Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 7703444, 12 pages
https://doi.org/10.1155/2022/7703444

Research Article

Research on Camouflaged Human Target Detection Based on Deep Learning

Wei Zhang, Qikai Zhou, Ruizhi Li, and Fu Niu

Academy of Systems Engineering of Academy of Military Science of Chinese PLA, Beijing 100166, China

Correspondence should be addressed to Fu Niu; niufu@vip.sina.com

Received 8 June 2022; Revised 7 August 2022; Accepted 14 September 2022; Published 12 October 2022

Academic Editor: Anandakumar Haldorai

In the task of camouflaged human target detection, the target is highly integrated with the complex environment background, which is difficult to identify and leads to false detection and missed detection. A detection algorithm MC-YOLOv5s is proposed for the characteristics of camouflaged targets. The algorithm takes YOLOv5s as the basic framework. First, a multispectral channel attention module is embedded in the backbone feature extraction network, which enhances the network's ability to extract camouflaged target features, weakens the attention to the surrounding background, and effectively improves the algorithm's antibackground interference. Second, the original upsampling operation is replaced by a lightweight general upsampling operator to achieve effective fusion of high-resolution low-level feature maps and low-resolution high-level feature maps. Finally, the K-means++ clustering method is used to optimize the anchor boxes of the dataset target, and the sizes of the generated priori boxes are allocated to each detection layer, which increases the matching degree between the priori boxes and the actual target boxes, and further improves the detection accuracy of the algorithm. The training and verification were carried out on the military camouflaged personnel dataset (MCPD), the precision (P), recall (R), and mean average precision (mAP) of the MC-YOLOv5s algorithm reached 97.4%, 86.1%, and 94%, respectively. Compared to the original YOLOv5s model, mean average precision (mAP) is increased by 3.7 percentage points. The improved algorithm has better detection effect, is more sensitive to camouflaged targets, and achieves accurate positioning and identification of camouflaged human targets. If the proposed MC-YOLOv5s is applied to personnel search and rescue in complex battlefields and natural disaster environments, it can greatly improve personnel search and rescue efficiency and life survival rate, and reduce the consumption of human and material resources in rescue.

1. Introduction

In recent years, target detection technology has made great achievements in all walks of life and has been widely used in military and civilian fields [1, 2]. At present, with the continuous development of high and new technologies such as artificial intelligence and big data, in the military field, modern warfare has developed from the previous mechanization to an unmanned, information-based, and intelligent mode, and various cutting-edge weapons and equipment in various countries in the world are gradually being developed. The development and improvement in the direction of technology and intelligence, especially the detection technology of military targets, has obviously become the key technology of battlefield situational awareness [3]. For intelligent weapons and equipment, the first problem to be solved is to make them have keen observation and identification capabilities. Therefore, in the complex and changeable battlefield environment, it is of great strategic significance for us to automatically and quickly identify potential enemy military targets.

In the civilian field, target detection algorithms have played an important role in the fields of autonomous driving, intelligent navigation, medical image recognition, and search and rescue [4, 5]. Especially when a natural disaster strikes unexpectedly, it may cause a large area of houses to collapse and cause casualties, so the top priority is to rescue the survivors in the rubble in time. At this time, rescue robots and search and rescue drones played a major role in emergency rescue. Rescue robots can quickly identify

and locate survivors and carry out rescue work in complex and changeable disaster sites, especially in narrow spaces where fire rescue personnel, search and rescue dogs, or other rescue equipment tools are difficult to reach quickly. Compared with traditional remote sensing images and high-altitude aerial photography technology, search and rescue drones with target recognition and positioning functions are more efficient. In the future, more advanced and mature target detection technology will be applied to equipment such as search and rescue robots or search and rescue drones, which can greatly improve the search and rescue efficiency of personnel.

At present, the existing target detection algorithms are not outstanding in the above-mentioned target detection effect in the complex environment, especially the detection of camouflaged human targets in the complex environment, which is more difficult than the conventional detection task. In addition, military camouflage technology mainly changes the surface characteristics of the target through color, texture, and various patterns to achieve the effect of integrating into the surrounding complex environment to achieve the purpose of stealth, and some camouflage effects even reach the level that the human eye cannot recognize [6]. Not only that, the field environment where military targets are usually located is relatively complex, such as mountains, snow, deserts, and bushes, which may contain various military targets, all of which have brought enormous challenges to existing target detection algorithms.

After research and analysis, the difficulty of camouflaged human target detection mainly has the following two points: (1) the concealment of the target is very high, and the recognition degree is very low. Due to the complex diversity of its own dyes, patterns, and lines, the target is extremely integrated with the surrounding environment, and the discrimination is low; (2) overlapping occlusion is more common. Since the loss of target feature information is proportional to the degree of occlusion, overlapping occlusion will cause the algorithm network to lose part of the camouflage feature information during the feature extraction process, which in turn makes the training of the neural network more difficult. Therefore, if a convolutional neural algorithm with high accuracy and strong generalization ability can be developed, it will have important practical significance for the rapid search and positioning of targets in the field of intelligent combat and battlefield search and rescue.

Traditional military target detection algorithms are mainly identified by HOG [7], SVM [8], DPM [9], and other methods. However, the feature extraction ability of such algorithms is not enough, which leads to the loss of most of the target feature information, so the detection task of military targets in complex environments is more difficult, especially in the task of camouflaged target detection, which generally shows poor results. Since 2012, with the continuous upgrading of computers, the performance of many hardware devices has become better and better, and the target detection technology based on deep learning has also ushered in its rapid development period. Great progress has been made. At present, the military camouflage target detection algorithms for deep learning can be simply divided into two categories: one is the two-stage detection model (two stage), which uses the method of first dividing the target candidate area, then classifying and regressing, and representing the algorithm. There are R-CNN, Fast R-CNN, Faster R-CNN [10], etc.; the second is a single-stage detection model (one stage), which uses a network to directly classify and regress the target, and the representative algorithm has SSD [11], EfficientNet [12], RetinaNet [13], and YOLO series [14–17].

At present, many studies have applied target detection algorithms based on deep learning to precise guidance, battlefield search and rescue, and other military fields, and have achieved good detection results. According to Wang et al., [18] applying the Faster R-CNN algorithm to the tank armor target detection task in a complex background, compared with the traditional classic algorithm, effectively improved the target recognition rate. According to Yu and Lv [19], in response to the problem of not high detection accuracy of the traditional detection algorithm, it further improved the detection accuracy of military targets by introducing methods such as improving the ResNet50-D residual network and dual attention mechanism. Deng et al. [20] introduced an attention mechanism to imitating human vision on the basis of the RetinaNet algorithm and increased the detection accuracy of the camouflage target to 93.1%. The above studies have achieved good results for the detection tasks of military camouflage, but there is still room for improvement in the detection accuracy.

