Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9110292, 10 pages
https://doi.org/10.1155/2022/9110292

Research Article

Research on the Detection Technique of Situation Elements in Obscure Overlapping Scenes

Jinlong Liu and Kangda Cheng

Harbin Institute of Technology, Harbin 150000, China

Correspondence should be addressed to Jinlong Liu; yq20@hit.edu.cn

Received 17 February 2022; Accepted 14 March 2022; Published 11 April 2022

Academic Editor: Mingqian Liu

In recent years, some scholars have proposed to apply single-stage target detection algorithms such as YOLO (You Only Look Once) to situational element detection, but the traditional YOLO algorithm is to treat the target detection process as a regression problem, which cannot distinguish well between overlapping objects and has defects such as less accurate bounding boxes and hard to distinguish objects from the background, and it is difficult to cope with problems such as the higher overlap of targets to be detected and stronger target camouflage ability in obscured overlapping scenes. In this paper, we propose to add the attention module CBAM to the backbone network of the YOLOv3 model, to construct a SEDNet with high accuracy and good robustness for situational element detection, and to apply it to the situational element detection in occlusion overlapping scenes. We use SEDNet to classify and localize ten elemental targets, respectively. The analysis of experimental results shows that the SEDNet target detection model can complete element detection in complex environments with strong target camouflage, achieve end-to-end detection, and lay the technical foundation for the formation of complete situational awareness.

1. Introduction

Situational awareness was first used to study pilots' awareness and understanding of their flight environment and flight status and to help train pilots to develop the ability to judge and react correctly to current and future flight situations over time [1]. Later, it was widely used in decision-making, network information security, system supervision, etc. In 1995, Endsley defined situational awareness as the awareness of situational elements within a certain time and space environment and the understanding of the information obtained, which leads to the formation of a prediction of the state of these situational elements at the next moment [2]. This definition divides situational awareness into three phases, awareness, understanding, and prediction, which are shown schematically in Figure 1. Among them, situational element detection is the basis for forming situational awareness.

The traditional target detection network is difficult to accomplish the task of situational target detection in the context of high overlap of targets to be detected and strong target camouflage capability and is not applicable to the problem of situational awareness in obscured overlapping scenarios.

In 2019, Peng et al. [3] introduced the YOLO (You Only Look Once) algorithm to situational element identification and localization and achieved better results. However, because the YOLO algorithm treats the target detection process as a regression problem, it cannot distinguish overlapping objects well and suffers from deficiencies such as the lack of accurate bounding boxes and difficulty in distinguishing objects from the background, which makes it difficult to cope with the problems of high overlap of detected targets and strong camouflage of targets to be detected in obscured overlapping scenarios [4].

Attention mechanism may be the key to solving this problem. Woo et al. [5] showed that attention mechanisms could widely and effectively improve the performance of convolutional neural networks. Chengji et al. [6] introduced the attention mechanism into the YOLO model and

demonstrated that it could effectively improve the detection accuracy of the YOLO model. This paper conducts an exploratory study on this topic. This paper first describes the detection principle of YOLOv3 and its network structure, then describes the principle and structure of an attention module, then proposes a target detection network SEDNet based on the combination of YOLOv3 and CBAM attention module specifically for solving the situational element detection problem, and improves the loss function and training strategy. Finally, this paper constructs a SA situational element dataset and conducts training and validation experiments on the SA dataset. The experimental results show that SEDNet can complete the identification and localization of situational elements in real time effectively, which provides effective technical support to solve the situational element detection problem.

2. Related Works

2.1. The Network Structure of Baseline Model YOLOv3. YOLO [7] is a single-stage target detection model that treats the target detection problem as a regression problem with target region prediction and category prediction. YOLO enables real-time, high-precision detection that identifies the location of objects in an image and their categories at a glance. The method uses a single neural network to directly predict the bounding box and category probability of a target, which enables end-to-end target detection. At the same time, this method is much faster than other target detection methods (e.g., Fast R-CNN) and is more suitable for obscured overlapping scenarios.

Figure 2 gives a schematic diagram of the YOLO target detection model. As can be seen from the figure, when the YOLO model performs target detection, the original image is first divided into cells, and if the center point of the target to be detected is in a cell, then that cell is responsible for detecting this target. Each cell needs to detect a bounding box and a category probability. Each bounding box contains the information of x, y, w, and h and confidence, which are the center coordinates of the bounding box and are the length and width of the bounding box, respectively, and the confidence indicates the probability of predicting the target category and the accuracy of the target at that location. Based on the target categories in each cell and the corresponding bounding information, the location information, center coordinates, length, and width of the whole segmented area where each target is located can be calculated. The target windows with low probability are then removed according to the threshold value, and the redundant windows are removed by nonmaximal suppression, leaving the detection results.

