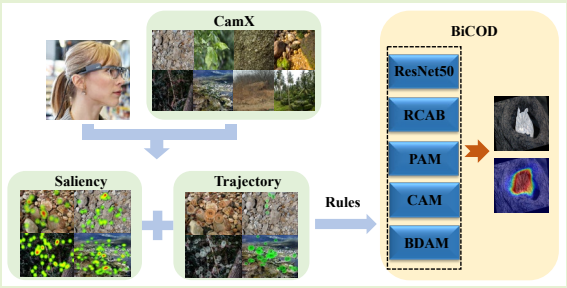
This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3343917

BiCOD: A Camouflaged Object Detection Method Directed by Cognitive Attention

Lianrui Xu, Xiong You*, Fenli Jia and Kangyu Liu

***Abstract*—Camouflaged object detection (COD) is a typical application of deep-coupled unmanned platform combat support which aims to detect objects that are highly similar to the background in terms of structure, details, and texture while improving the efficiency and accuracy of detecting camouflaged objects. Existing COD methods are built upon extraction and segmentation of image features and lack of theoretical interpretability. In this paper, the task of COD was revisited and analyzed. From the perspective of cognition, the cognitive laws of camouflaged objects were assessed through eye movement experiments to form an entire cognitive process, which serves as a guide for designing COD methods. Feature extraction, position attention, and channel attention modules were utilized as the basic framework. The residual-in-residual module was introduced to improve the accuracy of feature learning and transmission. Then a bidirectional attention module was added to guide the feedforward and feedback of attention features. And a closed loop was formed to achieve efficient feature transmission and use. As a result, the performance of our method BiCOD was promoted. In addition, a COD dataset containing both natural and artificial camouflage objects was compiled to evaluate the generalization ability of the camouflaged object recognition algorithm. The experiment results showed that BiCOD achieved an advanced level in quantitative results and visual comparisons in general, and the effectiveness and accuracy of the method in different environments were verified.**



***Index Terms*—Cognition, Attention, Camouflaged Object Detection**

I. Introduction

Humans have become curious about their attention since the beginning of cognitive activity. The brain has limited cognitive resources. However, the environment always present more information than the brain receives. Memory will always contain more features than the brain can process. And the brain will always provide more behavioral choices than people can perform. Therefore, attention has been very instrumental in human evolution and survival. From the perspective of cognitive psychology, attention focuses on the concentration and focus of mental activities that are selective, transferable, and easily deconstructible [1]. From the perspective of computer vision, the widely used attention mechanism differs from attention in terms of cognitive processes. It is a selective mechanism that has the processing power to allocate limited information. By constructing a real neural network model, attention behavior processed as soft, hard, or self-attention [2].

Research on attention continues to be conducted because attention can allocate limited information processing resources and provide perceptual selection ability. Visual attention helps humans select the most important and

Lianrui Xu is with Institute of Geospatial Information, Strategic Support Force Information Engineering University, Zheng Zhou 450052, China. (e-mail: Xulianrui124@163.com)
Xiong You is with Institute of Geospatial Information, Strategic Support Force Information Engineering University, Zheng Zhou 450052, China. (e-mail: youarexiong@163.com)
Fenli Jia is with Strategic Support Force Information Engineering University. (e-mail: fljia_ieu@126.com)
Kangyu Liu is with Department of Applied Psychology, Henan College of Science Technology and Communication，Kai Feng 475000.(e-mail: liukangyu0826@126.com)

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3343917

relevant information from a large amount of visual information. Bottom-up and top-down attention mechanisms jointly guide the visual attention behavior [3], as depicted in Fig. 1. Due to the importance of attention in efficient cognition and reducing cognitive burden, many scholars have simulated and designed human attention models from different perspectives and widely applied them in feature extraction, object recognition, algorithm evaluation, etc. [4][5][6]. With the continuous development of sensors and their application technologies, unmanned platforms—as the main platforms for sensor data acquisition and processing—face the same challenges as humans. The efficient extraction of most task-related features and identification of the most important objects from a large amount of data are difficult problems to solve at present. Visual attention mechanisms are of great significance in addressing such challenges.

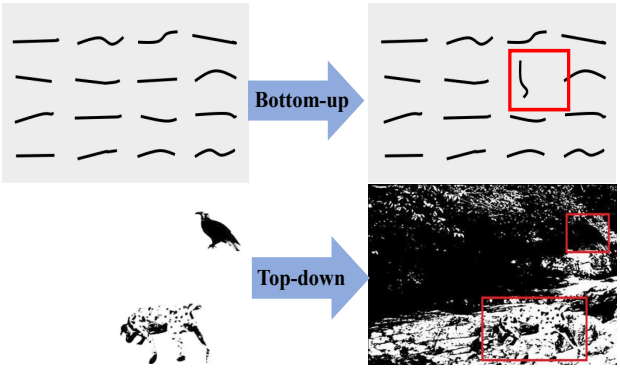


Fig. 1. Examples of bottom-up and top-down attention.

COD is a typical application of unmanned platforms for extracting the most relevant features and identifying the most important objects that are closely related to the cognitive attention process. Camouflage can be divided into natural and artificial camouflage. Its main feature is that the camouflaged object is highly similar to the background in terms of color, texture, and shape. Moreover, the visual recognition of the boundary and surrounding environment is poor, which poses a significant challenge to the saliency recognition method [7]. Although thermal imaging technology can achieve a high detection rate for camouflaged objects, its high cost limits its popularity and application. Visual COD based on visible light has advantages in terms of realization cost and reliability. Therefore, it remains the focus of many surveys. Currently, COD methods are mainly focused on computer vision. Bayes method, graph model, frequency domain analysis, pattern recognition, deep learning and other methods were used to extract camouflaged object features and design a significance detection module based on feature similarity theory, guide search and other models. Moreover, tests have been conducted on datasets such as COD10K [9], NC4K [10], and Camo [11] to improve the generalization of the detection methods, as depicted in Fig. 2. However, the existing methods for COD focus on designing detection modules and improving the detection rate of algorithms that lack explanation of the action mechanism of attention in the detection process, as well as analysis of how attention guides the method design. The improvement in the calculation indicators represents the actual recognition effect, and the process lacks interpretability. Based on the cognitive attention process, this paper provides analyses of cognitive experiment records, summarizes the rules, guides the design and module construction of COD methods, addresses the functional organization of attention at the conceptual and computational levels, and cross-validates it through empirical research in cognitive science and computer science.