Based on this, this paper makes the following three improvements on the basis of the YOLOv5s algorithm framework for the characteristics that the camouflaged human target is highly similar to the surrounding environment: first, the multispectral channel attention module (MCA) is introduced in the backbone feature extraction network of YOLOv5s. The attention mechanism is added to make the algorithm network, which has the same keen observation ability as the human eye, so as to improve the network's full extraction of the information of the camouflaged human target, strengthen the algorithm's attention and antibackground interference ability, and further solve the problem that the original algorithm is insufficient to extract the features of the camouflaged human target; the second is to use the lightweight general upsampling operator CARAFE to replace the original upsampling operation to strengthen the effective fusion of high-resolution low-level feature maps and low-resolution high-level feature maps, so that the upsampling operation focuses on camouflage target area and the background area, which are ignored to achieve the content perception effect, so as to reduce the false detection rate of the original algorithm; the third is to use the K-means++ clustering algorithm to optimize the target a priori frame, and reclustering the dataset to generate the size of the new a priori frame is allocated to each detection layer, which increases the matching degree between the a priori frame and the actual target frame, and further enhances the network's ability to detect overlapping occluded targets. Through the above three improvements, this paper proposes a human target detection algorithm MCA-CARAFE-

YOLOv5s (hereinafter referred to as "MC-YOLOv5s") for camouflage characteristics, which effectively improves the algorithm's detection accuracy of camouflaged human targets in complex environments.

The scheme in this paper is applied to the complex and changeable battlefield environment. It can realize the recognition ability of various weapons and equipment similar to human eyes, automatically and quickly identify potential enemy military targets, and assist commanders in making battlefield decisions and achieve the goal of defeating the enemy. At the same time, when human-machine natural disasters and man-made disasters come, rescue robots and search and rescue drones using this technology can quickly search and rescue survivors trapped in the rubble, and rescue robots can quickly find and locate in complex environments. Survivors replace or assist manual rescue at the disaster site, especially in small spaces (such as holes less than 1 meter in diameter) that cannot be reached by humans, dogs, or other detection equipment, and the rescue safety factor is higher. It can automatically search and locate lost ships and personnel from a high altitude and a wide range, which greatly improves the efficiency of personnel search and rescue and the survival rate of life, and greatly reduces the consumption of manpower and material resources. In conclusion, this solution has important practical application value in both military and civilian fields.

2. The Detection Algorithm for Camouflaged Human Objects Based on Deep Learning

2.1. The Basic Principle of YOLOv5 Algorithm. The YOLOv5 algorithm is a detection algorithm with excellent detection speed and detection accuracy. The algorithm consists of four parts: the Input, the Backbone, the Neck and the Prediction. Its network structure is shown in Figure 1. The first is the input terminal, Input, the size of the input picture is generally 640×640. The second part is the backbone feature extraction network, Backbone, which is mainly composed of Focus, Conv, C3, Spatial Pyramid Pooling (SPP) [21], and other structures. The Focus structure is mainly the process of slicing and splicing feature maps in order to retain more feature information, and Conv is the basic convolution process, which mainly performs three operations such as 2-dimensional convolution, regularization, and activation on the input feature map. The C3 module consists of several Bottleneck structures. The third part is the feature fusion network Neck, which consists of Feature Pyramid Networks (FPN) [22] and Path Aggregation Networks (PANet) [23] structures. The combination of the two further improves the model's ability to target Feature attention. The fourth part is the detection layer Prediction. The three detection layers correspond to the detection and recognition of large-, medium-, and small-scale targets, respectively. The YOLOv5 algorithm model has four versions: *s*, *m*, *l*, and *x* according to the number of network layers. The change in the number of structural layers is mainly realized by changing the two parameters of depth multiple and width multiple. Among them, YOLOv5s has the least number of network layers, a simpler structure and the fastest speed.

2.2. A Camouflaged Human Target Detection Model Based on Improved YOLOv5. The YOLOv5 detection algorithm is one of the typical and excellent single-stage target detection algorithms. However, in the task of camouflage target detection under complex background, because the complex environment background in the picture usually occupies more, so in the process of extracting feature information, it is easy to generate redundant environment background information, which makes it difficult for the network to effectively extract the target feature information, and with the deepening of the network structure, the target feature information is seriously lost, resulting in poor target detection results. Therefore, this paper proposes a human target detection algorithm MC-YOLOv5s aiming at the characteristics of camouflaged targets, such as high concealment, low discrimination, and overlapping occlusion, which effectively improves the shortcomings of the original YOLOv5 algorithm and realizes the accurate identification and rapid positioning of camouflaged human targets in complex environments.

The network structure of the improved algorithm MC-YOLOv5s is shown in Figure 2. Input is the input terminal, and the size of the input picture is generally 640×640. Backbone is the backbone feature extraction network, which consists of modules such as Focus, Conv, C3, SPP, and multispectral channel attention module (MCA). The Focus structure is to slice and splice the input feature map, and add a value at every pixel point in the horizontal and vertical directions. More image information is preserved. Conv is the basic convolution unit of YOLOv5, which sequentially performs two-dimensional convolution, regularization, activation, and other operations on the input. C3 is composed of several Bottleneck modules. Bottleneck is a classic residual structure. After the input passes through two convolution layers, the Add operation is performed with the original value to complete the residual feature transfer without increasing the output depth. SPP is a spatial pyramid pooling layer. SPP performs three maximum pooling operations of different sizes on the input and contacts the output results to splicing the output results, so that the output depth of the network is the same as the input depth. The MCA module is one of the improved methods in this paper. In the improved algorithm, the MCA module is embedded into the backbone feature extraction network to improve the algorithm's antibackground interference ability, focusing on the original algorithm's ability to extract the features of camouflaged human targets.

