m$), which considers both the global average of image and local pixel matching.

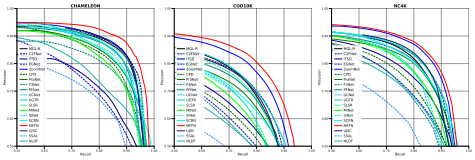
**Experiment environment.** The entire mod code is implemented based on PyTorch. In the feature extraction stage, ResNet-50 model pretrained on ImageNet is adopted.

**Hyperparameter setting.** To verify the validity of the model itself, we followed the same hyperparameter settings as most of the comparison models. SGD with a momentum of 0.9 and a weight decay of 0.0005 was chosen as the optimizer. We initialize the learning rate to 0.01 and follow the cosine preheat decay strategy. In addition, we set batch size to 8, we trained our model in the training set, and evaluated it in the independent validation set every three epochs. When 60% results of the evaluation metrics of the model on the validation set did not exceed the previous evaluation results, the training was stopped. For more detailed information, please see Appendix D.

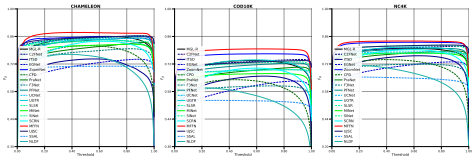
### 4.2. Comparisons with State-of-the-Arts

Due to the wide application value of COD, researchers have published a number of models with excellent performance in top academic conferences and journals. We selected models that have appeared in mainstream computer vision conferences in recent years for comparison and applied the published results. In addition, during the training process, we follow the same backbone and hyperparameters' settings (*i.e.* batch size, learning rate, and optimizer) as most models.

As shown in Tab. 1, MFFN achieves the best performance in all three test datasets without extra training data, especially for the four indicators of $S_m$, $F_\beta$, $F_\beta^w$, $E_m$. Compared with model MGL [48], which introduced boundary detection as an auxiliary task with interactive learning and graph neural network, it is obvious that our method has superior performance by capturing boundary information and separating background simultaneously. Compared




(a) PR curves of the proposed MFFN and recent SOTA algorithms over CHAMELEON, COD10K and NC4K.




(b) $F_\beta$ curves of the proposed MFFN and recent SOTA algorithms over CHAMELEON, COD10K and NC4K.

Figure 5: Results of PR and $F_\beta$ curves. Red line represents our proposed MFFN.

with ZoomNet [30] which achieved the second best performance, our model improves $S_m$ by 0.8%, $F_\beta^w$ by 1.6%, $F_\beta$ by 1.6% and $E_m$ by 0.6% in the COD10K test set. Similarly, in the NC4K dataset test results, $S_m$ is improved by 0.3%, $F_\beta^w$ is improved by 0.7%, $F_\beta$ is improved by 0.9% and $E_m$ is improved by 0.3%. We draw precision-recall (PR) curve and $F_\beta$ curve. As shown in Fig. 5a and Fig. 5b, the PR curve of MFFN surrounds the previously proposed model, and the $F_\beta$ curve also presents an almost horizontal shape, which represents that MFFN has more accurate detection results. The visualization results for the different methods are shown in Fig. 6. We select 8 samples with obvious differences in object size, background interference and color for analysis. The comparison results show that our method can still obtain clear prediction boundaries and region shapes under the circumstance of highly blurred boundary and highly similar background. For model complexity, although we increase the input images with the multi-view design, our model still has the least number of parameters compared with single-view models, as shown in Tab. 2. This indicates that with multi-view design, we are able to apply a simpler encoder (*i.e.* instead of ViT [47]) with less complex strategies (*i.e.* instead of joint SOD and COD [19], or joint mutual graph learning [48]).

Table 2: Comparison of the number of parameters of our proposed MFFN and other SOTA models.

| Method | MFFN(Ours) | UGTR [47] | UJSC [19] | ZoomNet [30] | PfNet [29] | MGL-R [48] | SLSR [27] |
|---|---|---|---|---|---|---|---|
| Parameters | 36.554M | 48.868M | 217.982M | 32.382M | 46.498M | 63.595M | 50.935M |

### 4.3. Ablation Studies

In this section, we conduct ablation studies on the combination of different views, the mode to interact of multiple

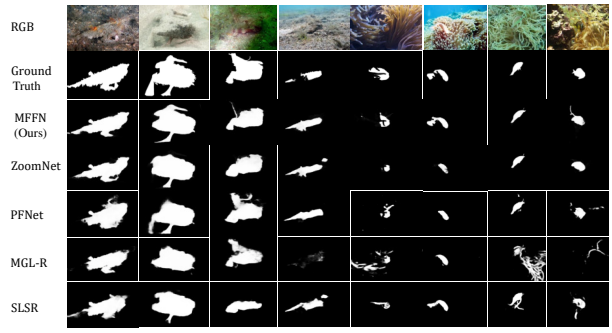



Figure 6: Visual comparisons of some latest research algorithms and our proposed MFFN in some typical images. We can find that the prediction results of MFFN have clearer boundary and region shape.

views and the CFU. Considering the representativeness of the dataset, different model design methods were used to evaluate the two large datasets COD10K and NC4K.