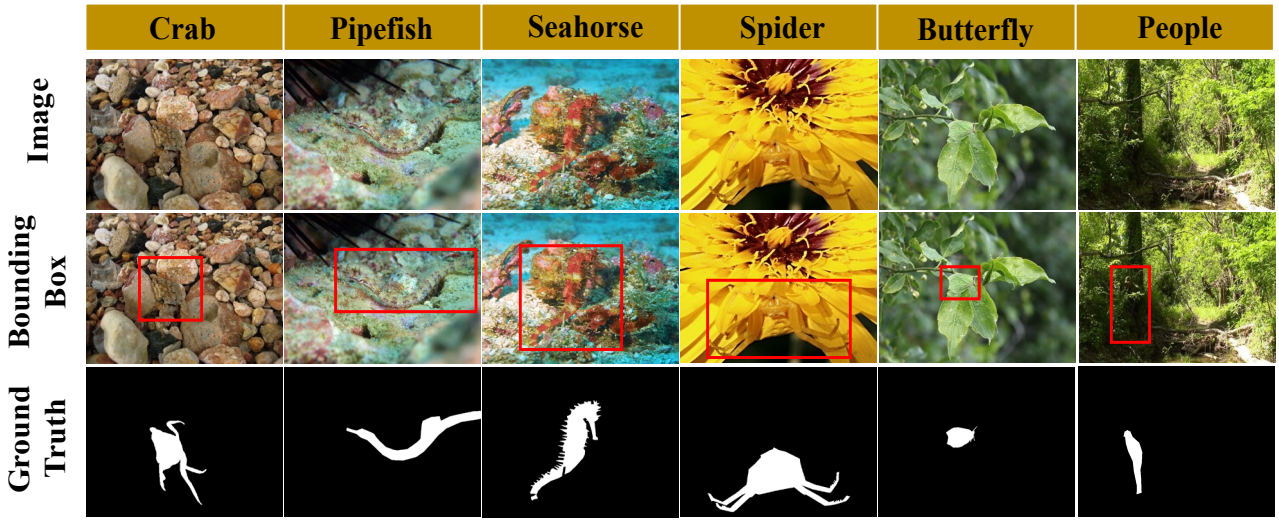


Fig. 2. Examples of camouflaged objects.

The main contributions of this paper are as follows. (1) The process of cognitive attention is analyzed in combination with an eye movement experiment of COD, and the rules of cognitive attention are summarized. (2) It provides an analysis of the characteristics of camouflaged objects, incorporates the cognitive attention law, introduces the residual into the residual module, presents the design of a bidirectional attention module, and provides a COD method called BiCOD. (3) It presents a new dataset CamX for cognitive attention-based COD tasks, including artificial camouflage and natural camouflage that effectively verifies our method.

II. Related Work

Currently, most attention models are inspired directly or indirectly by cognitive concepts and reality. Treue and Martánnez Trujillo et al. [12] proposed the feature similarity theory of attention based on red-green stimulation experiments in cognitive physiology and believed that the intensity of attention gain and inhibition depends on the contrast between the object and the background. Treisman and Gelade et al. [13] put forward the feature integration theory, which emphasizes the importance of visual features and combines them with practical tasks. Based on the feature integration theory, Koch and Ullman [14] proposed using the feedforward model to associate features and introduce the concept of saliency maps to represent the concern degree of spatial location. The saliency model described by Itti [15] uses linear filtering to perform shallow feature extraction of color, brightness, and direction which obtains feature maps by calculating the center-periphery difference and normalization. He utilizes cross-scale merging to transform feature maps into saliency maps and employs return suppression to complete the focus shift. In a computational study of attention models, the feature integration theory proposed by Treisman [13], guide search utilized by Wolfe [16], and closed-loop attention model [17] all explored the theoretical basis of computational attention systems in depth. And they have been successfully applied in object recognition, image matching, robot navigation and other fields [18][19].

Early works on COD can be traced back to 1988. In essence, COD is the processing, labeling and classification of image pixels. Single-modality image segmentation is easily limited by information diversity. Hong et al. [20][21] developed multi-modality learning frameworks and integrated different network architectures to enhance the generalization capability of image processing in different datasets. It can reduce the impact of image differences on pixel segmentation algorithms. Meanwhile, the unsupervised method based on few-shot learning can effectively reduce image noise and dependence on labels, and still maintain the accuracy of pixel classification under the condition of missing samples, which is suitable for different types of downstream tasks including COD[22]. The self-supervised method based on the foundation model can give full play to the advantages of computing power by generating a large parameter model, and improve the accuracy of image processing and segmentation[23]. Based on the processing of image intensity values. Tankus et al. [24] proposed a non-edge area-of-interest detection mechanism to detect and identify artificially camouflaged objects in both natural and combat environments. Subsequently, methods for COD using visual features such as color, texture, optical flow, and convexity have been proposed [25][26]. However, when the contrast between the foreground and background is low and the structure of the camouflaged object is highly similar to that of the surrounding environment object, these methods have problems, such as low feature extraction efficiency, poor algorithm generalization, and low object recognition accuracy. In addition, COD method and some technologies are also widely used in other research fields such as medical image classification and food image recognition[27][28]. The combination of semi-supervised and weakly supervised methods connected with COD also has great development potential [29].

With the rise of deep learning methods, the field of COD has developed significantly. The COD method based on deep learning can extract the deep features of a camouflaged object in an image and the interrelation of the corresponding pixels through multilayer image convolution and other operations, as well as continuously model the feature structure of the camouflaged object independently to improve COD accuracy. Fan et al. [9] inspired by animal hunting, proposed a combination of a texture-enhanced module, neighbor connection decoder, and group-reversal attention (GRA). The COD search network SINet utilizes GRA and sophistication of the network was verified on a large dataset COD10K. Zhong et al. [30] introduced the frequency domain to improve the detection accuracy of camouflaged objects and constructed a frequency-domain enhancement module. Then they estimated the frequency-domain transform gradient of pixels based on the discrete cosine transform, fused the RGB domain and frequency-domain features with the method of feature alignment, and

processed the camouflage features through the higher-order relation module. Li et al. [31] proposed a joint salient object and COD model aimed at the uncertainties generated during the labeling of camouflaged objects and used a full convolution discriminator to evaluate the confidence of the model prediction results. Simultaneously, the confidence was modeled explicitly based on adversarial generation training to guide the prediction results more effectively.