The Neck part is mainly composed of feature pyramid network FPN and path aggregation network PANet. The FPN and PANet structures realize the fusion and complementation of high-level features and low-level features, so that the network can obtain more original image information and more accurate target recognition. The FPN structure transfers the category features of the high-level large objects to the lower layers, and the PANet structure transfers the location features of the low-level large objects and the category and location features of the small objects upward. Among them, the lightweight general upsampling operator CARAFE (Content-Aware ReAssembly of FEatures) is one of the improved methods in this paper. In the improved algorithm, the CARAFE module is used to replace

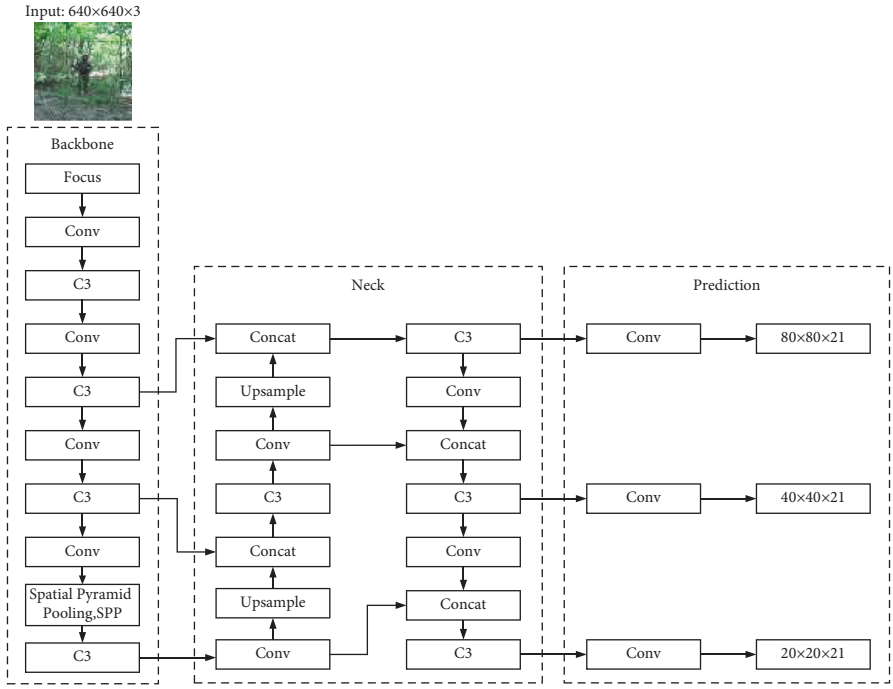

Figure 1: YOLOv5s detection framework.

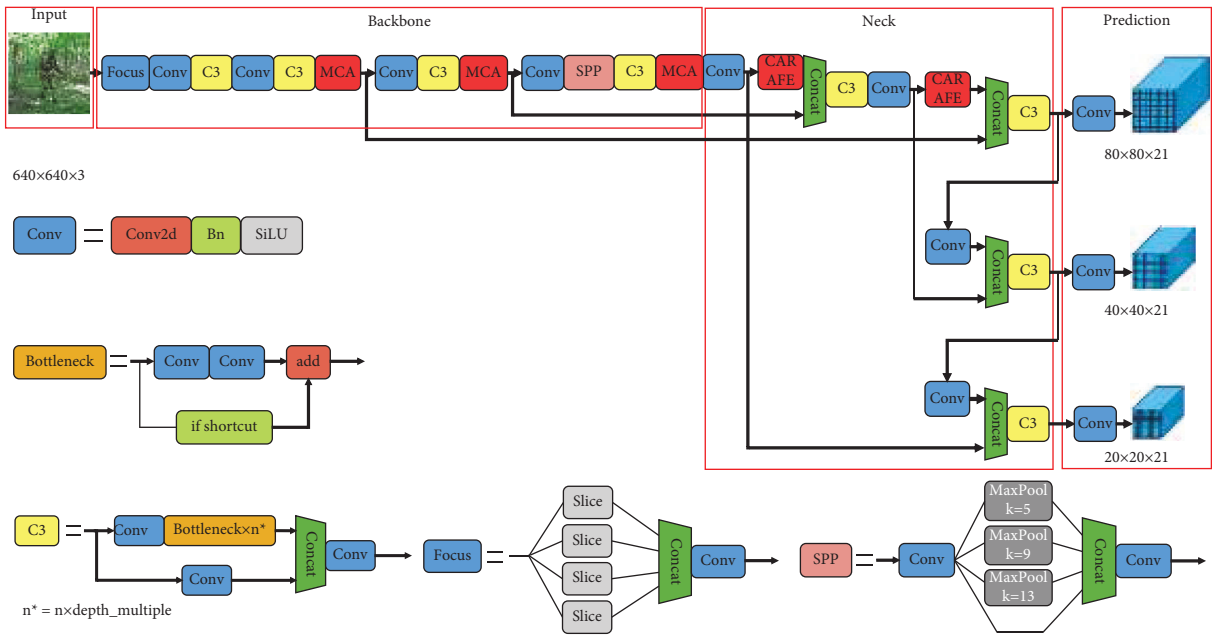

Figure 2: Network structure of MC-YOLOv5s.

the upsampling operation in the original algorithm, and the high-resolution low-level feature maps and low-level feature maps are strengthened. The fusion of high-resolution feature maps can avoid the loss of target feature information, thereby reducing the probability of false detection and missed detection.

The Prediction part is the detection structure of the improved algorithm. Since the input image size of the network in this paper is 640×640, and the downsampling is used 5 times in the improved algorithm, the final feature map size is 20×20, 40×40, 80×80, input it into the Detect module, which is used to identify large-, medium-, and small-scale targets, corresponding to 640×640, and the receptive field of each feature map is $640/20 = 32 \times 32$, $640/40 = 16 \times 16$, and $640/80 = 8 \times 8$ size.

2.2.1. Multispectral Channel Attention Module. The traditional channel attention module mainly lies in setting the weight function of various channel importances. For example, the channel attention module (Squeeze and

Excitation Network, SENet) [24] performs Global Average Pooling (GAP) on channels. The weight of each feature channel is automatically obtained by means of learning, and then, the features of the target area are strengthened and the background information irrelevant to the target is weakened according to the learned weight. If there is information loss, the GAP operation can only obtain general features. Although this selection is simple and efficient, GAP cannot fully obtain the rich semantic information of the target area, and it is easy to cause the problem of false detection and missed detection of overlapping occlusion targets.