The network structure of YOLOv3 [8] is mainly composed of two parts, Darknet-53 feature extraction network, and feature pyramid, with 75 convolutional layers and 3 detection layers, and its network structure is shown in Figure 3, where Inputs denotes input images, Conv2D denotes performing a 2D convolution operation, Residual Block denotes residual module Concat denotes concatenating the data, and UpSampling2D denotes performing an upsampling operation opposite to the pooling operation.

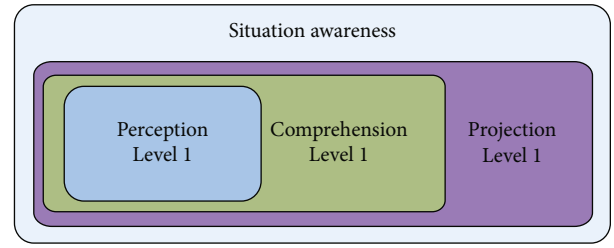

Figure 1: Schematic of the Endsley situational awareness model.

As shown in Figure 3, YOLOv3 uses Darknet53 as its backbone feature extraction network, whose main feature is the residual network, which is easy to optimize and can improve the accuracy by increasing the network depth, and its internal residual blocks use jump connections to alleviate the gradient disappearance problem caused by increasing the network depth in deep neural networks. Each convolution part of Darknet53 uses the unique DarknetConv2D structure and performs L2 regularization operation at each convolution and batch normalization and Leaky ReLU operation after completing the convolution. The mathematical expression of Leaky ReLU is shown.

$$y_i = \begin{cases} x_i, & \text{if } x_i \geq 0, \\ \dfrac{x_i}{\alpha_i}, & \text{if } x_i < 0. \end{cases} \tag{1}$$

In the YOLOv3 target detection model, the image input is first subjected to feature extraction through a five times downsampling process. YOLOv3 uses multiscale prediction to extract three feature layers for target detection, which are located in the middle layer, lower-middle layer, and bottom layer of the Darknet53. After completing the feature layer extraction, the three feature layers are processed by five convolutions, and the processed data are used to output the prediction results corresponding to this feature layer on the one hand and are used for upsampling after the deconvolution operation with UmSampling2d on the other hand, and the upsampled results are stacked with another feature layer to construct a feature pyramid, which can be used for multiscale feature extraction to obtain more effective features.

2.2. The Loss Function of Baseline Model YOLOv3. The loss function of YOLOv3 can be divided into three parts: coordinate error loss function, IoU error loss function, and classification error loss function.

$$L = L_1 + L_2 + L_3, \tag{2}$$

where L denotes the total loss function in the training process, L_1 denotes the coordinate error loss function, L_2 denotes the IoU error loss function, and L_3 denotes the