In deep learning for camouflaged objects, many modules use attention mechanisms. The introduction of the attention mechanism in the computer vision can capture the global context information more fully [32]. Lv et al. [10] employed the ResNet50 [33] framework to extract image and camouflaged object deep features, associated pixel information based on channel attention and position attention modules, and finally used a residual channel attention module to achieve integrated transmission of extracted and analyzed features. The D^2CNet proposed by Wang et al. [34] is based on a partial decoder component, with global and residual attention added to enhance the feature extraction process and U-Net combined to refine deep features to transmit the feature information of the camouflaged object efficiently and accurately. Sun et al. proposed a context-aware cross-level fusion network (C^2F-Net) [35]. They used multiscale and multichannel attention to guide the aggregation of cross-level features and made full use of global and local information to improve the detection performance of multiscale objects.

In summary, the existing COD methods are mostly based on the image itself or computer vision. For example, the object detection framework based on Transformer adopts encoder-decoder architecture and obtains rich feature through Multi-Head Attention. It can learn the long-distance dependence relationship and improve the success rate of object detection. The mainstream semantic segmentation methods based on supervised learning use manually annotated pixels to provide a large number of detailed semantic information and local features, which realize feature extraction and transmission through the Full Convolutional Neural Network, Recurrent Neural Network, Generative Adversarial Network and other ways. Weakly supervised learning lacking annotation information can construct the relationship between pixels and image labels by means of multiple-instance learning, and obtain the precise features required for semantic segmentation efficiently. However, these methods ignore the importance and interpretability of cognitive guidance to a certain extent. Methods based on deep learning are very sensitive to parameters such as learning rate and weight decay, that are easy to lose local feature information. Therefore, we proposed BiCOD from the cognitive perspective, extracted attention rules through cognitive experiment. Based on the above work, we designed BiCOD framework which greatly enhanced the interpretability of the method at the cognitive level. Meanwhile, the Bidirectional Attention Module(BAM) is designed to simulate the feedforward, feedback and global regulation of features between different modules. The excellent performance of BiCOD is demonstrated by the cross-validation of cognitive experiments and computer vision.

III. Proposed Methods

A Cognitive Law Extraction

In the cognitive experiment, a Tobii Pro Glasses 2 eye tracker was used as the data-acquisition device. Based on the principle of telemetry corneal reflex, data such as individual eye movement trajectories, line-of-sight changes, and eye movement states were recorded synchronously, and visual processing characteristics were analyzed. The sampling frequency was 50 Hz, and the scene camera resolution was 1920 × 1080 pixels, as depicted in Fig. 3.

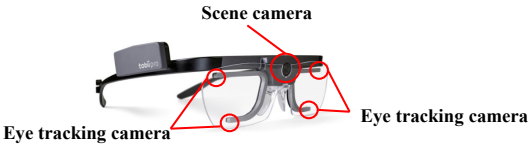


Fig. 3. Tobii Pro Glasses 2 eye tracker.

Participants： In this experiment, 28 participants with different experience in COD were selected. The detailed information is depicted in Tab. 1. All participants were in good physical condition. They had normal vision or corrected vision and no color blindness or weakness. At the end of the experiment, 2 participants were released due to insufficient sample rate of eye movement data. As a result, 26 participants(14 male and 12 female) successfully completed the experiment. The 26 participants ranged in age from 19 to 36, with an average age of 26, 10 with experience and 16 with no experience.

In the cognitive experiment, 16 camouflaged object images from CamX were randomly arranged, of which 8 were naturally camouflaged and 8 were artificially camouflaged, as depicted in Fig. 4. 26 participants were selected for COD cognitive experiments. The positions and durations of the fixation points on different images were recorded using an eye tracker. The rules were

comprehensively analyzed and summarized to guide the design of COD method.

TABLE I

INFORMATION STATISTICS FOR PARTICIPANTS

Number	Age	Sex	Experience
1	25	Male	Yes
2	26	Male	No
3	24	Male	No
4	24	Female	Yes
5	21	Female	Yes
6	19	Male	No
7	29	Female	No
8	23	Female	No
9	34	Male	Yes
10	26	Male	No
11	25	Female	No
12	36	Male	No
13	27	Female	No
14	25	Female	Yes
15	27	Male	No
16	34	Male	Yes
17	30	Male	No
18	20	Female	No
19	34	Female	No
20	22	Male	Yes
21	28	Male	Yes
22	22	Female	Yes
23	21	Male	No
24	24	Female	No
25	26	Male	Yes
26	24	Female	No

Fig. 5 shows saliency maps of eye tracking results in COD experiment . The superposition of fusion map for participant 3, participant 14, and participant 21 is obtained by eliminating the overlay of noise data with a velocity threshold<30 degrees/second using Tobii I-VT algorithm, which can reflects participants' general film fixation areas. The color depth in the saliency map is proportional to the fixation time of the participants. The analysis of the saliency map of COD revealed that objects that have lower camouflaged level and close to the image center were more likely to attract the attention of participants. Simultaneously, the fixation time of objects was longer, the attention degree of the background was lower. For objects with high camouflaged levels, the dispersion degree of the fixation area of objects was greater.

Additionally, when comparing participants in terms of COD experience, the fixation points and search scope of the experienced participants were less than those of the inexperienced participants, indicating that they could identify the camouflaged objects by checking fewer suspected objects and could quickly determine the camouflaged objects by increasing the fixation time. The gender of participants had no significant impact on the detection results.

For different positions, the fixation time was reduced. And the fixation area was closer to the area with bright colors and high contrast in the center of the image, which was well reflected in the fusion saliency maps of participants.

Further analysis in the fixation trajectory diagrams of representative participants was performed in Fig. 6, where the circle size represents the fixation time. Numbers in circles represent the gaze order. When most participants detected camouflaged objects, they firstly scanned the central area of the image to form overall feature perception. When no object was found, the image was divided into potential areas and searched in different areas in a certain order to complete the search transfer from global to local. When a suspicious object was found, the similarity between the edge and background of the suspicious object was compared to confirm the camouflaged object and complete the identification and detection. It can be seen that COD from the cognitive perspective will focus on the image at different levels to achieve feature extraction and satisfy the global–local–global search and comparison processes. Simultaneously, the feedforward process of attention includes a summary and feedback of the feature information. Therefore, we used this cognitive law to guide the COD network design.