The literature FcaNet [25] proves that GAP is a special form of discrete cosine transform (DCT), that is, the lowest frequency component, and only using GAP is equivalent to ignoring many other useful frequency components in the feature channel. Among them, the multispectral channel attention module (MCA) and other attentions have the same starting point, but the multispectral channel attention module not only retains the global average pool but also uses in addition to the global average pool. The frequency components can solve the problem of information loss caused by only focusing on a single frequency, make the network model to pay more attention to important features, and filter out redundant features. For this reason, in view of the problem that the military camouflaged human target is highly integrated with the surrounding environment background and has a low degree of discrimination, this study adds a multispectral channel attention module (MCA) to the YOLOv5s backbone feature extraction network to help the model effectively capture-rich feature information. More accurately, we identify and locate military camouflage personnel. The MCA module structure is shown in Figure 3.

The calculation of the general two-dimensional discrete cosine transform is shown in formula (1). Among them, $f_{h,w}^{2d}$ is the spectrum of the two-dimensional DCT, and $x_{i,j}^{2d}$ is the input; H and W represent the height and width of the input feature map, and the latter part is the DCT weight.

$$f_{h,w}^{2d} = \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right),$$

$$s.t.\ h \in \{0,1,\cdots,H-1\}, w \in \{0,1,\cdots,W-1\}. \tag{1}$$

Assuming that h and w in the 2-dimensional DCT are 0, formula (2) can be obtained. Among them, $\cos(0) = 1$, the left half is the lowest frequency component of the 2D DCT. Therefore, the following formula can prove that it is proportional to GAP.

$$\begin{aligned} f_{0,0}^{2d} &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} \cos\left(\frac{0}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{0}{W}\left(j+\frac{1}{2}\right)\right), \\ &= \sum_{i=0}^{H-1}\sum_{j=0}^{W-1} x_{i,j}^{2d} = \mathrm{gap}\left(x^{2d}\right)HW. \end{aligned} \tag{2}$$

In order to simplify the operation and facilitate the description, B is used to represent the weight component of the 2D DCT, namely,

$$B_{h,w}^{i,j} = \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right). \tag{3}$$

Combined with the above formula, the inverse transform of the 2D DCT is rewritten into the following form:

$$\begin{aligned} x_{i,j}^{2d} &= \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} f_{h,w}^{2d} \cos\left(\frac{\pi h}{H}\left(i+\frac{1}{2}\right)\right)\cos\left(\frac{\pi w}{W}\left(j+\frac{1}{2}\right)\right), \\ &= f_{0,0}^{2d}B_{0,0}^{i,j} + f_{0,1}^{2d}B_{0,1}^{i,j} + \cdots + f_{H-1,W-1}^{2d}B_{H-1,W-1}^{i,j}, \\ &= \mathrm{gap}\left(x^{2d}\right)HW \cdot B_{0,0}^{i,j} + + f_{0,1}^{2d}B_{0,1}^{i,j} + \cdots + f_{H-1,W-1}^{2d}B_{H-1,W-1,}^{i,j} \\ &s.t \quad i \in \{0,1,\cdots,H-1\}, j \in \{0,1,\cdots,W-1\}. \end{aligned} \tag{4}$$

It can be seen from the above formula that the features of the input image can be divided into combinations of different frequency components, and the GAP operation is only one of the frequency components. To introduce more information, the multispectral channel attention module adopts 2D DCT to fuse multiple frequency components. The specific operation is to first divide the input feature map X into n parts according to the number of channels, $C' = C/n$, where n must be divisible by the number of channels C. Then, each part has a two-dimensional DCT frequency component corresponding to it, and F represents the preprocessing result of the multispectral channel attention module, which can be expressed as follows:

$$\begin{aligned} F^i &= 2DDCT^{u,v}\left(X^i\right) = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{:\ h,w}^{i} B_{h,w}^{u,v} \\ &s.t \quad i \in \{0,1,\cdots,n-1\}, \end{aligned} \tag{5}$$

where $F^i \in R^{C'}$ is the preprocessed dimensional multispectral vector, and $[u, v]$ is the frequency component index corresponding to X^i. Then, the frequency components of different parts are combined, and the specific calculation is shown in

$$F = \mathrm{cat}\left(\left[F^0, F^1, \cdots, F^{n-1}\right]\right), \tag{6}$$

where cat represents the vector concatenation, and $F \in R^C$ represents the multi-spectral vector obtained. Finally, the entire MCA attention module can be represented by m, and its specific calculation is expressed as formula (7), where sigmoid and f_c represent the activation function and the mapping function, respectively.

$$m = \mathrm{sigmoid}\left(f_c(F)\right). \tag{7}$$

Since the MCA module can incorporate frequency components containing different kinds of information into the attention processing, more feature information can be extracted. With the deepening of the number of neural network layers, the feature information of the target to be detected is gradually lost. In this paper, considering the

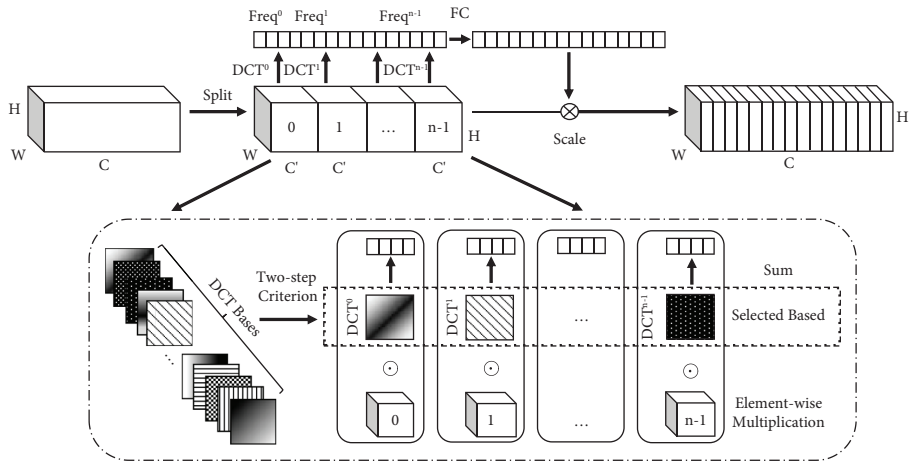

Figure 3: Multispectral channel attention module.

impact of different frequency components in the multispectral channel attention on the performance of the YOLOv5s detection framework, the first k frequency components with better performance are selected for comparative experiments, and the value of k is 4, 8, 16, or 32. Table 1 presents the ablation experiments of the parameters of the MCA module.

MAP in Table 1 represents the mean average precision. Two points can be found from the experimental data: first, the YOLOv5s model with multispectral channel attention (MCA) embedded in the Backbone part, and the detection performance is higher than the original YOLOv5s network model, which fully verifies the correctness of using multiple frequency components in channel attention. Second, when the frequency component in the MCA module is 16, the network can obtain the best detection effect. Therefore, the frequency components of the MCA module in the subsequent experiments in this paper are all taken as 16.

