JEI Journal of Electronic Imaging

RESEARCH PAPER

Camouflage target detection based on strong semantic information and feature fusion

Junhua Yan,[a,b,*] Xutong Hu,[a,b] Yun Su,[c,*] Yin Zhang,[a,b] Mengwei Shi,[a,b] and Yinsen Gao[a,b]

[a]Nanjing University of Aeronautics and Astronautics, Ministry of Industry and Information Technology, Key Laboratory of Space Photoelectric Detection and Perception, Nanjing, China
[b]Nanjing University of Aeronautics and Astronautics, College of Astronautics, Nanjing, China
[c]Beijing Institute of Space Mechanics and Electricity, Beijing, China

ABSTRACT. Aiming at the detection difficulties in camouflage target detection, such as the high similarity between the target and its background, serious damage to the edge, and strong concealment of the target, a camouflage target detection algorithm YOLO of camouflage object detection based on strong semantic information and feature fusion is proposed. First, the attention mechanism convolutional block attention module (CBAM) is constructed to highlight the important channel features and target spatial locations to further aggregate rich semantic information from the high-level feature map. Then the atrous spatial pyramid pooling module is constructed to repeatedly sample the multiscale feature maps to expand the receptive field of the neural network, reduce feature sparsity in the process of convolution, and ensure dense features and multiscale contextual semantic information enter the feature fusion module. Finally, the attention skip-connections are constructed based on the CBAM module for fusing the original feature maps extracted by the backbone network to the corresponding detection outputs so as to eliminate the redundant features as well as enrich the target information of the network outputs. In order to fully verify the performance of the proposed algorithm, a camouflage target detection dataset named strong camouflage efficiency target dataset (SCETD) is constructed. Experimental results on SCETD show that the precision and recall of the proposed algorithm achieve 96.1% and 87.1%, respectively. The $\mathrm{AP}_{0.5}$ and $\mathrm{AP}_{0.5:0.95}$ achieve 92.3% and 54.4%, respectively. The experimental results prove the effectiveness of the proposed method in detecting camouflage targets.

[DOI: 10.1117/1.JEI.32.6.063019]

Keywords: camouflage target; strong semantic information; feature fusion; attention mechanism; receptive field; convolutional neural network

Paper 230039G received Jan. 14, 2023; revised Nov. 5, 2023; accepted Nov. 13, 2023; published Nov. 29, 2023.

1 Introduction

Camouflage targets are difficult to detect because of adopting camouflage strategies such as background matching, edge disruption, and surface disruption,[1] making them share great similarities with the background. The detection accuracy is always reduced when detecting camouflage targets with former algorithms. Thus it is necessary to carry out relative research on camouflage target detection.

At present, a few research teams have begun to study this problem and developed some camouflage target detection methods. Most of them usually considered the camouflage patterns

*Address all correspondence to Junhua Yan, yjh9758@126.com; Yun Su, suedul@163.com as special texture regions. Bhajantri and Nagabhushan[2] cut the image containing the original camouflage target into blocks and calculated the gray-level co-occurrence matrix of each image block to build a tree graph and realized the preliminary detection of the camouflage targets through cluster analysis. Sengottuvelan et al.[3] also proposed a camouflage target detection method based on texture analysis using the gray-level co-occurrence matrix and tree graph. Pan et al.[4] made a three-dimensional convex analysis of the original image containing camouflage targets and proposed adopting the Darg operator to detect the camouflage targets by utilizing the gray difference of the convex structure in camouflage. Wu et al.[5] further optimized the three-dimensional convex analysis method by combining the spatial smoothing filter and improved the detection accuracy on camouflage soldiers hidden in the jungle background. However, the above methods only use the shallow features of the camouflage targets and are sensitive to environmental noise, resulting in limited detection effects.

In recent years, deep learning and convolutional neural networks (CNNs) have been applied to camouflage target detection. Zheng et al.[6] proposed a dense deconvolution neural network (DDCN), in which the pooling method is replaced by deconvolution for upsampling and dense connections between multiscale feature maps are constructed to obtain more semantic information about the targets, thus improving the effectiveness of the segmentation on camouflage people under various backgrounds. However, due to the "checkerboard effect" caused by deconvolution, the detection accuracy of this algorithm is limited. Fang et al.[7] built a strong semantic dilatation network (SSDN), in which series dilatation convolutions are introduced to expand the receptive field of the neural network to obtain more semantic information about the camouflage targets. However, dilatation convolutions in series structure usually extract discrete target information during sampling and the extracted target features may be sparse, resulting in the loss of important features. Therefore, the detection accuracy of this algorithm is limited for camouflage targets. Gupta et al.[8] proposed an image acquisition scheme hardware based compressed acquisition scheme based on the compressed acquisition scheme and deep neural networks. Downsampling, bit truncation, and JPEG are used for image compression, and a deep restoration network deep restoration network for hardware based compressed acquisition scheme is then built based on super-resolution technology to restore the details of the compressed image, reducing the effort spent on the process of image acquisition as well as ensuring users get clear images of high resolution for viewing. The proposed method can be applied to the research on camouflage target detection at low resolution. Camouflage targets share great similarities with the background, which can be solved by using super-resolution technology to enrich the image details and extract more fine-grained features to highlight the difference between the camouflage target and the background, thus improving the detection performance. Deng et al.[9] and Wang et al.[10] proposed the improved RetinaNet and the improved YOLOv5 model, respectively, for camouflage target detection. By introducing the attention mechanism and fusing the information between different channels of the feature maps, the proposed algorithms can suppress the influence of the background noise and other redundant features, which improves the capacity of the algorithms for extracting camouflage texture features. However, the multiscale semantic information of the camouflage targets is not fully utilized, so the detection accuracy of the algorithms on large-scale camouflage targets is limited. Wu et al.[11] proposed a camouflage target detection algorithm based on the improved YOLOv3 network. First, sizes of the anchor boxes are reclustered to better fit the targets. The cascading method of the residual network is changed from single-pole skipping to multilevel skipping, and the channel attention mechanism is integrated, which enhances the ability of the algorithm to extract camouflage features and improve the recall rate of the algorithm. However, the shallow spatial location information and deep semantic information of the targets are not fully utilized, so the precision rate and the overall detection accuracy of the algorithm are limited. Liang et al.[12] proposed a semantic segmentation algorithm CSS-Net for camouflage targets, which combines the multiscale feature extraction method and multilevel attention mechanism to obtain the multiscale representation information and channel information of images. The proposed method improves the segmentation effect of the algorithm on camouflage targets in complex natural environments. However, the real-time performance of the proposed algorithm is not good, making it difficult to meet the real-time requirements of engineering applications.

Camouflage targets have strong concealment and are highly similar to the background, making them difficult to be detected. To solve the problems mentioned above, the YOLOv5[13] is adopted as the baseline algorithm and a camouflage target detection algorithm YOLO of camouflage object detection (COD-YOLO) based on strong semantic information and feature fusion is proposed in this paper. The convolutional block attention module (CBAM) module[14] is constructed to effectively aggregate the channel information and spatial information of low-, middle-, and high-level feature maps. The atrous spatial pyramid pooling (ASPP) module[15] is constructed to densely sample the low-, medium-, and high-level feature maps to obtain the multiscale contextual semantic information of camouflage targets. The channel information, spatial information, and contextual semantic information are utilized to construct the strong semantic information of the camouflage targets. In this paper, an attention skip-connection structure is proposed based on the CBAM module. The attention skip-connections are constructed to connect the detection outputs of the multilevel feature fusion with the corresponding low-, middle-, and high-level feature maps extracted by the backbone of the fully convolutional network, so as to fuse the camouflage features of each layer and enrich the target information of the outputs. COD-YOLO can still detect camouflage targets even when the targets are highly similar to the background, greatly improving the detection accuracy. In order to verify the detection performance of the proposed method, a camouflage target dataset named strong camouflage efficiency target dataset (SCETD) is constructed in this paper.