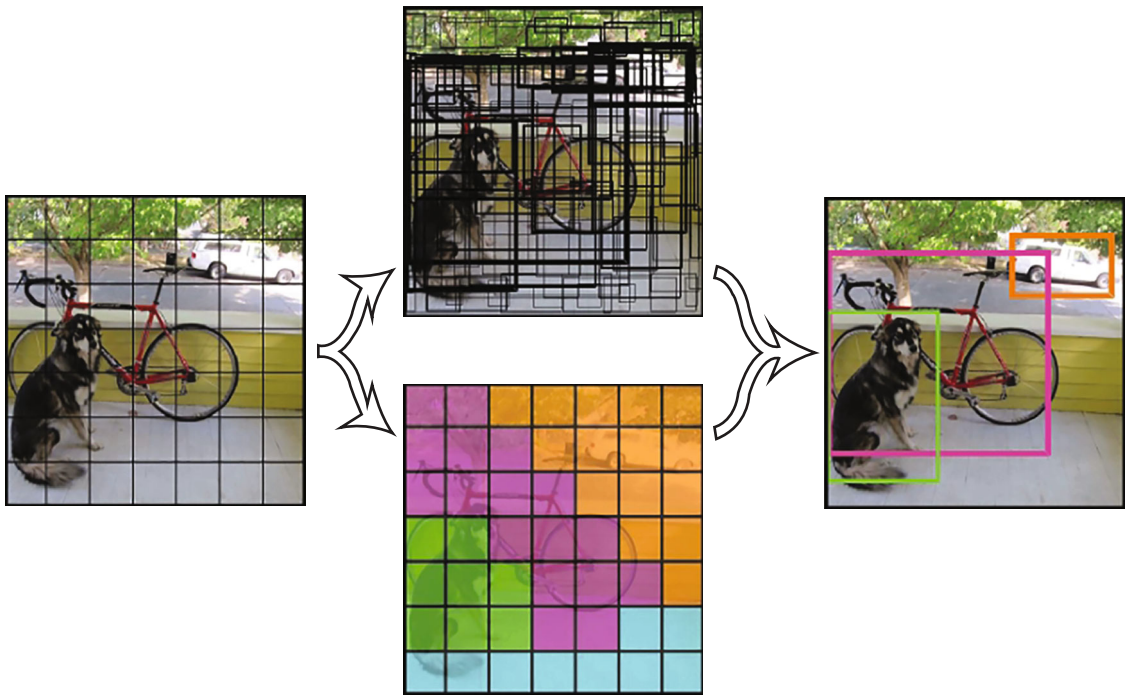

Figure 2: YOLO target detection model detection process schematic.

classification error loss function. L_1 can be calculated using

$$L_1 = \lambda_{\text{coord}} \sum_{i=0}^{s^2} \sum_{j=0}^{B} I_{ij}^{\text{obj}} \left[(x_i - x_i)^2 + (y_i - y_i)^2 \right] + \lambda_{\text{coord}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{\text{obj}} \left[\left(\sqrt{w_i} - \sqrt{\widehat{w}_i} \right)^2 + \left(\sqrt{h_i} - \sqrt{\widehat{h}_i} \right)^2 \right], \tag{3}$$

where x_i, y_i, w_i, and h_i are the coordinate values and target proportions of the targets in the prediction frame, respectively, and $\widehat{x}_i, \widehat{y}_i$, $\widehat{w}_i$, and $\widehat{h}_i$ denote the coordinate values and target proportions of the targets in the actual frame, where $\lambda_{\text{coord}} = 0.5$; I_{ij}^{obj} indicates that the target has no target in the prediction frame of the ith raster. The square root of w and h is used to reduce the effect of w and h on the size targets.

IoU is the intersection-over-union of the prediction frame and the label frame, and the IoU error loss function L_2 can be calculated using

$$L_2 = \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{\text{abj}} (C_i - C_i)^2 + \lambda_{\text{noobj}} \sum_{i=0}^{S^2} \sum_{j=0}^{B} I_{ij}^{\text{noobj}} (C_i - C_i)^2, \tag{4}$$

where C is the set of all categories, c is an element in C that denotes a specific category, and $Pi(C)$ denotes the confidence probability of the category on the ith cell. In addition, for the specificity of images, YOLOv3 adds two data scaling methods, scaling and rotation-based contrast enhancement and Gaussian white noise, which can mitigate the overfitting problem.

In the detection process, the image is passed through the YOLOv3 target detection model to obtain the feature map, and each position on the feature map is used to identify the attributes and positions of the targets through prediction frames with different aspect ratios and scales. The prediction frame is a clustering selection of regions with real labeled targets in the training set, and the dimension of the cluster center is selected as the dimension of the prediction frame.

YOLOv3 can perform target detection in real time, but the traditional YOLO model cannot well separate the foreground region from the background region and has difficulties in distinguishing overlapping objects. In situation element detection tasks, the target to be detected is often camouflaged according to the environment, causing the target to overlap with the environmental background or the target and the environmental background to be highly similar, so the classical YOLOv3 target detection model is difficult to be applied to the situation element detection problem.

Recent studies have shown that adding an attention mechanism to convolutional neural networks (CNNs) can exchange a very small increase in computational effort for a significant increase in accuracy. Therefore, to solve the above problem, this paper introduces the attention mechanism in the Darknet53 backbone network of YOLOv3 and constructs the Situation Element Detection Network (SED-Net) based on the attention mechanism.