**The effect of different views.** The proposed model aims to complement and fuse the multi-view information to obtain precise and accurate boundary information and semantic correlation. We expand on the *distance view*, *perspective view*, *angle view* and the different combinations of them. The experimental results shown in the Tab. 3 and Tab. 5 reveal the significance of different views for feature capture, and we choose the best combination of views.

**The effect of two-stage attention in CAMV.** In our method, we introduce CAMV to interact with multi-view feature maps, enhancing the semantic expression of foreground and background. In order to better analyze the effect of two-stage attention on model performance improvement, we analyze the two stages respectively.

**The effect of CFU.** Considering the deficiency in context semantic association of feature maps after multi-view fusion, we design the CFU to further potential mine clues of the feature tensors obtained after CAMV. CFU module mainly includes channel expansion and interaction and context extraction modules. We perform an ablation analysis on the two main parts of the two CFU. Based on the results in the Tab. 4, it is discovered that obtaining potential context clues through CFU is critical.

## 5. Conclusion

In this paper, we propose the MFFN model by imitating the multi-view observation mechanism of biology, which makes the features captured from different views complement and interact with each other. MFFN makes up for

Table 3: Comparisons of different views and their combinations using different CAMV on COD10K. V-O: original view; V-F: far view; V-C: close view; V-A: angle view; V-P: perspective view.

| View | CAMV | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ |
|---|---|---|---|---|---|---|
| V-O | – | 0.797 | 0.649 | 0.063 | 0.704 | 0.875 |
| V-F | One-stage | 0.808 | 0.678 | 0.033 | 0.721 | 0.884 |
| V-C | One-stage | 0.841 | 0.736 | 0.029 | 0.772 | 0.909 |
| V-A | One-stage | 0.812 | 0.686 | 0.034 | 0.729 | 0.892 |
| V-F&C | One-stage | 0.844 | 0.735 | 0.028 | 0.769 | 0.907 |
| | Two-stage | 0.842 | 0.735 | 0.028 | 0.771 | 0.911 |
| V-A&F | One-stage | 0.807 | 0.675 | 0.034 | 0.717 | 0.886 |
| | Two-stage | 0.805 | 0.673 | 0.035 | 0.717 | 0.882 |
| V-C&P | One-stage | 0.827 | 0.714 | 0.036 | 0.759 | 0.901 |
| | Two-stage | 0.838 | 0.725 | 0.031 | 0.764 | 0.907 |
| V-A&P | One-stage | 0.796 | 0.649 | 0.042 | 0.689 | 0.881 |
| | Two-stage | 0.802 | 0.660 | 0.037 | 0.707 | 0.886 |
| V-A&C | One-stage | 0.835 | 0.727 | 0.03 | 0.766 | 0.906 |
| | Two-stage | 0.846 | 0.745 | 0.028 | 0.782 | 0.917 |

Table 4: Influence of CFU module on performance.

| Dataset | Method | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ |
|---|---|---|---|---|---|---|
| COD10K | no-CFU | 0.844 | 0.73 | 0.03 | 0.771 | 0.917 |
| | CFU | 0.846 | 0.745 | 0.028 | 0.782 | 0.917 |
| NC4K | no-CFU | 0.854 | 0.78 | 0.045 | 0.819 | 0.915 |
| | CFU | 0.856 | 0.791 | 0.042 | 0.827 | 0.915 |

Table 5: Comparisons of different views and their combinations using different CAMV on NC4K.

| View | CAMV | $S_m \uparrow$ | $F_\beta^\omega \uparrow$ | $MAE \downarrow$ | $F_\beta \uparrow$ | $E_m \uparrow$ |
|---|---|---|---|---|---|---|
| V-F | One-stage | 0.839 | 0.764 | 0.046 | 0.801 | 0.902 |
| V-C | One-stage | 0.854 | 0.788 | 0.043 | 0.823 | 0.912 |
| V-A | One-stage | 0.839 | 0.764 | 0.047 | 0.802 | 0.903 |
| V-F&C | One-stage | 0.86 | 0.793 | 0.042 | 0.824 | 0.914 |
| | Two-stage | 0.857 | 0.79 | 0.042 | 0.823 | 0.913 |
| V-A&F | One-stage | 0.834 | 0.757 | 0.048 | 0.795 | 0.9 |
| | Two-stage | 0.833 | 0.755 | 0.049 | 0.795 | 0.9 |
| V-C&P | One-stage | 0.843 | 0.774 | 0.049 | 0.806 | 0.897 |
| | Two-stage | 0.852 | 0.782 | 0.046 | 0.817 | 0.909 |
| V-A&P | One-stage | 0.821 | 0.742 | 0.054 | 0.780 | 0.886 |
| | Two-stage | 0.835 | 0.753 | 0.050 | 0.792 | 0.899 |
| V-A&C | One-stage | 0.845 | 0.774 | 0.047 | 0.812 | 0.906 |
| | Two-stage | 0.856 | 0.791 | 0.042 | 0.827 | 0.915 |