Fig. 4. Images in COD cognitive experiment.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3343917

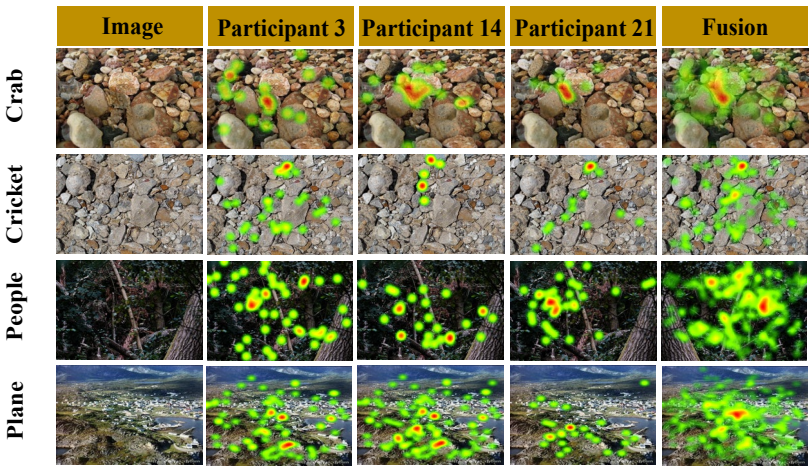


Fig. 5. Saliency and fusion maps of eye tracking results for participant 3,14,21.

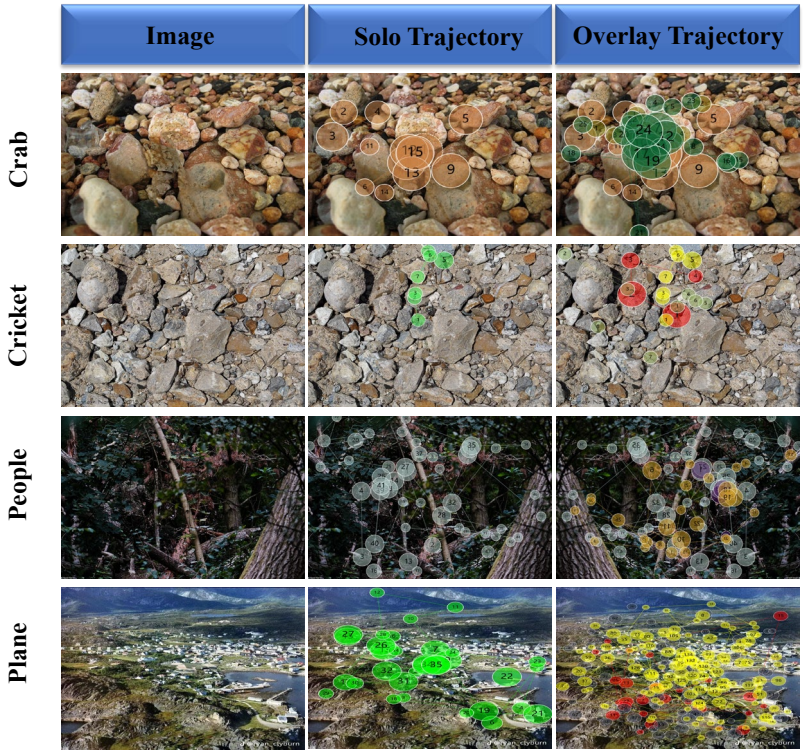


Fig. 6. Eye movement trajectories in COD cognitive experiments.

B. Network Design

According to the attention rules summarized in cognitive experiments, we firstly selected ResNet50 as the backbone to design BiCOD, as depicted in Fig. 7. After the original image was obtained, it was passed into the backbone network, and a feature map of different channel was obtained through convolution. In addition, we designed a convolutional layer that adjusts the feature size and channel to maintain an output of the same size as those of different layer in backbone network.

In the encoder, features obtained from the backbone network are extracted through bilinear up-sample with scale parameters of 8, 4, 2, and 0.5, the features of which are input into the residual attention module. The accuracy of feature extraction is improved through operations such as convolution, pooling, convolution, and channel attention [36]. Subsequently, the features of Residual Channel Attention Block(RCAB) [37] are input into the two-channel attention module, the channel attention representations of $B \times C \times C$ and $B \times (H \times W) \times (H \times W)$ and location attention representations are obtained. The attention is output with the input features to obtain the fusion-discriminant features.

The decoder design was based on the Dense Atrous

Spatial Pyramid Pooling module[38], which accepts multiscale discriminant features, obtains the region that needs to be discriminated, compares it with marked truth value, and introduces a loss function to quantify different branches. In this paper, to strengthen the cognitive characteristics of detecting the global position of camouflaged objects, we designed a bidirectional attention module to guide the propagation and learning of attention. Firstly, the fused discriminant feature F is obtained. Then, the reverse attention is obtained based on $E-F$ and multiplied the backbone feature, which is transmitted back to the encoder network by a short skip connection. The attention bias of the residual attention module and the double attention module is guided, saliency is superimposed with the original image to obtain the area of the image that requires attention, and COD is realized.

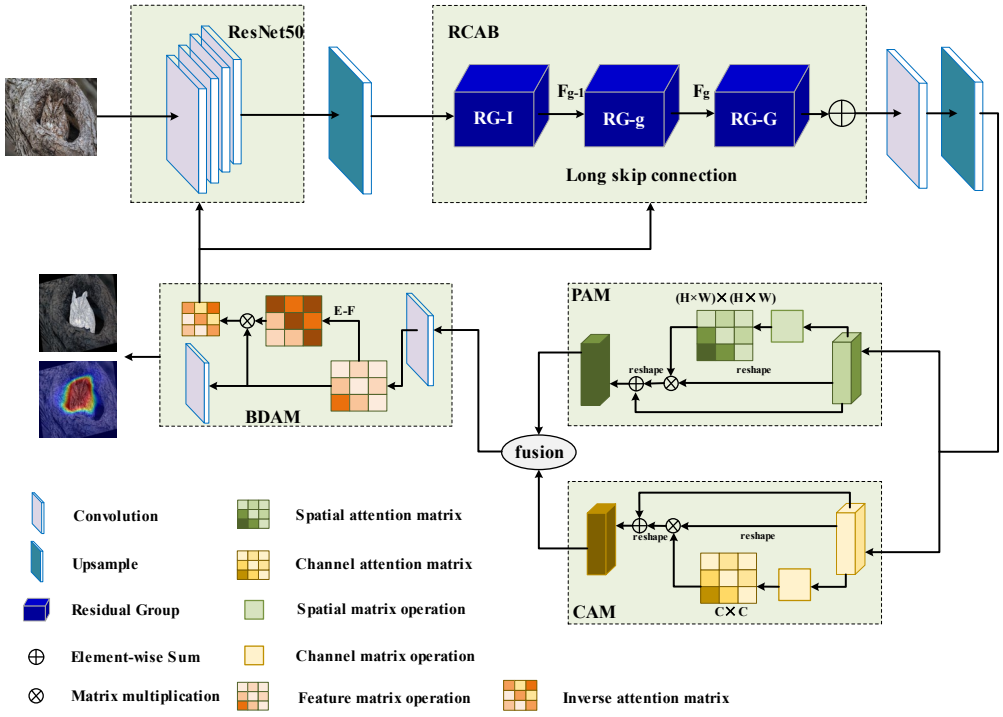


Fig. 7. Framework of BiCOD method.