2.3. Attention Mechanisms. The attention mechanism [9] is originally a mechanism unique to human vision, which is

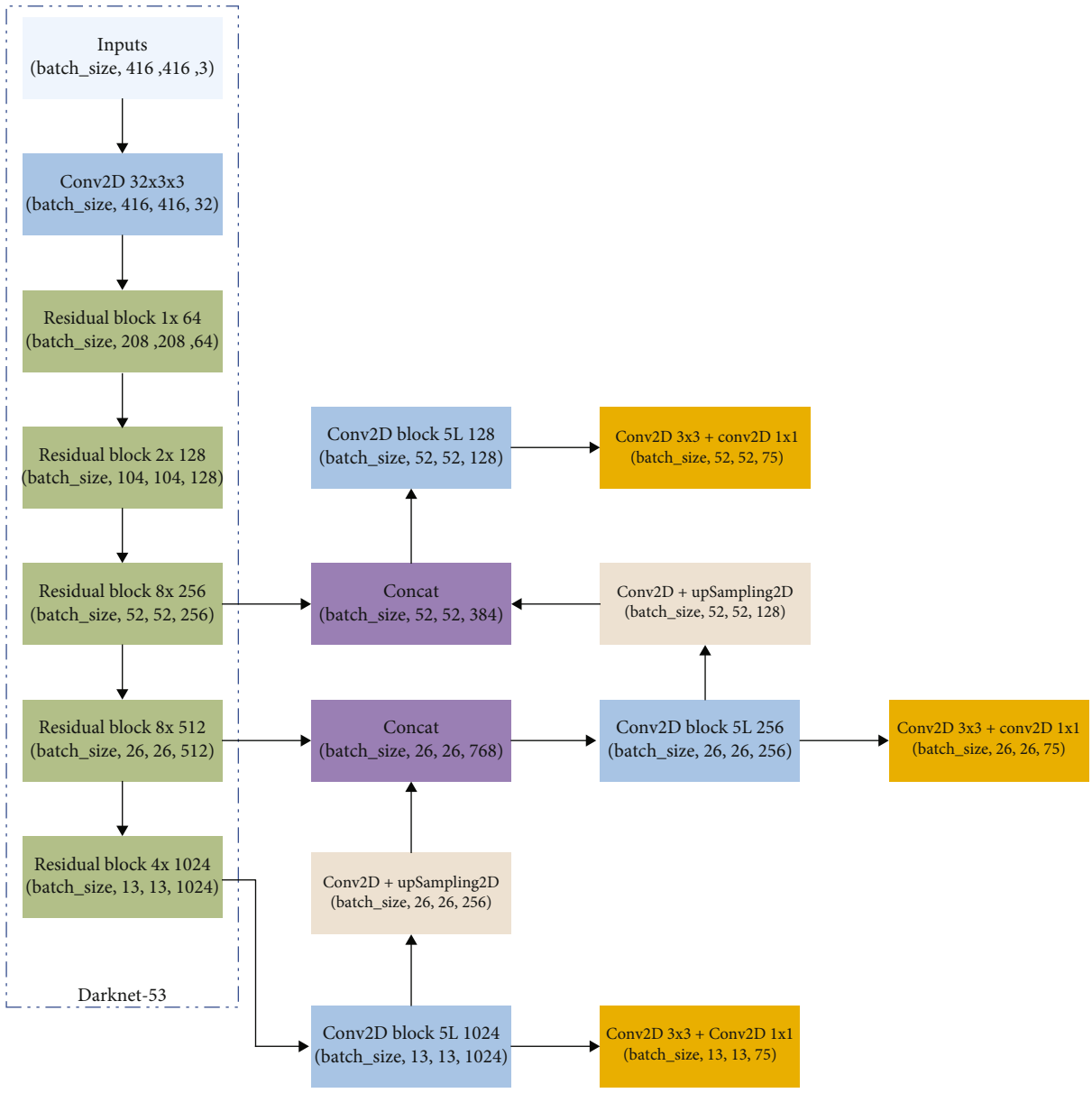

Figure 3: YOLO target detection model structure schematic.

the pointing and focusing of mental activity and consciousness on certain objects, and its basic function is to filter the information of important elements in the environmental elements. When humans observe a global image, they quickly identify the target area that needs attention and then devote more attention resources to this area and suppress information from other useless areas to obtain more detailed information about the target area that needs attention; this is the ability of humans to use limited attention resources to quickly filter out high-value information from a large amount of information. Attention mechanism greatly improves the accuracy and efficiency of human visual information processing, which is a unique advantage of humans.

The attention mechanism in the field of deep learning is borrowed from the attention mechanism of human vision by quickly scanning the global image, acquiring the target region that needs to be focused on, and then devoting more computational resources to that region to obtain more detailed information to be captured and suppressing the useless information in other regions.

Attention mechanisms can extract key local information from global information and can be easily embedded in existing network architectures to help extract key local features quickly. It can overcome the problems of difficult target localization, difficult target classification, and background interference caused by the detection of camouflaged targets in the situational awareness process, so it is important to conduct research on attention mechanisms in the field of situational awareness.