Table 1: Performance analysis of different frequency components in MCA module.

Number	Backbone	Frequency	mAP/%
1	Backbone		90.2
2	Backbone + MCA	4	90.7
3	Backbone + MCA	8	90.5
4	Backbone + MCA	16	92.9
5	Backbone + MCA	32	91.1

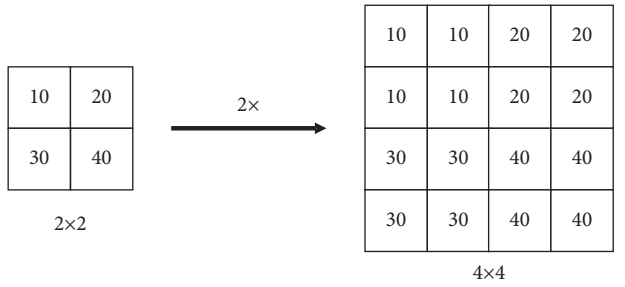

Figure 4: Nearest neighbor upsampling.

2.2.2. Upsampling Algorithm Improvements. The upsampling operation is an important step in the multiscale feature fusion network to fuse high-resolution low-level feature maps and low-resolution high-level feature maps. Effective upsampling operations can bring better multiscale prediction. The upsampling operation in the original YOLOv5s model is the method of nearest neighbor upsampling, as shown in Figure 4. The nearest neighbor upsampling algorithm uses the spatial position of the pixel to determine the upsampling kernel, only considers the subpixel neighborhood, and does not make full use of the target feature information. It can be regarded as a kind of "uniform" upsampling, and the perception range is general. Are very small, the size of the receptive field is only 1×1 , so a lot of feature map information is lost, and the rich semantic features required for camouflage target detection cannot be well captured.

Therefore, this paper replaces the original upsampling operation by introducing a lightweight general upsampling operator (Content-Aware ReAssembly of FEatures, CARAFE) [26], in order to strengthen the high-resolution low-level feature maps and low-resolution high-level feature maps. Fusion is to avoid the loss of target feature information, thereby reducing the probability of false detection and missed detection, and improving the detection accuracy of the algorithm for camouflaged targets in complex environments. CARAFE is mainly divided into two modules: the Kernel prediction module and the content-aware reassembly module.

The network structure of CARAFE is shown in Figure 5.

Assuming $X \in R^{C\times H\times W}$ is the input feature map, then the process of generating the target feature map $Y \in R^{C\times\sigma H\times\sigma W}$ can be divided into the following steps: first, for the $k_{\text{encoder}} \times k_{\text{encoder}}$ neighborhood $N(X_l, k_{\text{encoder}})$ of each position $l = (i, j)$ in the input feature map, we use the Kernel prediction module to predict the recombination kernel $W_{l'}$ of each l', and the specific calculation is as shown in the following formula:

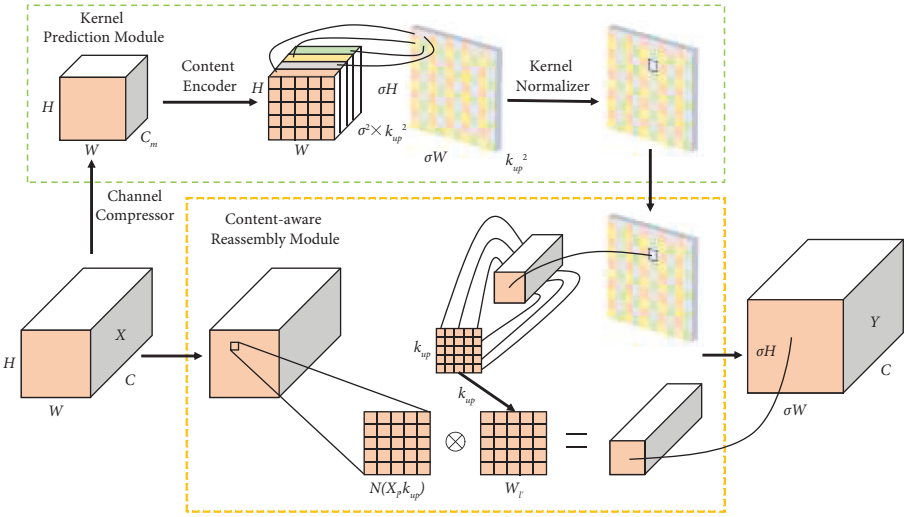

Figure 5: Network structure of CARAFE.

$$W_{l'} = \varphi\left(N\left(X_{l,}k_{\text{encoder}}\right)\right). \tag{8}$$

Then, the Kernel prediction module will use the 1×1 convolution to compress the number of channels of the input X from C to C_m, and the channel reduction will not only cause no loss of feature information but also reduce the amount of network parameters and computation. The specific operation of the Content Encoder is to use $k_{\text{encoder}} \times k_{\text{encoder}}$ convolution to convert the number of channels from C_m to C_{up}, $C_{up} = \sigma^2 \times k_{up}^2$, and σ is the upsampling multiple. Using a larger k_{encoder} will expand the receptive field to capture semantic information in a large range, but the network complexity increases with the increase of k_{encoder}. The default parameters selected in this paper are as follows: $\sigma = 2$, $k_{\text{encoder}} = 3$, and $k_{up} = 5$. The resulting feature map of $H \times W \times \sigma^2 \times k_{up}^2$ is then restructured into a feature map of $\sigma H \times \sigma W \times k_{up}^2$ and then normalized using the Softmax function for the channel values at each location.

Finally, the weighted summation is performed on each $k_{up} \times k_{up}$ square area $N(X_{upl}, k_{up})$ centered on $l = (i, j)$ in the feature map $\sigma H \times \sigma W \times k_{up}^2$ after the expansion to obtain the feature of the position $l'(i', j')$ of the new feature map. Then, we perform feature reorganization to obtain the output feature map, and the specific calculation is shown in the following formula, where $r = \lfloor k_{up}/2 \rfloor$.

$$Y_{l'} = \sum_{n=-r}^{r} \sum_{m=-r}^{r} W_{l'(n,m)} \cdot X_{(i+n,j+m)}. \tag{9}$$

2.2.3. K-Means++ Algorithm to Optimize Anchor Box. In YOLOv5, the method of adaptive anchor box is used to calculate the anchor box matching the target of the dataset. The core principle is to cluster all the target real boxes in the COCO dataset through the K-means clustering algorithm and finally get 9 cluster centers: (10, 13), (16, 30), (33, 23), (30, 61), (62, 45), (59, 119), (116, 90), (156, 198), and (373, 326), and each 3 is assigned to feature maps of large, medium, and small scales. The military camouflage personnel dataset constructed in this paper is mainly based on human targets, of which the target height is greater than the width of the vast majority. Obviously, the above a priori frame is difficult to meet the actual situation of human target detection, so the size of the a priori frame needs to be resized.