2 Methods

This paper improves the basic detection algorithm YOLOv5 and proposes a camouflage target detection algorithm COD-YOLO based on strong semantic information and feature fusion, as shown in Fig. 1. The CBAM module is constructed to integrate the channel attention and spatial attention, which effectively highlights the channel features and spatial locations of the camouflage targets in low-, middle-, and high-level feature maps. The ASPP module is constructed to densely sample the low-, middle-, and high-level feature maps, which expands the receptive field of the neural network to obtain the multiscale contextual semantic information of the camouflage targets and enrich the target features for multilevel feature fusion. The channel information, spatial information, and multiscale contextual semantic information can enrich the semantic features and construct the strong semantic information of camouflage targets. Strong semantic information can establish the mapping between the shallow features and the target attributes, which is helpful in detecting camouflage targets when the shallow features are severely damaged.[6,7,16] The attention skip-connection based on the CBAM module is proposed in this paper. The attention skip-connections are constructed between different levels of the fully convolutional network and the corresponding outputs of the multilevel feature fusion, so as to fuse camouflage target features of each layer and enrich the features of the outputs.

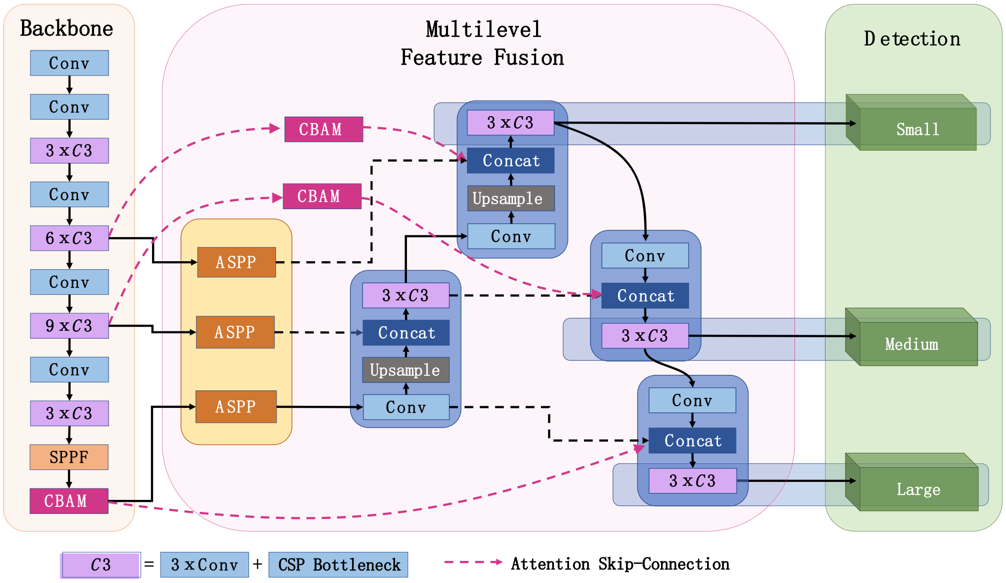

Fig. 1 Structure of COD-YOLO.

The proposed COD-YOLO has a larger receptive field and can extract and retain richer multiscale contextual semantic information for sufficient feature fusion. It can still detect camouflage targets even when the difference between targets and background is little, greatly improving the detection accuracy on camouflage targets.

2.1 Construction of the CBAM-Optimized Backbone

In this paper, the CBAM module is constructed to effectively aggregate the channel information and spatial information of the feature map to optimize the backbone network of YOLOv5. The channel information includes the relevant attributes and categories of the targets, and the spatial information includes the location of the targets. The CBAM module can effectively highlight the important channel features and spatial locations of the target[14] so as to aggregate more semantic information for the subsequent multilevel feature fusion module. Previous studies[17–19] have shown that the low-level feature map contains more shallow features, such as location, edge, and texture, whereas the high-level feature map contains richer semantic information. Camouflage targets are usually hidden in the surrounding environment, and the edge of the targets is always seriously damaged. By constructing the CBAM attention mechanism, COD-YOLO can more accurately capture the dependency between the channel information and spatial location, better extract and retain semantic information of the target in the feature map, and improve the detection performance on camouflage targets.

The function mode of the channel and spatial attention module CBAM in COD-YOLO is shown in Fig. 2. Channel information is aggregated by channel attention, which is followed by spatial attention to highlight target locations. The attention module CBAM can improve the ability of CNNs on feature representation and nonessential feature suppression, thus guiding CNNs to focus on key areas during the training process.[14]

The diagram of CBAM is shown in Fig. 3. C, H, and W represent the number of channels, height, and width of the input feature map, respectively. "Residual" is the residual network module, and "⊗" represents the elemental multiplication operation between the generated channel or spatial attention map and the original feature map.

The CBAM module consists of channel attention and spatial attention. First, the channel attention map is generated. Given the input feature map U, the average pooling and maximum pooling are used to process the input feature map, respectively. The spatial information of the input feature map is aggregated and compressed along the horizontal and vertical directions, and one-dimensional vectors $\mathbf{U}^c_{\text{avg}} \in \mathrm{R}^{C\times 1\times 1}$ and $\mathbf{U}^c_{\text{max}} \in \mathrm{R}^{C\times 1\times 1}$ are output. Both of the generated vectors are then forwarded to a shared network, and the element-level addition operation is performed to generate the channel attention map $M_c \in \mathrm{R}^{C\times 1\times 1}$. The shared network is composed of a multilayer perceptron (MLP) with one hidden layer. Finally, the channel attention map M_c is multiplied with the input feature map U to obtain the output $U' \in \mathrm{R}^{C\times H\times W}$. The above process can be expressed as

$$\begin{aligned} M_c &= \delta(\mathrm{MLP}(\mathrm{AvgPool}(U)) + \mathrm{MLP}(\mathrm{MaxPool}(U))) \\ &= \delta(\mathrm{MLP}(\mathbf{U}^c_{\text{avg}}) + \mathrm{MLP}(\mathbf{U}^c_{\text{max}})), \end{aligned} \tag{1}$$

$$U' = M_c \otimes U, \tag{2}$$

where δ is the sigmoid activation function. The MLP consists of two convolutional transformations F_0 and F_1, in which the convolution kernel sizes are both 1×1. Taking $\mathbf{U}^c_{\text{avg}}$ as an example, the function mode of MLP can be expressed as

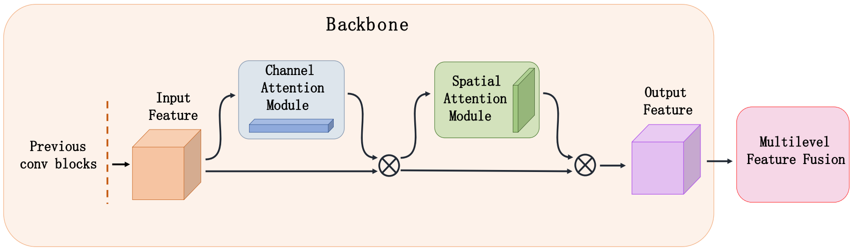

Fig. 2 Function mode of CBAM.

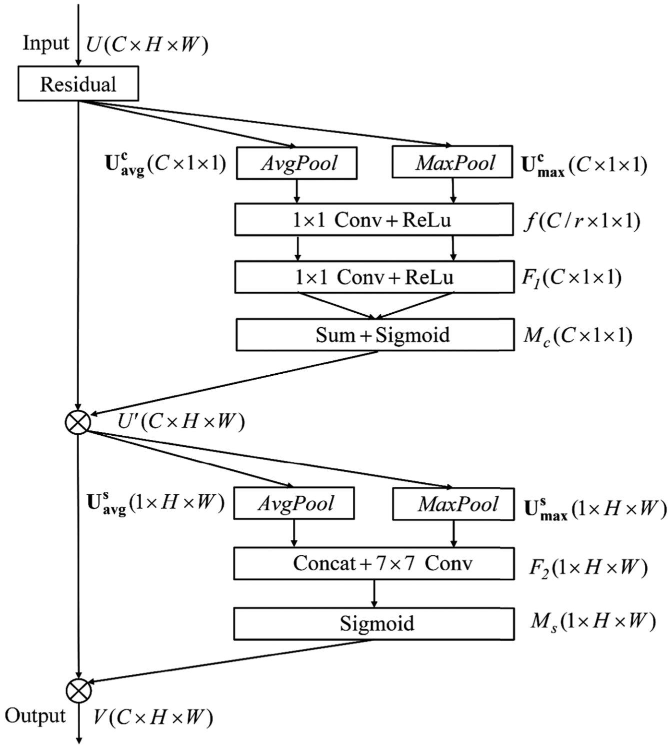

Fig. 3 Diagram of CBAM.

$$\mathrm{MLP}(\mathbf{U}^{c}_{\mathrm{avg}}) = F_1(f) = F_1(\sigma(F_0(\mathbf{U}^{c}_{\mathrm{avg}}))), \tag{3}$$

where σ is the nonlinear activation function ReLu. $f \in \mathrm{R}^{C/r\times1\times1}$ is the intermediate feature map of the one-dimensional channel vector $\mathbf{U}^{c}_{\mathrm{avg}}$. r is the reduction ratio and is set to 16 to reduce the parameters.