Recent studies have confirmed that attention mechanisms have great potential to improve the performance of deep convolutional neural networks (CNNs), and how to introduce attention modules into convolutional neural networks has attracted a lot of attention. There is no unique

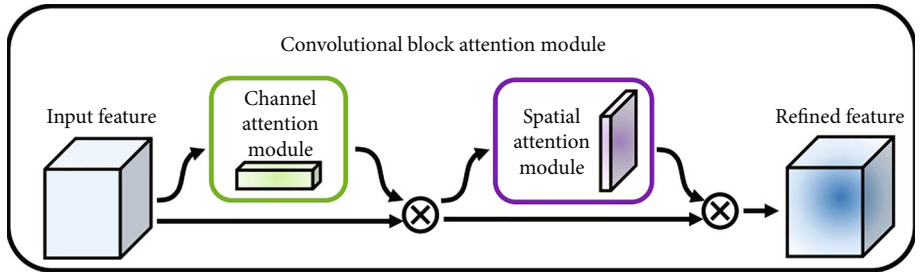

Figure 4: Convolutional block attention module structure schematic.

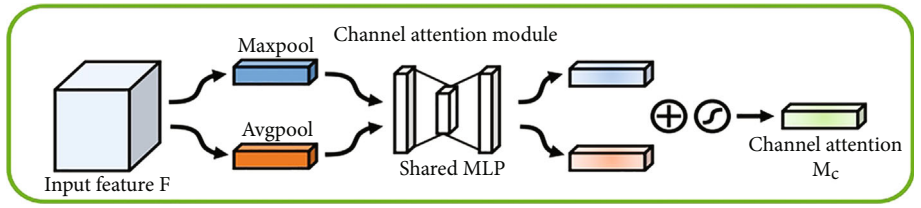

Figure 5: Schematic diagram of CBAM's channel attention module structure.

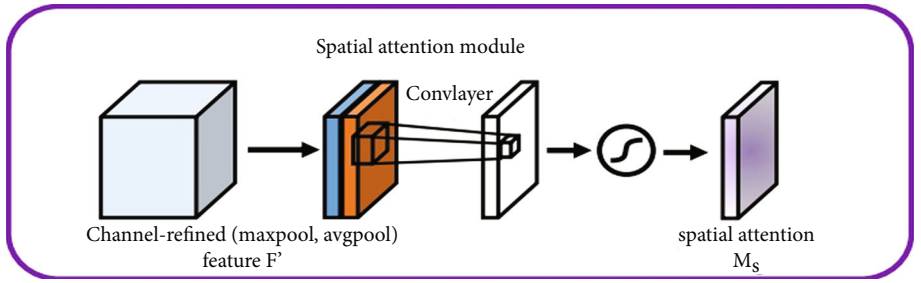

Figure 6: Schematic diagram of CBAM's spatial attention module structure.