The K-means algorithm is a simple and practical clustering algorithm, which is easy to implement and has a good clustering effect. All values in the dataset were clustered according to known K values. The resulting cluster center is used to represent the center of each class and is obtained by averaging all the values in that class. However, the K-means algorithm has the following shortcomings: first, the K value needs to be given in advance, but this value is very difficult to estimate; second, the initial clustering center is randomly selected, and different initial values may produce different clustering effects. Once the initial value is not chosen well, it is likely to cause the algorithm to fail to converge, and eventually, only a local optimal solution can be obtained. The K-means++ clustering algorithm improves K-means. This method sets the distance between the initial cluster center points farther, which further reduces the number of operations, speeds up the algorithm, and can obtain more accurate clustering.

To this end, this paper uses the K-means++ clustering algorithm to re-optimize all the target real boxes in the military camouflage personnel dataset to improve the matching degree between the prior frame and the actual human target frame, and further accelerate the convergence speed of the network model. We improve the detection effect of military camouflaged human targets. In this paper, the 9 priori boxes generated by the K-means++ clustering algorithm are equally divided into 3 different prediction layers, and the size of the newly generated a priori boxes are shown in Table 2.

3. Dataset and Evaluation Indicators

3.1. Dataset. This paper uses the military camouflaged personnel dataset (MCPD) for experiments. The dataset mainly uses more than 60 military camouflage videos on the Internet in complex field environments as the original

Table 2: The priori box sizes.

Prior box size	Feature map scale		
	Big	Middle	Small
Anchor box 1	(32, 95)	(62, 241)	(96, 273)
Anchor box 2	(42, 138)	(75, 207)	(75, 207)
Anchor box 3	(53, 181)	(76, 29)	(143, 414)

material, and initially captures 10,000 high-definition images of camouflage human targets (size of 1280×720) through video frame capture. Camouflage patterns in many countries. The target detection technology based on deep learning mainly relies on three important factors: data, algorithm, and computing power, in which high-quality data are the premise for the algorithm to accurately identify the target. Due to the randomness in the video clipping process, it has a great impact on the overall quality of the images in the dataset. Therefore, in order to avoid the influence of the quality of the dataset on the algorithm itself, this paper considers many factors such as multidirectional, multi-angle and target size, attitude, and clarity to screen the dataset images layer by layer, and finally retains relatively high-quality military camouflage target image.

After strict screening, the military camouflage personnel dataset (MCPD) contains a total of 1000 high-definition images, each image contains 1 to 3 camouflaged human targets, and then uses the image labeling tool LabelImg to calibrate the human targets in each image one by one. There are two types of human target and environmental background. Several characteristics of the military camouflaged personnel dataset (MCPD) constructed in this paper: (1) the target has high concealment and low discrimination; (2) the target scales are different, including human targets of various scales such as large, medium, and small; (3) various target postures, including upright, half-squat, lying down, frontal, back, and sideways; (4) the environmental background is complex, including 6 kinds of wild environments such as jungle, rainforest, mountain, desert, snow, and urban ruins, and many factors such as season, weather, light, and occlusion are also considered. Some image samples in the MCPD dataset are shown in Figure 6.

3.2. Evaluation Indicators. This paper selects the following four indicators as the evaluation criteria of the model: precision (P), recall (R), mean average precision (mAP), model size (MB), and processing image frames per second number of FPS (frames per second). Model inference speed FPS (frame/s) is obtained by averaging the detection time of 200 images in the test set under the server Tesla P100 graphics card environment. Among them, the calculation formulas of P, R, and mAP are shown in (10)–(13). Among them, TP (true positive) refers to samples that were originally positive and were classified as positive; FP (false positive) refers to samples that were originally negative but classified as positive, and FN (false negative) refers to samples that were originally positive but are classified as negative class samples.

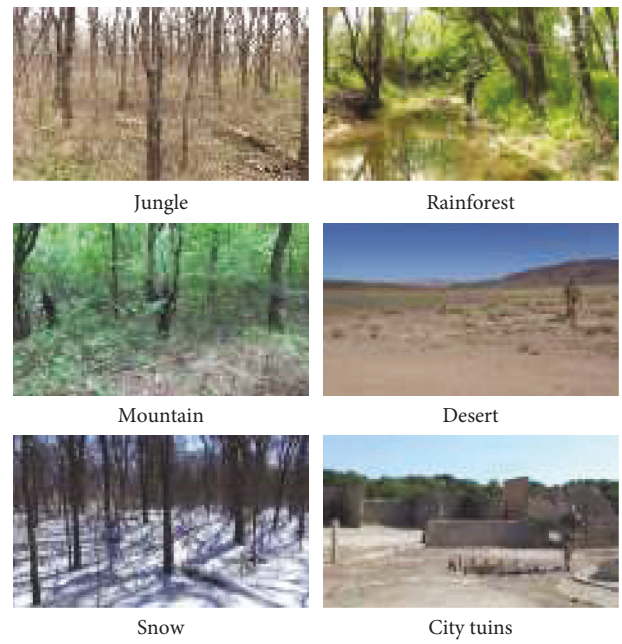

Figure 6: Military camouflage target sample.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \tag{10}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \tag{11}$$

$$AP = \frac{1}{n}\sum_{i=1}^{n} Pi = \frac{1}{n}P_1 + \frac{1}{n}P_2 + \ldots + \frac{1}{n}P_n, \tag{12}$$

$$mAP = \frac{1}{C}\sum_{k=1}^{C} AP_k. \tag{13}$$

In the above formula, the definition of accuracy P is as follows: the proportion of samples that are actually positive and predicted to be positive to all predicted positive samples, P pays more attention to the situation of wrongly classifying negative samples as positive samples, and the definition of recall R is as follows: actually, it is the proportion of samples that are positive and predicted to be positive to all actual positive samples. R pays more attention to classifying positive samples as negative samples. The average precision (AP) refers to the average of all the accuracies (i.e., the area under the PR curve) obtained under all possible values of recall. Mean average precision (mAP) refers to the mean of AP for each category. In formula, C (10) represents the number of classes, and k represents one class.