The spatial attention module takes the output of the channel attention module as the input and processes it with the average pooling and maximum pooling along the channel axis to aggregate the channel information and obtain the two-dimensional tensors $\mathbf{U}^{s}_{\mathrm{avg}} \in \mathrm{R}^{1\times H\times W}$ and $\mathbf{U}^{s}_{\mathrm{max}} \in \mathrm{R}^{1\times H\times W}$. The obtained two tensors are concatenated and convolved to generate the spatial attention map $M_s \in \mathrm{R}^{1\times H\times W}$. Element multiplication operation is performed on the spatial attention map M_s and the original input to get the final output $V \in R^{C\times H\times W}$. The pooling operation along the channel axis can effectively highlight the spatial information of the target in the feature map.[20] The above process can be expressed as

$$\begin{aligned}\mathrm{M}_s &= \delta(F_2([\mathrm{AvgPool}(U');\mathrm{MaxPool}(U')])) \\ &= \delta(F_2([\mathbf{U}^{s}_{\mathrm{avg}};\mathbf{U}^{s}_{\mathrm{max}}])),\end{aligned} \tag{4}$$

$$V = M_s \otimes U', \tag{5}$$

where δ is the sigmoid activation function. F_2 is the convolutional transform, in which the convolution kernel size is 7×7. $[;]$ represents the concatenate operation.

2.2 Construction of ASPP to Capture Multiscale Contextual Semantic Information

Large receptive field enables CNN to learn more contextual semantic information about the targets, so it can capture more semantic features conducive to camouflage targets.[21,22] The relationship between the receptive field of CNN and multiscale contextual semantic information of camouflage targets is shown in Fig. 4. Boxes with different colors represent receptive fields of different sizes, which contain different scales of contextual semantic information. When the receptive fields are small (such as the green and red boxes), little contextual semantic information of the target can be obtained, making it difficult to determine whether the objects in the box are

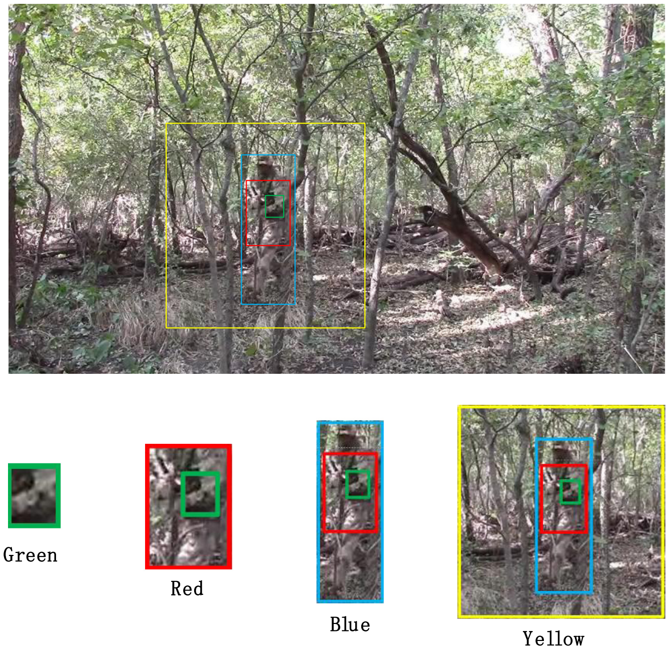

Fig. 4 Relationship between the receptive field of CNN and the multiscale contextual semantic information of the camouflage target.

camouflage targets. When the receptive field is the ground truth (such as the blue box containing the whole target), the obtained contextual semantic information about the target is still very little and it is thus difficult to distinguish the boundary between the target and the surrounding environment. When the receptive field is large (such as the yellow box), it contains not only the target but also many surrounding environments with more contextual semantic information of the background, making the camouflage target easy to be found. Therefore, the ASPP module[15] is constructed in this paper to expand the receptive field of the CNN and capture more multiscale contextual semantic information about the camouflage targets.

Atrous convolution can exponentially expand the receptive field of the kernel of the convolutional layers without downsampling[23] and thus obtain more contextual semantic information from the input feature map. For a convolution kernel with size $k \times k$, the size of the resulting atrous convolution kernel can be formulated as $K = k + (r - 1)(k - 1)$,[14,24] where K is the kernel size of the obtained atrous convolution, r is the atrous rate. $r = 1$ represents a standard convolution. By changing r, the atrous convolution can extract contextual semantic information under different receptive fields.

In this paper, the ASPP module is constructed to process the feature maps generated by the backbone and capture the multiscale contextual semantic information of the camouflage targets. Due to the high similarities between the camouflage targets and the natural environment, CNN with a small receptive field pays more attention to local features, which easily leads to misjudgment between the target and the background.[7] The construction of the ASPP module can expand the receptive field of COD-YOLO and improve the ability of the algorithm to aggregate and represent the target semantic information. The structure of the ASPP module is shown in Fig. 5. A standard convolution with a kernel size of 1×1 and three parallel atrous convolutions with different rates ($r = 6$, 12, and 18) are adopted to repeatedly sample the input feature map under different receptive fields to capture the contextual semantic information of different scales. The global average pooling is used to obtain the global semantic information of the input feature map, and the output is processed by the bilinear interpolation method to recover the spatial resolution after pooling. Outputs are concatenated to obtain richer semantic information about the camouflage targets.

The parallel structure of atrous convolutions in the ASPP module enables CNN to densely sample the feature maps without downsampling and extract more intensive target features at different stages of the network,[15,25,26] providing more abundant features for the subsequent multilevel feature fusion. As shown in Fig. 6, compared with the original input, the feature map after

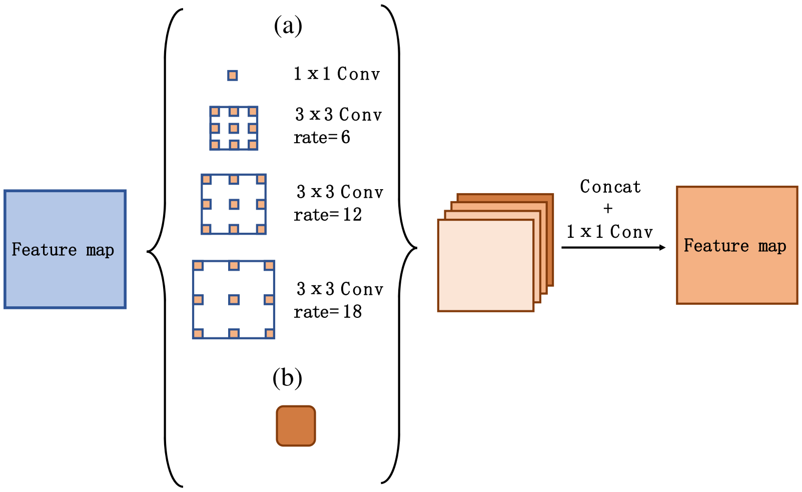

Fig. 5 Structure of ASPP module: (a) atrous spatial pyramid pooling and (b) global avgpool.