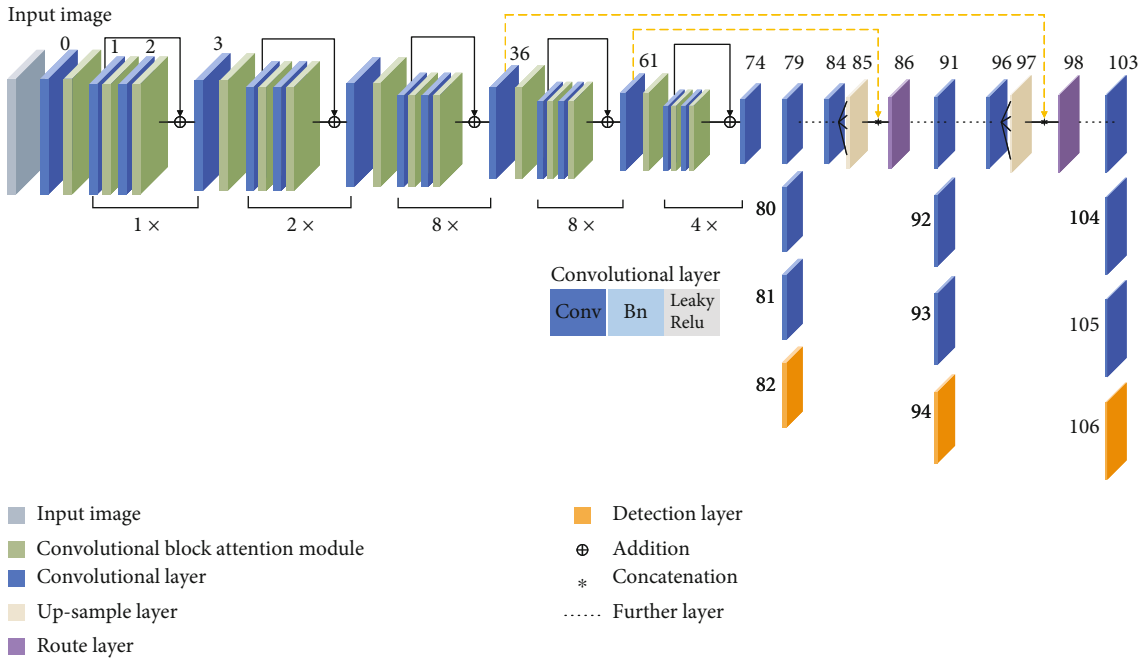

Figure 7: Schematic diagram of the overall network structure of the SEDNet model.

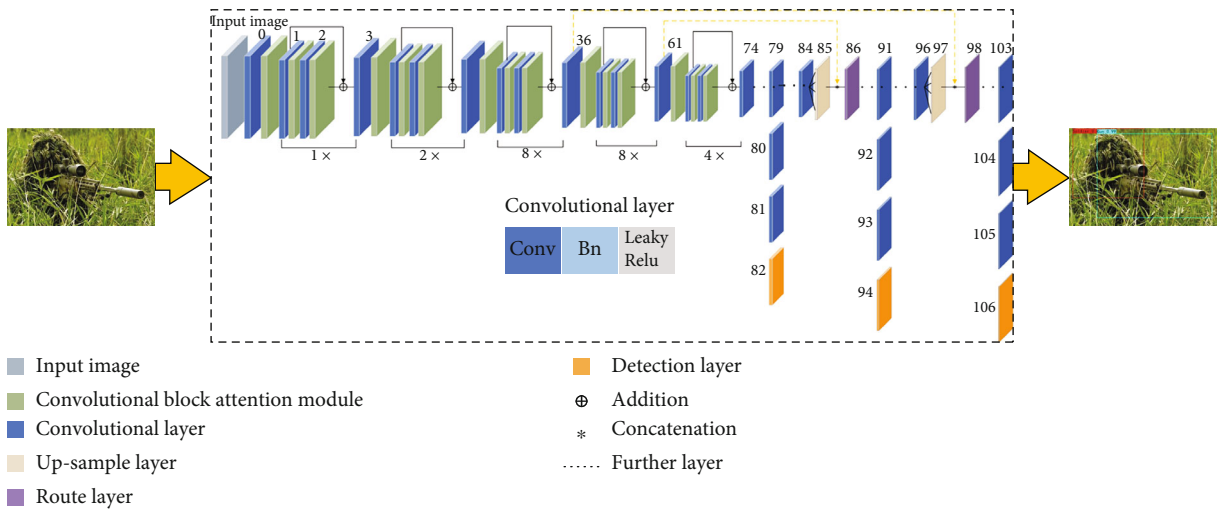

Figure 8: Situational element awareness model diagram.

Table 1: Comparison of the average accuracy rates of YOLOv3, SE-YOLO, and SEDNet.

	YOLOv3(%)	SE-YOLO(%)	SEDNet(%)
Submarine	89.09	94.10	99.12
UAV	77.20	80.00	92.93
Dog	90.03	90.86	90.65
Tank	65.71	79.87	89.20
Warship	72.96	83.10	87.92
Fighter	72.52	76.58	82.46
Soldier	61.34	73.63	80.38
Helicopter	55.98	75.33	78.49
Truck	48.87	67.41	72.55
Gun	38.53	48.90	64.64
mAP	67.22	76.98	83.83

way to introduce attention mechanisms into convolutional neural networks, either by adding attention mechanisms in the channel dimension or by adding attention mechanisms in the spatial dimension or by adding attention mechanisms in both the spatial and channel dimensions.

In 2018, Hu et al. [10] proposed the classical Squeeze-Excitation Network with feature compression and featured weight adjustment in the feature channel dimension. The results of applying it on the ResNet network showed that the technique not only significantly reduces the classification error rate but also effectively improves the mAP.

In 2019, Redmon et al. [7] introduced the attention mechanism into the YOLO algorithm, using weighted and filtered feature vectors to replace the original feature vectors for residual fusion and reducing the information loss in the fusion process by adding second-order terms. Experimental results on COCO and PASCAL VOC datasets show that the introduction of the attention mechanism in the YOLO model can effectively reduce the bounding box localization error and improve the detection accuracy.