4. Results and Discussion

4.1. Experimental Environment. The experiments in this paper are based on the official open source project of YOLOv5 (version 5.0) using the YOLOv5s model as the basic configuration, the computer operating system is Ubuntu 18.04, the *Python* version is 3.8, the deep learning

framework is Pytorch1.8.0, and the CPU processor is Intel (R) Xeon(R) Gold 6130 (memory size is 31 GB). All experiments were performed on a Tesla P100 (15 GB video memory), and CUDA (compute unified device architecture) and cuDNN (CUDA deep neural network library) were used for accelerated training in the experiments to improve the computing power of the computer.

4.2. Network Training. Before training, the military camouflaged personnel dataset (MCPD) is randomly divided into training set, validation set, and test set with a ratio of 6 : 2:2. In order to fully train the improved network model and prevent the neural network from overfitting, a variety of data enhancement methods are used to expand the dataset during training, including color transformation, left-right flip, translation zoom, and mosaic.

The number of training times, Epochs, is set to 100, the batch size is set to 32, the size of the input image is 640×640, the initial learning rate is 0.0001, the learning rate momentum and the weight decay coefficient are set to 0.937 and 0.0005, respectively, and the optimization strategy is Adam (adaptive moment estimation) [27] optimizer. After outputting the prediction results, the nonmaximum suppression algorithm is used to screen the prediction boxes.

4.3. Experiment Verification and Result Analysis. In this paper, two groups of experiments are designed for verification, which are different improved partial ablation experiments and different algorithm model comparison experiments.

4.3.1. Ablation Experiment. In order to test the detection effect of the improved algorithm, an ablation comparison experiment was carried out: the MCA module, the CARAFE module, and the K-means++ algorithm were introduced into the original YOLOv5s model, respectively. The comparison of the ablation results is shown in Table 1. Among them, "√" indicates that the corresponding improvement method is used in the network model, and "×" indicates that the corresponding improvement method is not used in the network model.

From the data in Table 3, it can be seen that the introduction of multispectral channel attention (MCA) in the backbone feature extraction network of YOLOv5s increases the model size by only 18.7 MB. The precision (P), recall (R), and mean average precision (mAP) is increased by 0.7%, 0.6%, and 2.6%, respectively, indicating that the MCA module can enhance the attention to the target area of interest in complex environments, and effectively improve the original YOLOv5 algorithm network without attention preference and and effectively improve the problem that the original YOLOv5 algorithm network has no attention preference which leads to the loss of camouflage target feature information. The lightweight general upsampling operator CARAFE is introduced to replace the original upsampling operation, so as to strengthen the fusion of high-resolution low-level feature maps and low-resolution high-level feature maps, effectively solve the problem of false detection and missed detection, and make mean average precision (mAP) of the algorithm is increased by 0.7%, which effectively improves the detection accuracy. On the basis of the above improvements, the K-means++ clustering algorithm is used to re-optimize all the a priori boxes of the dataset, so that the size of the a priori boxes is more suitable for the real box of the human target, and the experimental data in the table show that the anchor frame optimization improves the detection effect of the network model to a certain extent. Compared with the original algorithm, the combination of the three has greatly improved the evaluation indicators, which fully shows that in the detection task of camouflaged human targets, adding the multispectral channel attention module can effectively weaken the complex background information in the picture and improve the resistance of the network. Background interference ability and the introduction of a lightweight general upsampling operator CARAFE can further strengthen the full fusion of low-level and high-level feature maps, so as to achieve a good detection effect. In summary, the improved algorithm achieves accurate detection and recognition of camouflaged human targets in complex environments.

4.3.2. Comparison of Different Algorithm Models. In order to further verify the detection effect of the improved method in this paper, the improved algorithm is compared with several mainstream object detection models: RetinaNet, YOLOX, YOLOv5s, YOLOv5m, and YOLOv5l. The dataset and platform configuration conditions used in the experiment are the same, and the performance comparison results are shown in Table 4. The bold font is the optimal value of the project.

From the data in Table 4, it can be seen that the model size of the improved algorithm is only 33.1 MB, which is 18.7 MB higher than the original YOLOv5s algorithm, and the detection speed FPS is reduced by 31.2. However, the detection accuracy of the improved model for military camouflage personnel is improved by 3.7%. Compared with the four excellent algorithms such as RetinaNet, YOLOX, YOLOv5m, and YOLOv5l, the MC-YOLOv5s algorithm model is more lightweight and is superior to the appeal algorithm in terms of detection accuracy and inference speed. Overall, the improved algorithm has the best detection performance for camouflaged human targets.

4.4. Test Results. Figure 7 shows the detection results of camouflage personnel under typical complex backgrounds by five algorithms, including MC-YOLOv5s, YOLOv5s, YOLOX, YOLOv5m, and YOLOv5l. From the detection results in the figure, it can be seen that the improved algorithm has higher detection accuracy, can effectively

Table 3: Comparison of ablation experiment results.

Model	MCA	CARAFE	K-means++	P/%	R/%	mAP/%	Size/MB
YOLOv5s	×	×	×	94.1	83.3	90.3	14.4
YOLOv5s	√	×	×	94.8	83.9	92.9	32.8
YOLOv5s	×	√	×	88.4	87.6	92.4	14.6
YOLOv5s	×	×	√	92.5	85.7	90.8	14.4
YOLOv5s	√	√	×	92.3	88.5	93.6	33.1
YOLOv5s	√	√	√	97.4	86.1	94	33.1

Table 4: Detection performance of different models.

Module	mAP/%	FPS	Size/MB
RetinaNet	81.4	6.4	138
YOLOX	88.7	13.4	34.3
YOLOv5s	90.3	**125**	**14.4**
YOLOv5m	92.6	45.5	40.4
YOLOv5l	93.4	43.5	93.7
MC-YOLOv5s	**94**	93.8	33.1

Figure 7: Comparison of detection results of different models.

capture the feature information of camouflage human targets, and has better detection performance.