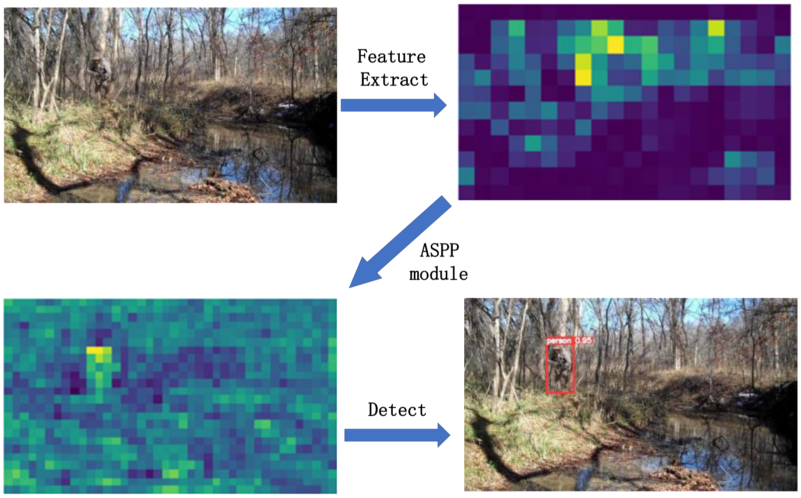

Fig. 6 Dense sampling by the ASPP module.

dense sampling by the ASPP module has clearer contour of the camouflage target and more specific and intensive features, which improves the accuracy of target detection.

2.3 Construction of the Attention Skip-Connection to Fuse Target Features of Each Layer

By constructing skip-connections, original feature maps of different layers extracted by the backbone network are fused to the corresponding output layer of the same scale so as to further enrich the target features such as edge, shape, and location contained in the final prediction layers and enrich the semantic information as well.[27–30] The original feature map is relatively rough,[22] and the similarity between the camouflage targets and the background is high. Directly constructing the skip-connections may introduce redundant features to the outputs, thus affecting the detection accuracy on camouflage targets. Therefore, in this paper, the attention skip-connection is proposed based on the CBAM attention module, as shown in Fig. 7. The CBAM module integrates the channel attention and spatial attention, which can effectively aggregate the channel information and spatial information of feature maps and suppress redundant features. Camouflage targets are usually hidden in the surrounding environment. Utilizing the CBAM module to process the original feature map extracted by the backbone network can retain more semantic and location information about the target when fusing it to the corresponding detection output so as to enrich the target features. The original feature map extracted by the backbone network is processed by the CBAM module and then sent through a skip-connection to participate in the concat fusion

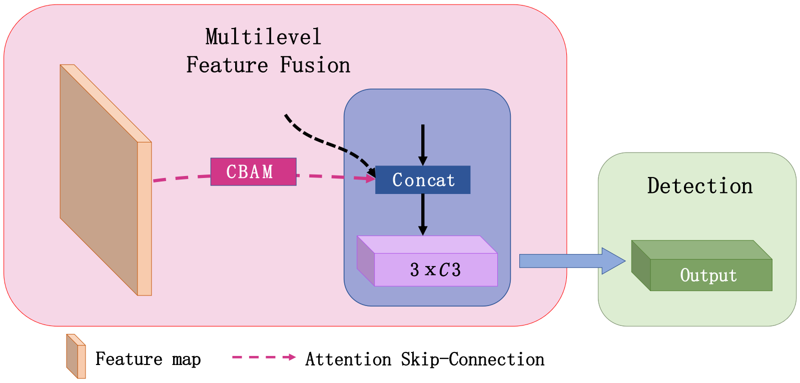

Fig. 7 Attention skip-connection based on the CBAM module.

with the corresponding detection output of multilevel feature fusion. After the convolution conducted by the $C3$ module of the output, the final feature map for target detection is generated.

3 Strong Camouflage Efficiency Target Dataset

At present, there are relatively few datasets for the detection of camouflage targets. A public dataset named camouflage people detection dataset (CPDD) has been constructed[6,7] and some camouflage target detection algorithms[9–12] have conducted experiments based on CPDD, which contains targets with different camouflage efficiency in the image. In order to verify the detection accuracy of the proposed COD-YOLO algorithm on targets with strong camouflage efficiency, a strong camouflage efficiency target detection dataset named SCETD is constructed based on CPDD in this paper.

3.1 Preliminary Selection of Images in CPDD

The public dataset CPDD contains 2600 images with a size of 854×480 pixels and contains camouflage targets of different poses and sizes in various natural scenes. Some images in CPDD have problems, such as poor camouflage efficiency, black shadow of targets, and extremely low-target resolution, as shown in Figs. 8(b)–8(d). In Fig. 8(a), the camouflage efficiency of the target in the image is good; in Fig. 8(b), although the target is wearing a snow camouflage clothing, the camouflage efficiency is poor and the target is obvious; in Fig. 8(c), due to the influence of shooting angle and lighting, the camouflage target turns into a black shadow with clear contour, making the camouflage efficiency poor; in Fig. 8(d), the target is completely covered by the

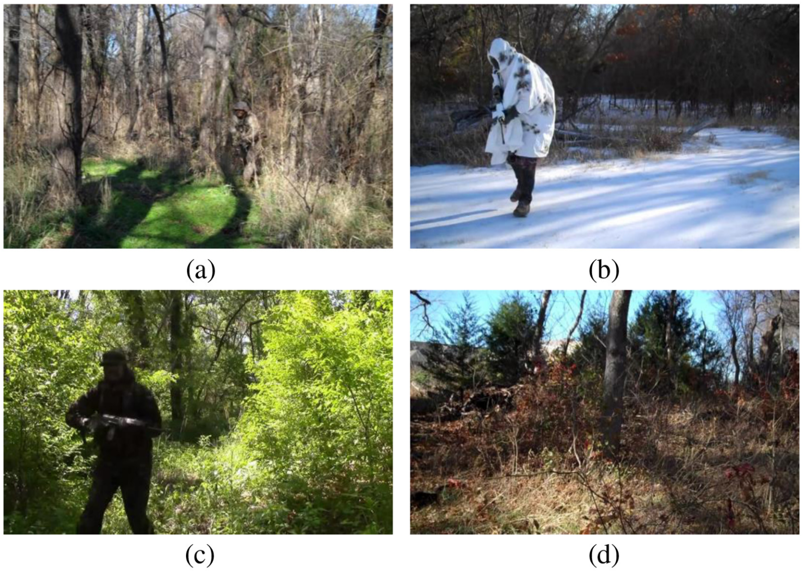
(a)
(b)
(c)
(d)

Fig. 8 Different camouflage efficiencies. (a) Target with good camouflage efficiency, (b) target with poor camouflage efficiency, (c) target with black shadow, and (d) target with low resolution.

bushes and the resolution is too low to label the target correctly. In order to solve the above problems and ensure the strictness and standardization of SCETD in camouflage targets, some images and duplicate images with poor camouflage efficiency, too fuzzy camouflage targets, and extremely poor resolutions are removed from the public dataset CPDD. There are 2458 images retained after preliminary selection.

3.2 Evaluation of the Target Camouflage Efficiency

Calculate the comprehensive similarity between the camouflage target and its background,[31,32] and the greater the comprehensive similarity is, the higher the camouflage efficiency of the target is. According to the comprehensive similarity, the camouflage efficiency of the remaining 2458 images is evaluated, and the results are statistically analyzed. The evaluation results show that the minimum value of the comprehensive similarity is 0.516, and the maximum value is 0.935. According to the evaluation results, the comprehensive similarity of all images is divided into intervals. The statistical results are shown in Table 1.

Images of different camouflage efficiency are shown in Figs. 9(a)–9(d). Figs. 9(a) and 9(b) belong to the interval (0.5, 0.7) and (0.7, 0.8), respectively, and Figs. 9(c) and 9(d) belong to the interval (0.8, 0.95). In Fig. 9(a), the camouflage target is covered by black shadows, making the camouflage texture not obvious. As a result, the camouflage efficiency of the target is low, making it easy to detect. In Fig. 9(b), the black shadow of the camouflage target is eliminated to an extent and the camouflage texture is more obvious. The camouflage efficiency of the target has been improved, but there is still a noticeable color difference between the target and the background, making the contour of the target clear and easy to detect. In Figs. 9(c) and (d), the similarities of color and brightness between the camouflage target and the background are high. Due to the presence of trees, grass, and other shielding conditions, targets have strong concealment and high camouflage efficiency, making them difficult to be detected.