Convolutional block attention module (CBAM) [7] is an attention mechanism that combines two dimensions of feature channels and feature spaces. Previous work has shown that CBAM can improve the performance of various deep CNN architectures, and it has shown good generalization ability in target detection and instance segmentation tasks. As shown in Figure 4, the CBAM module consists of a channel attention module and a spatial attention module.

CBAM uses not only average pooling to compress the feature map in the spatial dimension but also employs maximum pooling as a compliment. The structure diagram of the channel attention module of CBAM is shown in Figure 5, where MaxPool and AvgPool denote performing maximum pooling operation and average pooling operation, respectively, and MLP denotes multilayer perceptron.

The channel attention module of CBAM computes the feature mapping as shown in

$$M_c(F) = \sigma(\mathrm{MLP}(\mathrm{AvgPool}(F)) + \mathrm{MLP}(\mathrm{MaxPool}(F))), \quad (5)$$

$$M_c(F) = \sigma\left(W_1\left(W_0\left(F^c_{\mathrm{avg}}\right)\right) + W_1(W_0(F^c_{\mathrm{max}}))\right), \quad (6)$$

where F denotes the input feature maps and F^c_{avg} and F^c_{max} denote the feature maps that have been pooled on average and made large pooling, respectively. W_0 and W_1 denote the parameters of two of the layers in the multilayer perceptron, which use ReLU as the activation function for the neurons of the two-layer neural network. σ is the sigmoid function. At the time of calculation, F^c_{avg} and F^c_{max} share W_0 and W_1 in the multilayer perceptron.

Figure 6 shows the structure of the spatial attention module of CBAM. Unlike channel attention, spatial attention focuses on the location information of the target to be detected on the image. It first uses average pooling and maximum pooling in the channel dimension to obtain the information of the channel layer, then uses channel concatenation to concatenate the two-channel feature maps, and then uses a hidden layer containing a single convolution

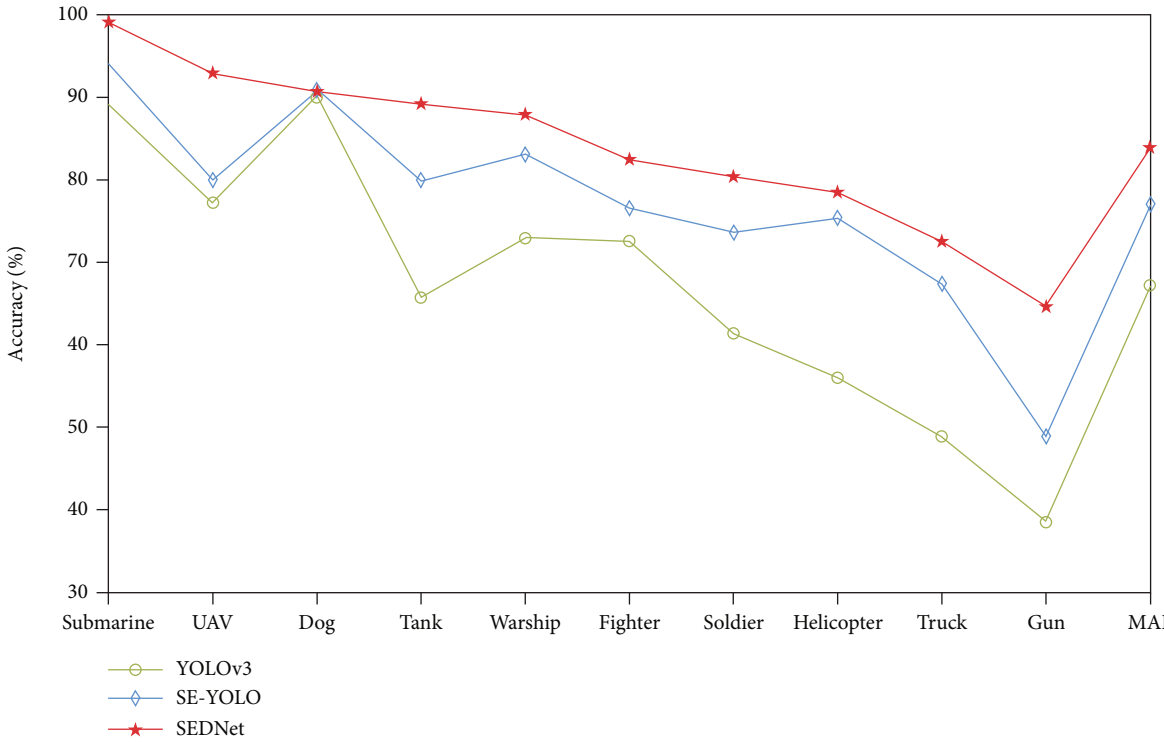

Figure 9: Comparison of the average accuracy rates of YOLOv3, SE-YOLO, and SEDNet.

kernel to convolve the feature maps after the above processing.