5. Conclusions

In this paper, in the task of camouflaged human target detection in complex battlefield environments, there are problems that the target and the complex environmental background are highly integrated, and the detection effect is poor due to low discrimination. A human target detection algorithm MC-YOLOv5s based on the characteristics of camouflage is proposed. The algorithm introduces the multispectral channel attention module (MCA) in the backbone part of the backbone feature extraction network in YOLOv5s, which improves the network's full extraction of the feature information of camouflaged human targets and

strengthens the ability to resist background interference; second, the lightweight general upsampling operator CA-RAFE is used to replace the upsampling operation in the original algorithm, and the high-resolution low-level feature map is further strengthened and the fusion of low-resolution high-level feature maps, reducing the false detection rate of the original algorithm. Finally, the K-means++ method is used to cluster the actual boxes of all objects in the dataset images to obtain the sizes of the more matching prior boxes. The combination of the three effectively solves the problem that the original YOLOv5s algorithm is difficult to fully extract the characteristics of camouflaged personnel and effectively integrate the underlying and high-level feature maps. At the same time, it improves the matching degree between the prior frame and the actual target frame, and strengthens the detection ability of the detection network to camouflage human targets in complex environments. The experimental results show that the MC-YOLOv5s algorithm has excellent detection effect on the MCPD dataset, and has stronger generalization ability. Compared with the original YOLOv5s algorithm model, the precision (P), recall (R), and mean average precision (mAP) are increased by 3.3%, 2.8%, and 3.7%, respectively. Detecting camouflaged personnel images in complex and changeable battlefield environments is better than algorithms such as RetinaNet, YOLOX, YOLOv5m, and YOLOv5l and significantly improves the two difficult problems of camouflaged human target detection mentioned above. If the proposed MC-YOLOv5s is applied to personnel search and rescue in complex battlefields and natural disaster environments, the efficiency of personnel search and rescue and the survival rate of life can be greatly improved, and the consumption of manpower and material resources in the rescue process can also be greatly reduced.

Data Availability

No data were used to support this study.

Conflicts of Interest

The authors declare that there are no conflicts of interest.

References

[1] H. M. Wang, X. G. Wang, and X. Y. Wang, "Research on target detection under complex background based on deep learning," *Control and Decision*, vol. 60, no. 5, pp. 1–8.

[2] X. R. Zhang, F. Xiang, H. J. Li, C. Zhang, S. C. Zhou, and Y. Zuo, "Design of Steel Coil End Face Defect Detection System Based on Deep Learning," *Computer Integrated Manufacturing System*, pp. 1–19, https://kns.cnki.net/kcms/detail/11.5946.TP.20220306.1105.002.html.

[3] Y. Wang, T. Y. Cao, and J. B. Yang, "Camouflaged object detection based on improved YOLOv5 algorithm," *Computer Science*, vol. 48, no. 10, pp. 226–232, 2021.

[4] G. Qi, Y. Zhang, K. Wang, and M. Neal, "Small object detection method based on adaptive spatial parallel convolution and fast multi-scale fusion," *Remote Sensing*, vol. 14, no. 2, p. 14, 2022.

[5] N. M. Alfaer, H. M. Aljohani, S. A. Khalek, S. A. Abdulaziz, and F. M. Romany, "Fusion-based deep learning with nature-inspired algorithm for intracerebral haemorrhage diagnosis," *Journal of Healthcare Engineering*, vol. 2022, 2022.

[6] X. Y. Liang, H. K. Lin, and H. Yang, "Construction of semantic segmentation dataset of camouflage target image," *Lasers and Optoelectronics Progress*, vol. 58, no. 04, pp. 214–220, 2021.

[7] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," *Proc Computer Vision & Pattern Recognition*, vol. 1, pp. 886–893, 2005.

[8] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and Their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[9] P. F. Felzenszwalb, R. B. Girshick, D. Mcallester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[10] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: towards real-time object detection with region proposal networks," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 6, pp. 1137–1149, 2017.

[11] W. Liu, D. Anguelov, D. Erhan et al., *SSD: Single Shot Multibox detector*, Springer, Cham, 2016.

[12] M. Tan and Q. V. Le, "EfficientNet: Rethinking Model Scaling for Convolutional Neural networks," 2019, https://arxiv.org/abs/1905.11946.

[13] T. Y. Lin, P. Goyal, and R. Girshick, "Focal loss for dense object detection[J]," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, no. 99, pp. 2999–3007, 2017.

[14] J. Redmon, S. Divvala, R. Girshick, and F. Ali, "You Only Look once: Unified, Real-Time Object detection," IEEE, 2016, https://arxiv.org/abs/1506.02640.

[15] J. Redmon and A. Farhadi, "YOLO9000: Better, Faster, stronger," pp. 6517–6525, IEEE, 2017, https://arxiv.org/abs/1612.08242.

[16] J. Redmon and A. Farhadi, "YOLOv3: An Incremental improvement," arXiv e-prints, 2018, https://arxiv.org/abs/1804.02767.

[17] A. Bochkovskiy, C. Wang, and H. M. Liao, "YOLOv4: Optimal Speed and Accuracy of Object detection," 2020, https://arxiv.org/abs/2004.10934.

[18] Q. D. Wang, T. Q. Chang, and L. Zhang, "Automatic detection and tracking system of tank armored targets based on deep learning algorithm," *Systems Engineering and Electronic Technology*, vol. 40, no. 09, pp. 2143–2156, 2018.

[19] B. W. Yu and M. Lv, "Application of improved YOLOv3 algorithm in military target detection," *Journal of Ordnance Engineering*, pp. 1–10.

[20] X. T. Deng, T. Y. Cao, and Z. Fang, "Research on detection of people with camouflage pattern via improving RetinaNet," *Computer Engineering and Applications*, vol. 57, no. 05, pp. 190–196, 2021.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[22] T. Y. Lin, P. Dollar, R. Girshick, H. Kaiming, H. Bharath, and B. Serge, "Feature pyramid networks for object detection," in *Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 936–944, IEEE, Honolulu, HI, USA, June 2017.

[23] S. Liu, L. Qi, H. F. Qin et al., "Path aggregation network for instance segmentation[C]//2018 IEEE," *CVF Conference on*

Computer Vision and Pattern Recognition, pp. 8759–8768, IEEE Press, Salt Lake, 2018.

[24] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7132–7141, IEEE, Piscataway, 2018.

[25] Z. Qin, P. Zhang, F. Wu, and L. Xi, "FcaNet: Frequency Channel Attention Networks," https://arxiv.org/abs/2012.11879v4.

[26] J. Wang, K. Chen, R. Xu, L. Ziwei, L. C. Change, and L. Dahua, "CARAFE: Content-Aware Reassembly of features," IEEE, 2020, https://arxiv.org/abs/1905.02188.

[27] D. P. Kingma and B. A. J. Adam, "A method for stochastic optimization," 2014, https://arxiv.org/abs/1412.6980.