Table 1 Statistical results of camouflage efficiency.

Intervals of comprehensive similarity	Number of images	Proportion (%)
0.5 to 0.7	322	13.10
0.7 to 0.8	916	37.27
0.8 to 0.95	1220	49.63

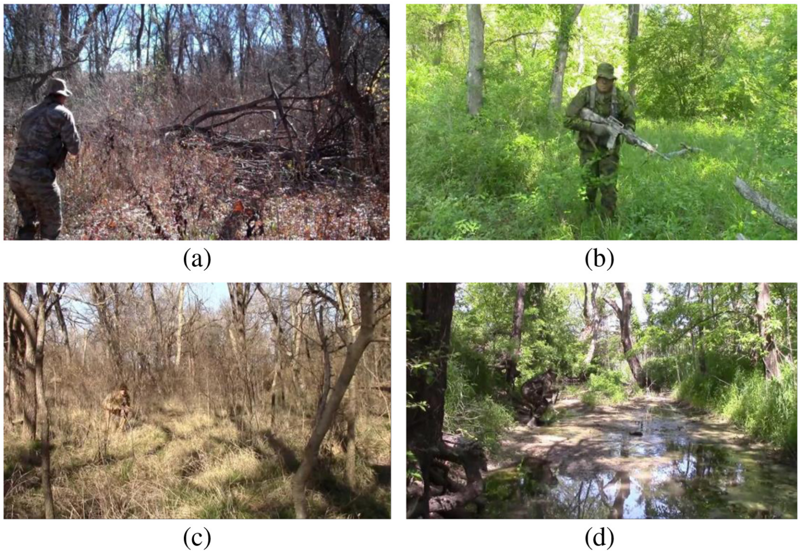

Fig. 9 Images of different camouflage efficiencies. (a) Target with camouflage efficiency of interval (0.5, 0.7), (b) target with camouflage efficiency of interval (0.7, 0.8), (c) target with camouflage efficiency of interval (0.8, 0.95), and (d) target with camouflage efficiency of interval (0.8, 0.95).

3.3 Construction of SCETD

In order to verify the detection accuracy of COD-YOLO proposed in this paper on targets with strong camouflage efficiency, the SCETD is constructed. The remaining 2458 images of CPDD are selected to construct the SCETD dataset according to the following principles.[7]

(1) According to the evaluation results of the camouflage efficiency, select the images of which the comprehensive similarity between the target and the background is >0.80.

(2) The type of the target camouflage and the background should match each other, including camouflage in various natural scenes, such as rainforests, grassland, snow mountains, and deserts.

(3) The dataset should contain camouflage targets in various positions and orientations, including front, back, side, lying down, squatting, standing upright, and running.

(4) The dataset should contain camouflage targets of various distances and sizes.

(5) The dataset should contain camouflage targets under different lighting and occlusion conditions.

Based on the above principles, the SCETD dataset constructed in this paper contains 1220 images totally, and the image size is 854 × 480 pixels. Compared with the original public dataset CPDD, targets in the constructed SCETD dataset have higher camouflage efficiency and are difficult to be detected. SCETD contains 26 camouflage types in different natural scenes. Some image samples and the corresponding camouflage names in SCETD are shown in Fig. 10.

4 Experiments and Analysis

The experiments in this paper mainly include three parts. The first part is the ablation experiment, which is performed on the SCETD dataset. The effectiveness of the CBAM module, ASPP module, and the attention skip-connection of the proposed COD-YOLO in this paper is verified step by step through the ablation experiment. The second part is the superiority verification experiment of COD-YOLO on the detection of strong camouflage efficiency targets. The YOLOv5 and the COD-YOLO are tested, respectively, on CPDD and SCETD to verify that the algorithm

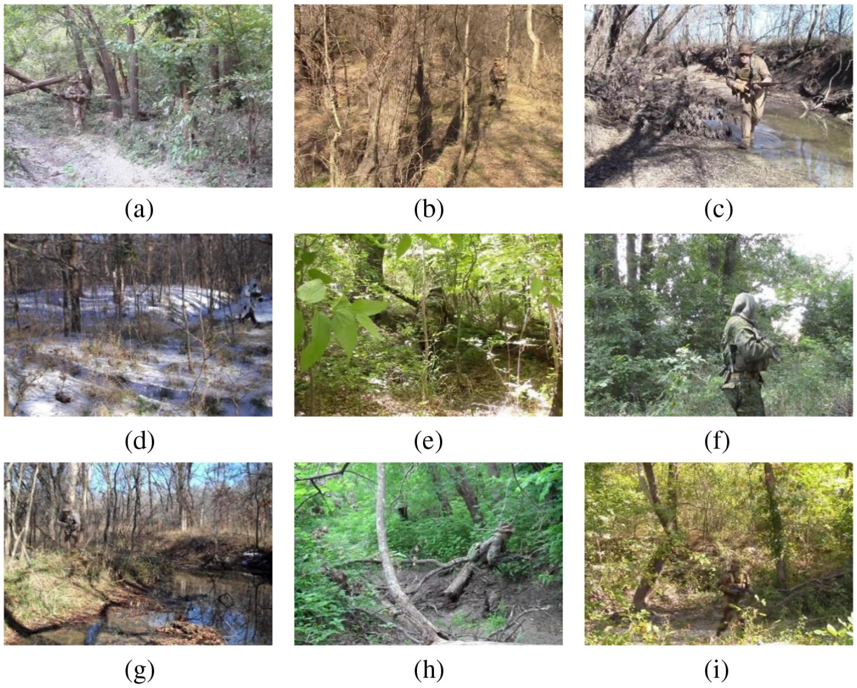

Fig. 10 Samples of SCETD: (a) Arid Fleck, (b) BGS Sumpfmuster, (c) Coyote Tan, (d) German Snow, (e) British DPM, (f) Danish M84, (g) Desert Digital MARPAT, (h) MARPAT Digital Woodland, and (i) Rhodesian Pattern.

proposed in this paper has the advantage of detecting strong camouflage efficiency targets more effectively. The third part is the comparison experiment of the proposed algorithm. The proposed COD-YOLO is compared with five other detection models DDCN,[6] SSDN,[7] YOLOv5,[13] DSSD,[33] and RefineDet[34] on the SCETD dataset to verify that COD-YOLO has higher detection accuracy on strong camouflage efficiency targets.

All the experimental platforms that have been used are Intel Core i5-9400F processor, NVIDIA GeForce RTX 2060 SUPER graphics card (32 GB memory, 8 GB video memory), and Win10 operating system. Programming environment is PyCharm2020, PyTorch1.7, and Python3.8.

The loss function of COD-YOLO includes three parts, namely bounding box regression loss L_{box}, object confidence loss L_{obj}, and object classification loss L_{cls}.L_{box} is calculated using the CIoU[35] method. L_{obj} and L_{cls} are calculated using the binary cross-entropy loss with a sigmoid active function. The total loss L_{total} is formulated as

$$L_{\text{total}} = \sum_{k=0}^{K}\left(\alpha_k^{\text{balance}}\alpha_{\text{box}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\text{II}_{\text{kij}}^{\text{obj}}L_{\text{box}} + \alpha_{\text{obj}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\text{II}_{\text{kij}}^{\text{obj}}L_{\text{obj}} + \alpha_{\text{cls}}\sum_{i=0}^{S^2}\sum_{j=0}^{B}\text{II}_{\text{kij}}^{\text{obj}}L_{\text{cls}}\right), \tag{6}$$

where K is the number of output feature maps at the final detection module and for COD-YOLO, K is 3. S^2 is the number of cells on the output feature map, which will be divided into an $S \times S$ grid. B represents the number of bounding boxes predicted by each grid cell. $\text{II}_{\text{kij}}^{\text{obj}}$ represents the j'th bounding box predicted by the i'th cell on the k'th output feature map. If the bounding box is a positive sample, $\text{II}_{\text{kij}}^{\text{obj}}$ takes the value of 1, otherwise, it is 0. α_* represents the weight of each item, where α_{box}, α_{obj}, and α_{cls} are set to be 0.05, 0.7, and 0.3, respectively. $\alpha_k^{\text{balance}}$ is the weight coefficient utilized to balance the output feature maps of different sizes and is set to be 4.0, 1.0, and 0.4, corresponding to the feature maps of size 80×80, 40×40, and 20×20, respectively, at the detection output of COD-YOLO.