C. Feature Extraction Module(FEM)

Owing to the hidden correlation between camouflaged objects and surrounding environment pixels, a multilayer network must be used to extract deep features effectively. To avoid gradient disappearance, ResNet50 is used as the backbone network for the feature extraction module. Given input image I, the convolution block and identity block are used to change the dimension and depth of the network, extract multilevel features of different scales from bottom-up, and obtain four feature mappings of scales s_1, s_2, s_3, s_4, with channel numbers of 64, 128, 256, 512, respectively. After convolution, batch normalization was used, and ReLU was selected as the activate function to move to the residual part and improve the accuracy of feature extraction.

D. Position And Channel Attention Modules

The position attention module(PAM) mainly encodes rich contextual information to form local features, and the adaptive aggregation features represent information, which is an important embody of global and local cognitive rules. Given a local feature $A \in R^{C\times H\times W}$, new features $B, C, and\ D$ are formed by convolution; B and C are transposed together; and the spatial attention map $S \in R^{N\times N}$ is calculated using SoftmMax function, where $N = H \times W$. The influence of position i on position j can be expressed as S_{ji}, as in (1):

$$S_{ji} = \frac{\exp(B_i \cdot C_j)}{\sum_{i=1}^{N} \exp(B_i \cdot C_j)} \tag{1}$$

Simultaneously, by multiplying feature D multiply by spatial attention S_{ji}, considering the influence of scale parameter α, the weighted sum of the associated feature and the original feature of each position pixel E_j is finally output, and similar features enhance each other to improve the classification consistency. E_j is given by (2):

$$E_j = \alpha \sum_{i=1}^{N} (S_{ji} D_i) + A_j \tag{2}$$

The channel attention module(CAM) mainly uses the interdependence between channels to calculate the channel attention $X \in R^{C\times C}$ directly from the original feature $A \in R^{C\times N}$, where $N = H \times W$. The transpose of A is multiplied times A, and the influence of channel i on channel j can be expressed as X_{ji}, as in (3):

$$X_{ji} = \frac{\exp(A_i \cdot A_j)}{\sum_{i=1}^{C} \exp(A_i \cdot A_j)} \tag{3}$$

Simultaneously, X is multiplied times the transpose of A. Considering the influence of the scale parameter β, the final output E_j can be expressed as in (4):

$$E_j = \beta \sum_{i=1}^{C} (X_{ji} A_i) + A_j \tag{4}$$

E. Residual Channel Attention Module

The residual attention module firstly extracts shallow features from the input using convolution and then inputs the extracted shallow features into the residual in residual (RIR) module for further feature extraction [37]. Specifically, each RIR includes residual groups (RGs) and long skip connections (LSCs). The RG in group g can be expressed as in (5):

$$F_g = H_g(H_{g-1}(\cdots H_1(F_0)\cdots)) \tag{5}$$

where F_g denotes the output of the gth RG, H_g is the gth RG function, and F_0 is the original input feature. Because simply stacking RG modules cannot improve the feature extraction accuracy, LSCs are introduced to stabilize the network while learning through residual error, ignoring too much low level information and focusing on high level information. Therefore, the high level feature F_{DF} can be expressed as in (6):

$$F_{DF} = F_0 + W_{LSC} = F_0 + W_{LSC} H_g(H_{g-1}(\cdots H_1(F_0)\cdots)) \tag{6}$$

where W_{LSC} is the weight of the convolution layer at the end of RIR. After obtaining the high level features, an additional convolution was used to reconstruct high level features, and L1 loss function was optimized based on a random gradient descent to improve the accuracy of feature extraction by the residual attention module. The loss function can be expressed as in (7):

$$L(\Theta) = \frac{1}{N} \sum_{i=1}^{N} ||H_{RCAB}(I_{LR}^{i}) - I_{HR}^{i}|| \tag{7}$$

where H_{RCAN} is the residual channel attention function, I_{LR} and I_{HR} correspond to the input and output of RCAB, respectively.

F. Bidirectional Attention Module

The feature extraction module uses multilayer convolution to extract features from the original images, which is a type of positive attention process that mimics the cognitive mode from global to local, where features of different regions are cascaded and combined with each other to form high level features representing objects. However, this commonly used positive attention ignores the structural and detailed features. In addition, it does not consider the cognitive mode of the camouflaged object search from local to global. Therefore, this section proposes BAM that includes forward and reverse feature transfers. The attention module was designed from the perspective of cognitive processes, and the learned local features were fed back to achieve local-global feature transfer and improve the detection accuracy of camouflaged objects.

Firstly, we obtained feature maps $s_1, s_2, s_3, and\ s_4$ of different scales using ResNet50, divided these features into multiple patches according to different channels, aligned the features through convolution and up-sample operations, and used batch normalization to obtain the overall representation F of the features. Based on the reverse operation, the feedback features $E - F$ of the reverse attention were obtained by subtracting all matrices E and F with elements of 1. These features were retransmitted to the forward feature extraction module through a short connection to guide the extraction of the original features as weight coefficients, forming a bidirectional attention cycle.

IV. Experiment Results

A. Experimental Design

Datasets: We evaluated our methods using three datasets: COD10K, NC4K, and CamX. COD10K is currently the largest camouflaged object detection dataset, with 3040 training images and 2026 test images. NC4K, which includes 4121 images downloaded from the network, is also widely used in camouflaged object detection experiments. CamX is a new dataset proposed by us, which is mainly divided into two categories: natural camouflaged objects and artificial camouflaged objects, including 700 training images and 300 test images. Multiple types of camouflaged objects can effectively improve the feature extraction and generalization abilities of detection networks.

Implementation details: The COD experiment used an Intel(R) Xeon(R) Gold 5218 CPU as the processor and four GeForce RTX 2080 graphics cards to compute multi-batch graphics data. First, the training images were clipped to a unified size of 360 × 360, and then a group of 32 images were input into ResNet50 backbone for feature extraction.

TABLE II

EXPERIMENT SETTINGS

Hyperparameter	Details
Backbone	ResNet50
Input image size	360 × 360
Upsample scales	8,4,2,0.5
Optimizer	Adam
Learning rate	2.5e–5
Batch size	32

After the long skip connection of RCAB, up-sample was conducted. The high level features obtained were passed into PAM and CAM. Then, the fusion features were input into the BDAM for reverse attention calculation, and the feedback obtained was transmitted back to the backbone network and RCAB through a short skip connection to guide the effective extraction of subsequent camouflaged object features. A total of 100 rounds were set up. The optimal training result parameters were taken to verify the verification set.