the omission of features in fixed view observation. Firstly, we obtain more compact features through multi-view attentional interaction design, which enhances the semantic representation ability of the feature maps to the object region and boundary, and well integrates the multi-view semantic information. In addition, the context association information of feature tensor, which is implied in the channel dimension, is further mined by the CFU. A large number of experimental results verify the high performance of this method in COD task, which is superior to the previous method. MFFN shows SOTA results in the COD task and is equally good in the SOD task, but our multi-view design concept still needs further development to achieve accurate detection performance in general object detection tasks.

## References


[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

[2] Ali Borji. What is a salient object? a dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing*, 24(2):742–756, 2015.

[3] R. C. Duarte, Aav Flores, and M. Stevens. Camouflage through colour change: mechanisms, adaptive value and ecological significance. *Philosophical Transactions of the Royal Society of London*, 372(1724):75–92, 2017.

[4] Dengping Fan, Gepeng Ji, Tao Zhou, Geng Chen, Huazhong Fu, Shen Jianbing, and Ling Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, October 2020.

[5] Dengping Fan, Ge-Peng Ji, Guolei Sun, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. Camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[6] Deng-Ping Fan, Ming-Ming Cheng, Yun Liu, Tao Li, and Ali Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, October 2017.

[7] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, July 2018.

[8] Jun Fu, Jing Liu, Haijie Tian, Yong Li, Yongjun Bao, Zhiwei Fang, and Hanqing Lu. Dual attention network for scene segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Fuentes, Yoon, Kim, SC, Park, and DS. A robust deep-learning-based detector for real-time tomato plant diseases and pests recognition. *SENSORS-BASEL*, 17:582–596, 2019.

[10] Shanghua Gao, Yongqiang Tan, Mingming Cheng, Chengze Lu, Yunpeng Chen, and Shuicheng Yan. Highly efficient salient object detection with 100k parameters. In Andrea Vedaldi, Horst Bischof, Thomas Brox, and Jan-Michael Frahm, editors, *Proceedings of the European Conference on Computer Vision (ECCV)*, August 2020.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[12] M. H. Hesamian, Wenjing Jia, Xiangjian He, and Paul Kennedy. Deep learning techniques for medical image segmentation: Achievements and challenges. *Journal of Digital Imaging*, 32:582–596, 2019.

[13] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip H. S. Torr. Deeply supervised salient object detection with short connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(4):815–828, 2019.

[14] Ge-Peng Ji, Deng-Ping Fan, Yu-Cheng Chou, Dengxin Dai, Alexander Liniger, and Luc Van Gool. Deep gradient learning for efficient camouflaged object detection. *arXiv preprint arXiv:2205.12853*, 2022.

[15] Huaizu Jiang, Jingdong Wang, Zejian Yuan, Yang Wu, Nanning Zheng, and Shipeng Li. Salient object detection: A discriminative regional feature integration approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[16] T. Kolda and B. Bader. Tensor decompositions and applications. *Siam Review*, 51:455–500, 2009.

[17] Trung-Nghia Le, Tam V. Nguyen, Zhongliang Nie, Minh-Triet Tran, and Akihiro Sugimoto. Anabranch network for camouflaged object segmentation. *Computer Vision and Image Understanding*, 184:45–56, 2019.

[18] Gayoung Lee, Yu-Wing Tai, and Junmo Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[19] Aixuan Li, Jing Zhang, Yunqiu Lv, Bowen Liu, Tong Zhang, and Yuchao Dai. Uncertainty-aware joint salient object and camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[20] Guanbin Li and Yizhou Yu. Visual saliency detection based on multiscale deep cnn features. *IEEE Transactions on Image Processing*, 25(11):5012–5024, 2016.

[21] Yin Li, Xiaodi Hou, Christof Koch, James M. Rehg, and Alan L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[22] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[23] Xiao Lin, Zhi-Jie Wang, Lizhuang Ma, and Xiabao Wu. Saliency detection via multi-scale global cues. *IEEE Transactions on Multimedia*, 21(7):1646–1659, 2019.