In the process of training, COD-YOLO is trained using stochastic gradient descent algorithm with momentum 0.937 and weight decay 0.0005. The learning rate is set to be 0.01 to warm up the training and will be linearly increased to 0.1 in the first 3 epochs and then annealed down using the cosine decay rule. The batch size is set to 8 and COD-YOLO is trained for 200 epochs in total.

4.1 Evaluating Index

Precision (P), recall (R), $\text{AP}_{0.5}$, and $\text{AP}_{0.5:0.95}$ are utilized to evaluate the effectiveness of the proposed algorithm COD-YOLO. Precision is the percentage of targets that are correctly detected among all detection results. Recall is the percentage of targets that are correctly detected among all targets. Average precision (AP) is the area surrounded by the P–R curve. P, R, and AP are formulated as

$$P = \text{TP}/(\text{TP}+\text{FP}), \tag{7}$$

$$R = \text{TP}/(\text{TP}+\text{FN}), \tag{8}$$

$$\text{AP}_{\text{IoU}} = \int_0^1 P(R)\mathrm{d}R, \tag{9}$$

where TP represents the number of correctly detected targets, FN represents the number of targets that are left out in the background, and FP represents the number of false alarms, which means background incorrectly detected as targets; P is the precision rate; R represents recall; intersection over union (IoU) indicates the threshold of IoU between ground truth and the generated bounding box when judged as a positive sample, i.e., when the value of IoU between ground truth and the bounding box is larger than the threshold, the generated bounding box is determined as a positive sample.

4.2 Ablation Experiment

The ablation experiment is performed based on the SCETD dataset. Images in SCETD are randomly divided into the training set and the validation set according to the ratio of 7:3. Ablation experiment results of the proposed COD-YOLO are evaluated by the precision (P), recall (R), $\text{AP}_{0.5}$, and $\text{AP}_{0.5:0.95}$, which are shown in Table 2.

Table 2 Ablation experiment results of COD-YOLO on the SCETD dataset. Bold values represent that COD-YOLO with all modules applied has achieved the best detection results.

	YOLOv5	COD-YOLO		
CBAM backbone	×	√	√	√
ASPP	×	×	√	√
Attention skip-connection	×	×	×	√
P (%)	89.1	91.3	94.1	**96.1**
R (%)	83.4	84.2	86.8	**87.1**
$\mathrm{AP}_{0.5}$ (%)	88.6	88.7	90.9	**92.3**
$\mathrm{AP}_{0.5:0.95}$ (%)	52.0	53.4	54.2	**54.4**

Note: "×" indicates the corresponding module is not applied, and "√" means the contrary.
Abbreviations: *P*, precision; *R*, recall; and AP, average precision;

As shown in Table 2, compared to the YOLOv5 algorithm in the first column, each evaluation index in the second column with the addition of CBAM module has been improved, among which P is increased by 2.2% and $\mathrm{AP}_{0.5:0.95}$ is increased by 1.4%. The evaluation indexes in the third column with the addition of ASPP module are also improved compared to the data in the second column, among which P and R are increased by 2.8% and 2.6%, respectively, and $\mathrm{AP}_{0.5}$ is increased by 2.2%. Compared to the data in the third column, the evaluation indices in the fourth column with the addition of the attention skip-connections are improved continually, among which P is increased by 2% and $\mathrm{AP}_{0.5}$ is increased by 1.4%. In conclusion, the CBAM module, ASPP module, and the attention skip-connection have all improved the detection accuracy of the proposed algorithm to a certain extent in detecting targets with strong camouflage efficiency.

4.3 Superiority Experiment on Detecting Target with Strong Camouflage Efficiency

The YOLOv5 algorithm and the COD-YOLO algorithm proposed in this paper are respectively tested on the original public dataset CPDD and the constructed dataset SCETD. Experiment results are evaluated by $\mathrm{AP}_{0.5}$ as shown in Table 3.

As shown in Table 3, with the improvement in target camouflage efficiency, COD-YOLO is superior to YOLOv5 by 1.6% to 3.7% for $\mathrm{AP}_{0.5}$. When the comprehensive similarity between the camouflage target and the background is >0.8, the $\mathrm{AP}_{0.5}$ of the YOLOv5 drops below 90%, whereas COD-YOLO remains above 90%, which proves that the COD-YOLO proposed in this paper can more effectively detect camouflage targets, and has superiority in detecting targets with strong camouflage efficiency. As one of state-of-the-art target detection algorithms at present, the detection accuracy of YOLOv5 is not much different from that of the proposed COD-YOLO when the overall camouflage efficiency of the targets in the dataset is not high. However, when the camouflage efficiency of targets increases, the gap between the detection accuracy of YOLOv5 and COD-YOLO becomes obvious.

Table 3 Detection results on targets with strong camouflage efficiency. Bold values represent that COD-YOLO has the best detection results compared to other five algorithms.

Dataset	Algorithm	$\mathrm{AP}_{0.5}$ (%)	(Δ%)
CPDD (comprehensive similarity > 0.5)	COD-YOLO	94.5	1.6
	YOLOv5	92.9	
SCETD (comprehensive similarity > 0.8)	COD-YOLO	92.3	3.7
	YOLOv5	88.6	

To further prove the superiority of COD-YOLO proposed in this paper on detecting strong camouflage efficiency targets, the feature maps of three detection layers at the output end of YOLOv5 and COD-YOLO are visualized. For the camouflage targets of large, medium and small sizes, the feature maps of three detection layers at the output end of YOLOv5 and COD-YOLO is compared, as shown in Fig. 11. Assuming that the size of the original input image is 1, sizes of the feature maps at three detection layers are 1/8, 1/16, and 1/32, respectively. The smaller