Evaluation index: The detection and evaluation of camouflaged objects are generally based on binary segmentation. Widely used evaluation indicators include the mean absolute error [39], mean F-measure [40], mean E-measure [41], and S-measure [42]. They are expressed as $\mathcal{M}, F_\beta^{mean}, E_\xi^{mean}, and\ S_\alpha$. S_α mainly evaluates the structural similarity between the truth graph and prediction graph, as in (8):

$$S_\alpha = \alpha S_o + (1-\alpha) S_r \tag{8}$$

where S_o and S_r represent the object perception feature and regional observation feature, respectively, and α and o are the weights.

E_ξ^{mean} evaluates the global and local accuracy of COD results by comparing the difference between the prediction and truth graphs, as in (9):

$$E_\xi^{mean} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \Phi(C(x,y) - G(x,y)) \tag{9}$$

where W and H are the width and height of the input image, respectively; Φ is the enhanced consistency matrix; and C and G are the prediction and truth graphs, respectively.

F_β^{mean} mainly calculates the relationship between the accuracy rate P and recall rate R and is reflected in the form of the mean, as in (10):

$$F_\beta^{mean} = \frac{(\beta^2+1)PR}{\beta^2 P + R} \tag{10}$$

$\mathcal{M}$ calculates the average absolute error of each pixel, as in (11):

$$\mathcal{M} = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} (C(x,y) - G(x,y)) \tag{11}$$

Comparison method: Considering the effectiveness and generalization of COD methods, as well as the similarity between some saliency detection methods and COD methods. SINet, SINET-V2, LSR, and JSCOD were selected as comparison methods. The selected methods are highly innovative and widely influential COD methods developed in recent years. This finding has good comparative significance.

B. Comparison Results

Quantitative results: Tab. 3 shows the specific values of the different indicators of our method and SINet [9], SINET-V2 [43], LSR [10], and JSCOD [31] for the COD10K, NC4K, and CamX datasets. It can be seen from Tab. 3 that our method obtained good results for the general indicators of $S_\alpha, F_\beta^{mean}, E_\xi^{mean}, and\ \mathcal{M}$, although there are still deficiencies in a single index of individual data sets. However, in general, compared with SINet, SINeT-V2, LSR, and JSCOD, our method exhibited significantly improved COD performance. The training of our dataset also indicates that our network has better generalization and can achieve a higher detection rate for different types of camouflaged objects.

Visual comparisons: As depicted in Fig. 8, we selected different categories of camouflaged objects for detection and used saliency maps to highlight the parts that required attention. Compared with other methods, our method based on cognitive attention guidance can more accurately detect camouflaged objects with subtle differences in scale, type, contrast, etc.; achieve better suppression of background noise and unrelated objects; process more complex backgrounds; and have a strong anti-interference ability. Moreover, BiCOD can compare the camouflaged and surrounding objects to assess their similarities. The distinction of subtle differences between other objects and the camouflaged objects highlights the superior camouflaged object detection capabilities of the proposed approach.

C. Ablation Studies

To compare the effectiveness of the proposed bidirectional attention module, we uncoupled and combined different modules for validation on the largest camouflaged object dataset COD10K, and our proposed CamX dataset, including the FEM, PAM, CAM, RIR, and BAM. In the ablation experiments with different modules, we maintained the same training parameters as those in Section 4.2. The

TABLE III

RESULTS OF DIFFERENT COD METHODS ON COD10K, NC4K AND CamX

Method	COD10K				NC4K				CamX			
	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	M ↓	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	M ↓	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	M ↓
SINet	0.733	0.588	0.768	0.069	0.779	0.696	0.800	0.086	0.728	0.627	0.768	0.065
SINet-V2	0.815	0.887	0.680	0.031	0.810	0.795	0.874	0.053	0.740	0.642	0.779	0.056
LSR	0.760	0.658	0.831	0.045	0.797	0.758	0.854	0.061	0.736	0.635	0.767	0.058
JSCOD	0.817	0.726	0.892	0.035	0.804	0.782	0.861	0.057	0.733	0.630	0.775	0.061
Ours	**0.820**	**0.827**	**0.879**	**0.032**	**0.827**	**0.793**	**0.865**	**0.055**	**0.741**	**0.639**	**0.783**	**0.054**

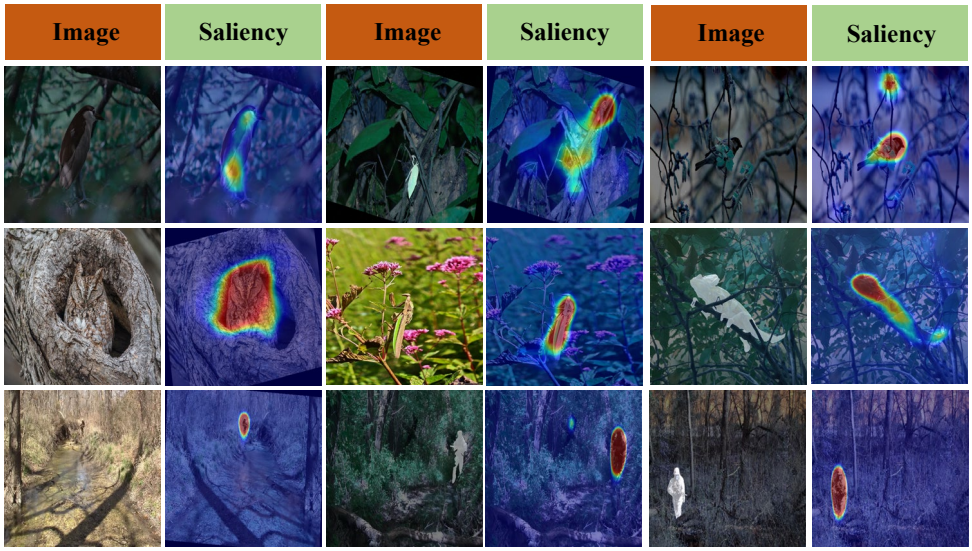


Fig. 8. Attention saliency maps of BiCOD method for different categories of camouflaged objects.

experimental results are listed in Tab. 4.