[24] Jiawei Liu, Jing Zhang, and Nick Barnes. Modeling aleatoric uncertainty for camouflaged object detection. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, January 2022.

[25] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[26] Nian Liu, Junwei Han, and Ming-Hsuan Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[27] Yunqiu Lv, Jing Zhang, Yuchao Dai, Aixuan Li, Bowen Liu, Nick Barnes, and Deng-Ping Fan. Simultaneously localize, segment and rank the camouflaged objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[28] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

[29] Haiyang Mei, Ge-Peng Ji, Ziqi Wei, Xin Yang, Xiaopeng Wei, and Deng-Ping Fan. Camouflaged object segmentation with distraction mining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8772–8781, June 2021.

[30] Youwei Pang, Xiaoqi Zhao, Tian-Zhu Xiang, Lihe Zhang, and Huchuan 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 (CVPR)*, June 2022.

[31] Youwei Pang, Xiaoqi Zhao, Lihe Zhang, and Huchuan Lu. Multi-scale interactive network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[32] Jongchan Park, Sanghyun Woo, Joon-Young Lee, and In So Kweon. Bam: Bottleneck attention module. In *British Machine Vision Conference (BMVC)*, September 2018.

[33] Jialun Pei, Tianyang Cheng, Deng-Ping Fan, He Tang, Chuanbo Chen, and Luc Van Gool. Osformer: One-stage camouflaged instance segmentation with transformers. In *Proceedings of the European Conference on Computer Vision (ECCV)*. Springer, 2022.

[34] Federico Perazzi, Philipp Krähenbühl, Yael Pritch, and Alexander Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2012.

[35] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. Basnet: Boundary-aware salient object detection. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, October 2015.

[37] P Skurowski, H Abdulameer, J Baszczyk, T Depta, A Kornacki, and P Kozie. Animal camouflage analysis: Chameleon database. In *Unpublished Manuscript*, 2018.

[38] Martin Stevens and Sami Merilaita. Animal camouflage: current issues and new perspectives. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1516):423–427, 2009.

[39] Jinming Su, Jia Li, Yu Zhang, Changqun Xia, and Yonghong Tian. Selectivity or invariance: Boundary-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[40] Yujia Sun, Geng Chen, Tao Zhou, Yi Zhang, and Nian Liu. Context-aware cross-level fusion network for camouflaged object detection. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, August 2021.

[41] Tom Troscianko, Benton Christopher P, Lovell P. George, Tolhurst David J, and Pizlo Zygmunt. Camouflage and visual perception. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 2009.

[42] Qilong Wang, Banggu Wu, Pengfei Zhu, Peihua Li, Wangmeng Zuo, and Qinghua Hu. Eca-net: Efficient channel attention for deep convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[43] Jun Wei, Shuhui Wang, and Qingming Huang. F3net: Fusion, feedback and focus for salient object detection. *AAAI Conference on Artificial Intelligence (AAAI)*, February 2020.

[44] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[45] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[46] Qiong Yan, Li Xu, Jianping Shi, and Jiaya Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[47] Fan Yang, Qiang Zhai, Xin Li, Rui Huang, Ao Luo, Hong Cheng, and Deng-Ping Fan. Uncertainty-guided transformer reasoning for camouflaged object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2021.

[48] Qiang Zhai, Xin Li, Fan Yang, Chenglizhao Chen, Hong Cheng, and Deng-Ping Fan. Mutual graph learning for camouflaged object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2021.

[49] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

[50] Jing Zhang, Xin Yu, Aixuan Li, Peipei Song, Bowen Liu, and Yuchao Dai. Weakly-supervised salient object detection via scribble annotations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[51] Xiaoning Zhang, Tiantian Wang, Jinqing Qi, Huchuan Lu, and Gang Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[52] Jiaxing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. Egnet: Edge guidance network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[53] Huajun Zhou, Xiaohua Xie, Jian-Huang Lai, Zixuan Chen, and Lingxiao Yang. Interactive two-stream decoder for accurate and fast saliency detection. In *Proceedings of the*

*IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

[54] Yuan Zhou, Ailing Mao, Shuwei Huo, Jianjun Lei, and Sun-Yuan Kung. Salient object detection via fuzzy theory and object-level enhancement. *IEEE Transactions on Multimedia*, 21(1):74–85, 2019.

[55] Hongwei Zhu, Peng Li, Haoran Xie, Xuefeng Yan, Dong Liang, Dapeng Chen, Mingqiang Wei, and Jing Qin. I can find you! boundary-guided separated attention network for camouflaged object detection. In *AAAI Conference on Artificial Intelligence (AAAI)*, February 2022.