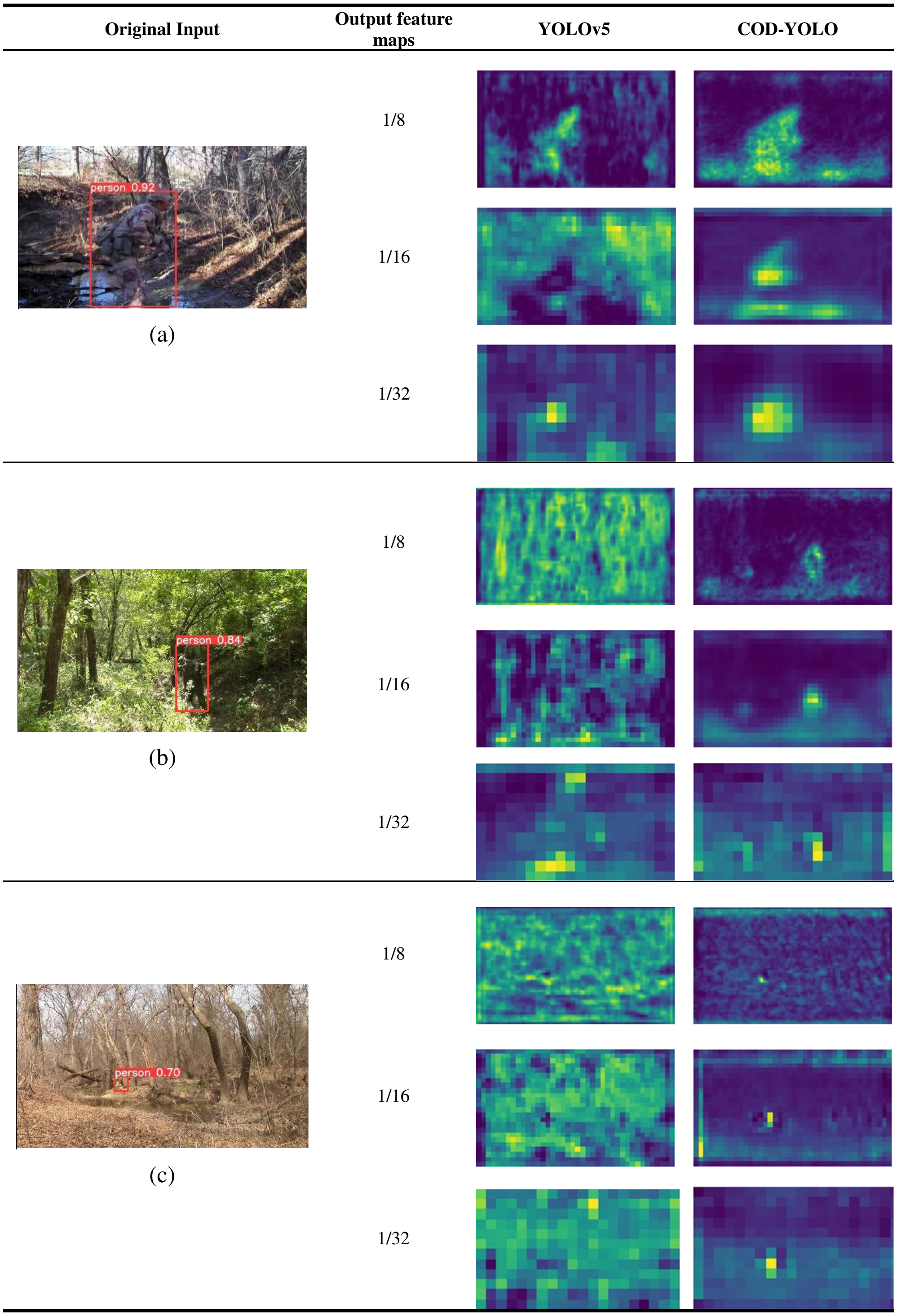

Fig. 11 Comparison results of the output feature maps under multiscale targets: camouflage target of (a) large size, (b) medium size, and (c) small size.

Table 4 Comparison results of COD-YOLO and other five algorithms.

Algorithm	Backbone	P (%)	R (%)	$\mathrm{AP}_{0.5}$ (%)	$\mathrm{AP}_{0.5:0.95}$ (%)	F_β (%)	MAE
SSDN	VGG-Net	—	—	—	—	66.1	0.005
DDCN	VGG16-Net	—	—	—	—	76.0	0.004
RefineDet	ResNet101	81.2	75.9	80.2	37.9	79.9	0.0038
DSSD	ResNet101	84.2	80.2	83.5	46.4	83.2	0.0034
YOLOv5	CSPDarkNet53	89.1	83.4	88.6	52.0	87.7	0.0029
COD-YOLO (This paper)	CSPDarkNet53	**96.1**	**87.1**	**92.3**	**54.4**	**93.8**	**0.0027**

Note: "—" indicates that the corresponding data are not provided in the original paper.

the size of the feature map is, the lower the resolution is, and the more abstract the representation of the target features is.

As shown in Table 4, the scene where the camouflage target is located contains dense and cluttered trees and grasslands, which makes the feature maps extracted by the CNN contain a lot of background noise. In the output feature maps of YOLOv5, the strong response area of neurons covers almost the entire feature map. It is thus difficult to distinguish the target from the background and easy to generate many missed detections and false alarms. In the output feature maps of COD-YOLO, the strong response area of neurons almost only appears in and around the target region. The target region is clear, making it easy to distinguish the target from the background and thus improving the detection accuracy of COD-YOLO on camouflage targets. The above analysis indicates that the CBAM module and attention skip-connections in COD-YOLO can guide the network to focus on key areas and filter out redundant features during the training process; the ASPP module can aggregate more low-level features, such as edge and texture as well as high-level semantic features of camouflage targets for multilevel feature fusion. The feature maps at the output end can thus be optimized to improve the detection accuracy of the proposed algorithm COD-YOLO on detecting targets with strong camouflage efficiency.

4.4 Comparison Experiment

The proposed COD-YOLO is compared with DDCN, SSDN, YOLOv5, DSSD, and RefineDet on the SCETD dataset constructed in this paper. DSSD is a representative one-stage target detection algorithm based on CNN, and RefineDet is a representative two-stage target detection algorithm based on CNN. SSDN is a strong semantic dilatation network and DDCN is a dense deconvolution neural network, both of which are designed for camouflage target detection. Utilize P, R, $\mathrm{AP}_{0.5}$, $\mathrm{AP}_{0.5:0.95}$, F measure ($\beta = 0.3$), mean absolute error (MAE), P–R curve to evaluate the detection results. The experimental dataset of DSSD, RefinDet, YOLOv5, and COD-YOLO algorithms is all SCETD dataset. Codes of the camouflage target detection algorithms SSDN and DDCN have not yet been open-sourced, so the original data of these two algorithms are directly taken from the corresponding articles for comparison in this paper. The comparison results are shown in Table 4, and the P–R curves of each algorithm are shown in Fig. 12.

As shown in Table 4, the detection performance of COD-YOLO proposed in this paper is best compared with the other five models, with $\mathrm{AP}_{0.5}$ of 92.3%, $\mathrm{AP}_{0.5:0.95}$ of 54.4%, F measure of 93.8%, and MAE of 0.0027. Experiment results show that COD-YOLO has higher detection accuracy for camouflage targets than other algorithms and can better detect camouflage targets in different natural backgrounds. In addition, the position of the target detection box output by the COD-YOLO algorithm is closer to the position of the ground truth, and the target detection box better fits the periphery of the target. As shown in Fig. 12, COD-YOLO has a larger lower surrounding area and is closer to the upper right corner compared with other five algorithms, indicating that COD-YOLO can still maintain good recall rates at higher confidence thresholds. The above analysis can prove that COD-YOLO proposed in this paper performs better on camouflage target detection.

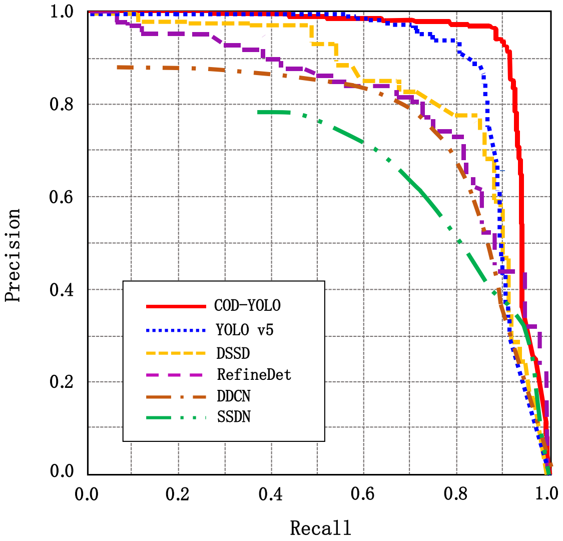

Fig. 12 Comparison of *P–R* curves.

5 Conclusion

In this paper, a detection algorithm COD-YOLO based on strong semantic information and feature fusion is proposed on detecting camouflage targets. The receptive field expansion and dense sampling module ASPP is constructed in COD-YOLO to extract multiscale semantic features of camouflage targets and provide richer target semantic information for multilevel feature fusion, improving the capabilities of the network in feature extraction and representation on camouflage targets. The CBAM attention module is constructed to sample the high-level feature map and the skip-connections based on CBAM are constructed to highlight the important channel features and spatial location features of the camouflage targets in the low-, middle-, and high-level feature maps and integrate them into the corresponding detection output, so as to suppress the background noise and further enrich the semantic information and other important target features of the outputs, which improve the detection accuracy of COD-YOLO for camouflage targets in cluttered natural backgrounds. COD-YOLO has a precision and recall rate of 96.1% and 87.1%, respectively, and achieves $\text{AP}_{0.5}$ of 92.3%, $\text{AP}_{0.5:0.95}$ of 54.4%, F measure of 93.8% and MAE of 0.0027 on the constructed strong camouflage efficiency target dataset SCETD, indicating that COD-YOLO has higher detection accuracy and location accuracy on camouflage targets.