Effectiveness of FEM: Based on the results in rows 1 and 2 in Tab. 4, the effectiveness of ResNet50 on the feature extraction backbone network was better than that of ResNet101 when the location and channel attention modules were the same. The S-measure exhibits improvements of 0.009 and 0.017 on the COD10K and CamX datasets, respectively, indicating that the ResNet50 feature extraction network performs better in camouflaged object extraction and consumes fewer computing resources during training and reasoning.

Effectiveness of RIR: Analysis of the results rows 1 and 3 in Tab. 4 shows that under the same settings of FEM, PAM, and CAM, the COD network with the RIR module performs better in the test dataset, with relatively large improvements in the four indices because RG and LSC in RIR can appropriately ignore low level feature information. The extracted effective features were stably transferred, and the accuracy of the extracted features was improved. Comparison of the index E_ξ^{mean}, which reflects the difference between the prediction and truth images, demonstrates that the network with the RIR module achieves improvements of 0.003 and 0.008 in the two datasets, verifying the effectiveness of the RIR module in improving the detection accuracy of camouflaged objects.

Effectiveness of BAM: Analysis of the results rows 3 and 4 in Tab. 4, under the same settings of FEM, PAM, CAM, and RIR, the proposed BAM performs better in COD, especially in terms of the absolute mean error $\mathcal{M}$ of the pixels. This result shows that the BAM module can effectively guide feature transfer and migration and reduce the loss of feature transfer. Based on attention feedforward,

TABLE IV

RESULTS OF DIFFERENT COD METHODS ON COD10K AND CamX

Method	COD10K				CamX			
	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	M ↓	$S_\alpha \uparrow$	$F_\beta^{mean} \uparrow$	$E_\xi^{mean} \uparrow$	M ↓
FEM(50)+PAM+CAM	0.771	0.784	0.849	0.046	0.715	0.618	0.762	0.065
FEM(101)+PAM+CAM	0.762	0.776	0.845	0.049	0.698	0.604	0.759	0.070
FEM(50)+PAM+CAM+RIR	0.787	0.799	0.852	0.045	0.723	0.627	0.770	0.062
FEM(50)+PAM+CAM+BAM	0.791	0.803	0.858	0.042	0.729	0.631	0.776	0.058
FEM(50)+PAM+CAM+RIR+BAM	0.820	0.827	0.879	0.032	0.741	0.639	0.783	0.054

attention feedback is realized, multiscale representation of features is mined, invalid processing of feature information is suppressed, more attention is paid to the structure and details of features, a transmission closed loop is formed through forward and reverse attention, and the detection efficiency and accuracy of camouflaged objects are improved. The experimental results also verified the generalization of the BAM module in different datasets that can facilitate the high speed, efficient, and high quality transmission of features, which provides a solid foundation for the transfer learning of camouflaged object features.

V. CONCLUSION

Based on the COD task, we designed cognitive experiments and summarized and extracted cognitive rules from the cognitive COD experiment to guide the design of COD networks BiCOD. By combining the FEM, PAM, CAM, RIR, and the proposed BAM module to form an overall framework, different types of camouflaged objects can be detected and recognized. The experimental results show that BiCOD achieves advanced performance in general and can be verified using different datasets. In addition, we proposed a new COD dataset CamX, which includes both natural and artificial camouflaged objects, to evaluate the generalization ability of COD more effectively.

However, there are still some problems that are not fully considered in this paper. First, the generalization of the proposed BiCOD method must be improved. Because of the relatively low number of samples in the dataset itself, the bidirectional feature transmission of the BAM module could only learn a limited number of low level features and high level features, so the effect in the verification of some datasets was not significant. Secondly, this paper did not strictly control variables in the cognitive rule extraction experiment, but only collected and analyzed the eye movement data of the participants from the perspective of engineering implementation. Consequently, some psychological experiment methods should be used for reference to control variables strictly and effectively and collect the required data more accurately. Finally, the rules extracted from the cognitive experiments in this study comprise more qualitative analysis while lacking strong data support.

In the future application of COD, we should pay attention to the following improvements. Firstly, it is necessary to strengthen the research on the interpretability of cognitive architecture and theoretical basis. Describing qualitative cognitive laws as quantitative model parameters and broadening the scope of attention mechanism from the perspective of cognition is significant. Secondly, COD ought to combine incremental learning, reinforcement learning and other methods that based on the idea of continuous learning. It can improve the generalization of COD in different datasets, which has better capability in feature learning, memory and application, and can be further transferred to the application of unmanned platforms. Thirdly, the approach connected with large model is no longer capable of real-time environmental perception and camouflaged object detection. In the future, the computing architecture and data transmission process of COD should be improved to reduce the dependence on the large model calculation, so as to have the ability to independently detect the camouflaged objection and make decisions. In conclusion, the proposed BiCOD method can accurately and effectively detect multiple types of naturally and artificially camouflaged objects, exhibiting strong application prospects. In the future, the cognitive rule can be further transferred, the algorithm

and the performance in COD tasks can be improved.

VI. Acknowledgements

This research was supported in part by the Key Project of the National Natural Science Foundation of China (42130112), Research Project of Central Plains scholar You Xiong Scientist studio (2020), and Research Project of Strategic Support Force Information Engineering University (1064201). We thank the reviewers for their feedback on our manuscript.

References

[1] Ding J H, Zhang Q, Guo C Y. Cognitive psychology[M]. China Renmin University Press, 2010.

[2] Guo M H, Xu T X, Liu J J, et al. Attention mechanisms in computer vision: A survey[J]. Computational Visual Media, 2022, 8(3): 331-368.

[3] Borji A, Itti L. State-of-the-art in visual attention modeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 185-207.

[4] Lateef F, Kas M, Ruichek Y. Saliency heat-map as visual attention for autonomous driving using generative adversarial network (GAN)[J]. IEEE Transactions on Intelligent Transportation Systems, 2021, 23(6): 5360-5373.

[5] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[6] Rangrej S B, Srinidhi C L, Clark J J. Consistency driven sequential transformers attention model for partially observable scenes[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 2518-2527.

[7] amouflaged Object Detection Based on Deep Learning[J]. Journal of Frontiers of Computer Science and Technology,2022,16(12):2734-2751.

[8] Borji A, Itti L. State-of-the-art in visual attention modeling[J]. IEEE transactions on pattern analysis and machine intelligence, 2012, 35(1): 185-207.

[9] Fan D P, Ji G P, Sun G, et al. Camouflaged object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 2777-2787.

[10] Lv Y, Zhang J, Dai Y, et al. Simultaneously localize, segment and rank the camouflaged objects[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 11591-11601.

[11] Le T N, Nguyen T V, Nie Z, et al. Anabranch network for camouflaged object segmentation[J]. Computer vision and image understanding, 2019, 184: 45-56.