In the future, COD-YOLO can be further optimized. The impact of detailed information carried by features of different levels will be studied, and more efficient feature fusion methods will be explored. Super-resolution method can be used to restore image details and generate more levels of features for efficient feature fusion, which is helpful to perform camouflage target detection even at low resolution. The adaptive mechanism of the receptive field is considered to be designed to adaptively adjust the size of the receptive field according to the depth of the network so as to obtain the multiscale contextual semantic information of camouflage targets more efficiently and improve the effectiveness of COD-YOLO in detecting camouflage targets on the current basis.

Code and Data Availability

We are glad to provide code and data to the interested researchers to replicate or interpret the findings reported in this paper. Anyone can contact Miss Hu via email and she will provide you with the code and data. Her email address is 1170741450@qq.com.

Acknowledgments

This work was supported by the National Defense Science and Technology Foundation Strengthening Program (Grant No. 2021-JCJQ-JJ-0834) and the Fundamental Research Funds for the Central Universities (Grant No. NJ2022025, NP2022450, and NJ2023029).

References

1. I. C. Cuthill, “Camouflage,” *J. Zool.* **308**(2), 75–92 (2019).
2. N. U. Bhajantri and P. Nagabhushan, “Camouflage defect identification: a novel approach,” in *9th Int. Conf. Inf. Technol. (ICIT’06)*, IEEE, pp. 145–148 (2006).
3. P. Sengottuvelan, A. Wahi, and A. Shanmugam, “Performance of decamouflage through exploratory image analysis,” in *First Int. Conf. Emerg. Trends in Eng. and Technol.*, IEEE, pp. 6–10 (2008).
4. Y. Pan et al., “Study on the camouflage target detection method based on 3D convexity,” *Mod. Appl. Sci.* **5**(4), 152 (2011).
5. G. Wu et al., “Application of three-dimensional convex analysis in pattern painting camouflage detection,” *J. PLA Univ. Sci. Technol.* **16**(6), 582–586 (2015).
6. Y. Zheng et al., “Detection of people with camouflage pattern via dense deconvolution network,” *IEEE Signal Process. Lett.* **26**(1), 29–33 (2018).
7. Z. Fang et al., “Camouflage people detection via strong semantic dilation network,” in *Proc. ACM Turing Celebration Conf.*, China, pp. 1–7 (2019).
8. P. S. Gupta, X. Yuan, and G. S. Choi, “DRCAS: deep restoration network for hardware based compressive acquisition scheme,” in *IEEE Int. Conf. Image Process. (ICIP)*, IEEE, pp. 291–295 (2020).
9. X. Deng et al., “Research on detection of people with camouflage pattern via improving RetinaNet,” *Comput. Eng. Appl.* **57**(5), 190–196 (2021).
10. Y. Wang et al., “Camouflage object detection based on improved YOLOv5 algorithm,” *Comput. Sci.* **48**(10), 226–232 (2021).
11. T. Wu, L. Wang, and J. Zhu, “Camouflage target detection based on an improved YOLOv3 algorithm,” *Fire Control Command Control* **47**(2), 114–120+126 (2022).
12. X. Liang et al., “Camouflage target segmentation algorithm using multi-scale feature extraction and multi-level attention mechanism,” *J. Comput.-Aid. Des. Comput. Graph.* **34**(5), 683–692 (2022).
13. J. Glenn, “YOLOv5,” https://github.com/ultralytics/yolov5 (2020).
14. S. Woo et al., “CBAM: convolutional block attention module,” *Lect. Notes Comput. Sci.* **11211**, 3–19 (2018).
15. L. C. Chen et al., “Rethinking atrous convolution for semantic image segmentation,” arXiv:1706.05587 (2017).
16. D. Bau et al., “Network dissection: quantifying interpretability of deep visual representations,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 6541–6549 (2017).
17. S. Liu et al., “Path aggregation network for instance segmentation,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 8759–8768 (2018).
18. T. Y. Lin et al., “Feature pyramid networks for object detection,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 2117–2125 (2017).
19. T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 3085–3094 (2019).
20. N. Komodakis and S. Zagoruyko, “Paying more attention to attention: improving the performance of convolutional neural networks via attention transfer,” in *ICLR* (2017).
21. H. Mei et al., “Don’t hit me! Glass detection in real-world scenes,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 3687–3696 (2020).
22. H. Mei et al., “Exploring dense context for salient object detection,” *IEEE Trans. Circuits Syst. Video Technol.* **32**(3), 1378–1389 (2021).
23. F. Yu and V. Koltun, “Multi-scale context aggregation by dilated convolutions,” arXiv:1511.07122 (2015).
24. P. Wang et al., “Understanding convolution for semantic segmentation,” in *IEEE Winter Conf. Appl. of Comput. Vis. (WACV)*, IEEE, pp. 1451–1460 (2018).
25. L. C. Chen et al., “DeepLab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs,” *IEEE Trans. Pattern Anal. Mach. Intell.* **40**(4), 834–848 (2017).
26. L. C. Chen et al., “Encoder-decoder with atrous separable convolution for semantic image segmentation,” *Lect. Notes Comput. Sci.* **11211**, 801–818 (2018).
27. S. Xie and Z. Tu, “Holistically-nested edge detection,” in *Proc. IEEE Int. Conf. Comput. Vis.*, pp. 1395–1403 (2015).
28. G. Bertasius, J. Shi, and L. Torresani, “DeepEdge: a multi-scale bifurcated deep network for top-down contour detection,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 4380–4389 (2015).
29. J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 3431–3440 (2015).
30. M. Tan, R. Pang, and Q. V. Le, “EfficientDet: scalable and efficient object detection,” in *Proc. IEEE/CVF Conf. Comput. Vis. and Pattern Recognit.*, pp. 10781–10790 (2020).
31. J. Dai, *Research on Camouflage Effect Evaluation Method Based on Similarity*, Xian Technological University (2018).

32. Z. Wang, *Research on Background Based Camouflage Design and Comprehensive Evaluation Method*, Northeastern University (2014).
33. C. Fu et al., "DSSD: deconvolutional single shot detector," CoRR,abs/1701.06659 (2017).
34. S. Zhang et al., "Single-shot refinement neural network for object detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, pp. 4203–4212 (2018).
35. Z. Zheng et al., "Distance-IoU loss: faster and better learning for bounding box regression," in *Proc. AAAI Conf. Artif. Intell.*, Vol. **34**, pp. 12993–13000 (2020).

Junhua Yan received her BSc, MSc, and PhD degrees from Nanjing University of Aeronautics and Astronautics in 1993, 2001, and 2004, respectively. She is a professor at Nanjing University of Aeronautics and Astronautics. She is the author of more than 60 journal papers and has 12 patents. Her current research interests include multisource information fusion; target detection, tracking, and recognition; and image quality assessment.

Xutong Hu received her BSc degree from Wuhan University of Technology in 2021. Now she is a master's student at Nanjing University of Aeronautics and Astronautics. Her main research interests include camouflage target detection and tracking.

Yun Su received his BSc degree from Beijing Institute of Technology in 2005 and his MSc degree from China Academy of Space Technology in 2008. He is currently a researcher at Beijing Institute of Space Mechanics and Electricity. He holds more than 90 patents. His research interests include space optical remote sensing, computational optical imaging, and optical system design.

Yin Zhang received his BSc degree from Jilin University in 2009 and received his MSc and PhD degrees both from Harbin Institute of Technology in 2011 and 2016, respectively. He is currently an associate professor at Nanjing University of Aeronautics and Astronautics. His main research interests include simulation and processing of photoelectric detection information and spectral radiation characteristics of photoelectric detection scene.

Mengwei Shi received her BSc degree from Nanjing University of Aeronautics and Astronautics in 2021. Now she is a doctoral candidate at Nanjing University of Aeronautics and Astronautics. Her main research interest includes target camouflage effectiveness evaluation.

Yinsen Gao received his BSc degree from the Southwest Jiaotong University in 2020. Now he is a master's student at Nanjing University of Aeronautics and Astronautics, and his main research interest includes camouflage target detection and identification.