[12] Treue S, Trujillo J C M. Feature-based attention influences motion processing gain in macaque visual cortex[J]. Nature, 1999, 399(6736): 575-579.

[13] Treisman A M, Gelade G. A feature-integration theory of attention[J]. Cognitive psychology, 1980, 12(1): 97-136.

[14] Koch C, Ullman S. Shifts in selective visual attention: towards the underlying neural circuitry[J]. Matters of intelligence: Conceptual structures in cognitive neuroscience, 1987: 115-141.

[15] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis[J]. IEEE Transactions on pattern analysis and machine intelligence, 1998, 20(11): 1254-1259.

[16] Wolfe J M, Cave K R, Franzel S L. Guided search: an alternative to the feature integration model for visual search[J]. Journal of Experimental Psychology: Human perception and performance, 1989, 15(3): 419.

[17] Van der Velde F, de Kamps M. CLAM: Closed-loop attention model for visual search[J]. Neurocomputing, 2004, 58: 607-612.

[18] Lagomarsino M, Lorenzini M, De Momi E, et al. An online framework for cognitive load assessment in industrial tasks[J]. Robotics and Computer-Integrated Manufacturing, 2022, 78: 102380.

[19] Niu Z, Zhong G, Yu H. A review on the attention mechanism of deep learning[J]. Neurocomputing, 2021, 452: 48-62.

[20] Hong D, Gao L, Yokoya N, et al. More diverse means better: Multimodal deep learning meets remote-sensing imagery classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2020, 59(5): 4340-4354.

[21] Hong D, Zhang B, Li H, et al. Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks[J]. Remote Sensing of Environment, 2023, 299: 113856.

[22] Yao J, Hong D, Wang H, et al. UCSL: Towards Unsupervised Common Subspace Learning for Cross-Modal Image Classification[J]. IEEE Transactions on Geoscience and Remote Sensing, 2023.

[23] Hong D, Zhang B, Li X, et al. SpectralGPT: Spectral

Foundation Model[J]. arXiv:2311.07113, 2023.

[24] Tankus A, Yeshurun Y. Detection of regions of interest and camouflage breaking by direct convexity estimation[C]//Proceedings 1998 IEEE Workshop on Visual Surveillance. IEEE, 1998: 42-48.

[25] Zhang X, Zhu C, Wang S, et al. A Bayesian approach to camouflaged moving object detection[J]. IEEE transactions on circuits and systems for video technology, 2016, 27(9): 2001-2013.

[26] Hall J R, Cuthill I C, Baddeley R, et al. Camouflage, detection and identification of moving objects[J]. Proceedings of the Royal Society B: Biological Sciences, 2013, 280(1758): 20130064.

[27] Ren Z, Kong X, Zhang Y, et al. UKSSL: Underlying Knowledge based Semi-Supervised Learning for Medical Image Classification[J]. IEEE Open Journal of Engineering in Medicine and Biology, 2023.

[28] Zhang Y, Deng L, Zhu H, et al. Deep Learning in Food Category Recognition[J]. Information Fusion, 2023: 101859.

[29] Ren Z, Wang S, Zhang Y. Weakly supervised machine learning[J]. CAAI Transactions on Intelligence Technology, 2023.

[30] Zhong Y, Li B, Tang L, et al. Detecting camouflaged object in frequency domain[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 4504-4513.

[31] Li A, Zhang J, Lv Y, et al. Uncertainty-aware joint salient object and camouflaged object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2021: 10071-10081.

[32] Chaudhari S, Mithal V, Polatkan G, et al. An attentive survey of attention models[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2021, 12(5): 1-32.

[33] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[34] Wang K, Bi H, Zhang Y, et al. D^2C-Net: A Dual-Branch, Dual-Guidance and Cross-Refine Network for Camouflaged Object Detection[J]. IEEE Transactions on Industrial Electronics, 2021, 69(5): 5364-5374.

[35] Sun Y, Chen G, Zhou T, et al. Context-aware cross-level fusion network for camouflaged object detection[J]. arXiv preprint arXiv:2105.12555, 2021.

[36] Fu J, Liu J, Tian H, et al. Dual attention network for scene segmentation[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3146-3154.

[37] Zhang Y, Li K, Li K, et al. Image super-resolution using very deep residual channel attention networks[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 286-301.

[38] Yang M, Yu K, Zhang C, et al. Denseaspp for semantic segmentation in street scenes[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3684-3692.

[39] Perazzi F, Krähenbühl P, Pritch Y, et al. Saliency filters: Contrast based filtering for salient region detection[C]//2012 IEEE conference on computer vision and pattern recognition. IEEE, 2012: 733-740.

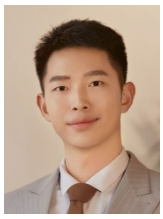
[40] Margolin R, Zelnik-Manor L, Tal A. How to evaluate foreground maps?[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 248-255.

[41] FAN D P, GONG C, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 698-704

[42] Fan D P, Cheng M M, Liu Y, et al. Structure-measure: A new way to evaluate foreground maps[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4548-4557.

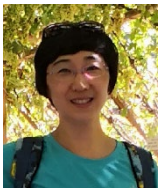
[43] Fan D P, Ji G P, Cheng M M, et al. Concealed object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6024-6042.

This article has been accepted for publication in IEEE Sensors Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JSEN.2023.3343917



Lianrui Xu received his bachelor degree in Environmental Engineering from the Strategic Support Force Information Engineering University in 2018 and his master degree in 2021, and now he is a doctor candidate focused in Mapping science and technology. His research interests include image processing, machine mapping and map construction based on cognitive attention.

Xiong You received the Ph.D. degree from the Institute of Surveying and Mapping in 1996. He is currently working at the Institute of Geospatial Information, Strategic Support Force Information Engineering University as a professor and doctoral supervisor. His research interests include pattern recognition, machine mapping and environmental simulation.



Fenli Jia received the Ph.D. degree in Cartography and Geoinformation Engineering in 2010. She is currently working at the Institute of Geospatial Information, Strategic Support Force Information Engineering University as an associate professor and master's supervisor. She is mainly engaged in scientific research, engineering application and teaching work in the fields of spatial cognition and machine map, holographic map and virtual geographic environment research.

Kangyu Liu received bachelor degree in Applied Psychology from Henan University in 2017 and master degree in Basic Psychology from Henan University in 2020. Since 2020, she has been working as a teaching assistant in Henan Institute of Science Technology and Communication. Her main research direction is cognitive psychology and application of basic psychology.