The spatial attention module of CBAM computes the feature mapping as shown in

$$M_s\left(F'\right) = \sigma\left(f^{7\times7}\left(\left[\text{AvgPool}\left(F'\right)\right]\right)\right), \tag{7}$$

$$M_s\left(F'\right) = \sigma\left(f^{7\times7}\left(\left[F'_{\text{avg}}; {}'F_{\max}\right]\right)\right), \tag{8}$$

where F' denotes the feature map processed by the channel attention module, F'_{avg} and ${}'F_{\max}$ denote the feature map after average pooling and maximum pooling, respectively, $f^{7\times7}$ denotes the convolution operation with a convolution kernel size of 7×7, and σ is the sigmoid function.

CBAM is stable, effective, and versatile. Introducing it into the YOLOv3 backbone network can effectively improve the detection accuracy of the model at the cost of a small amount of calculation. Therefore, this paper chooses to embed CBAM in YOLOv3 to improve the target detection ability of the model in a complex environment with high target overlap and strong target camouflage ability.

3. Situational Element Awareness Model

3.1. Network Structure of Situational Element Awareness Model. In this paper, while keeping the structure of the YOLOv3 network basically unchanged, CBAM is embedded into the five downsampling processes of YOLOv3, and a CBAM is added to the convolutional layers of the Darknet53 backbone network and after the two convolutional layers of each residual module, to construct the situation elements detection network.

The introduction of CBAM can help the model extract key local information from global information and invest more computational resources in key areas to obtain more key detailed information and suppress useless information, to overcome the problems of difficult target localization, difficult classification, and background interference caused by the strong camouflage ability of detected targets in the process of situational element detection and improve the accuracy rate of the model in detecting situational elements. The network structure of the SEDNet model is shown in Figure 7.

3.2. The Loss Function of Situational Element Awareness Model. YOLOv3 uses the cross-entropy loss function to calculate the bounding box confidence error and introduces IoU indirectly as the optimized loss term in the calculation of the error term of the bounding box confidence, but directly using IoU as the loss function may have the following problems: if there is no overlap between the prediction box and the real box, IoU is 0 and the backpropagation gradient is 0, and further optimization of the loss function is not possible. And when the IoU is the same, there is also a difference in the way the prediction frame overlaps with the real frame. To solve the above problems, this paper uses the GIoU loss function. The GIoU loss function can be calculated by

$$\text{IoU} = \frac{A \cap B}{A \cup B}, \tag{9}$$

$$\mathrm{GIoU} = \mathrm{IoU} - \frac{C/(A \cup B)}{C}, \tag{10}$$

$$L_{\mathrm{GIoU}} = 1 - \mathrm{GIoU}, \tag{11}$$

where A denotes the prediction box and B denotes the real box. The GIoU has the same scale invariance as the IoU, and it solves the problem that the IoU cannot reflect the overlap mode and the loss function cannot be optimized when the IoU is zero. For obscured overlapping scenarios with strong target camouflage and high target overlap, using $(1 - \mathrm{GIoU})$ as the bounding box regression loss function can achieve better results.

To address the problem that difficult samples cannot be effectively learned due to the imbalance of positive and negative samples in single-stage target detection algorithms such as YOLOv3, the focal loss is used in this paper. Focal loss is calculated as shown in

$$\mathrm{FL} = -\alpha(1 - P_t)^{\gamma} 1g(P_t), \tag{12}$$

where FL denotes the focal loss function and P_t denotes the probability that the target is correctly classified. The focal loss adds a weight decay term to the traditional cross-entropy loss function, which makes the size of the weights decreases with the growth of P_t and increases with the decrease of P_t. The traditional cross-entropy loss function does not distinguish between easy and hard samples but adds up the same weights to obtain the total loss, which leads to the positive samples cannot being trained effectively. The focal loss effectively solves this problem.

In this paper, the loss function L' of the SEDNet model is constructed using the traditional cross-entropy loss function as the classification error loss function, the focal loss function as the bounding box object confidence loss function, and $(1 - \mathrm{GIoU})$ as the bounding box regression loss function [11], and L' can be calculated using

$$L' = L'_1 + L'_2 + L'_3, \tag{13}$$

$$L'_1 == \sum_{0}^{\mathrm{cellnumber} * B} I^{\mathrm{object}} \times \left(1 - \mathrm{GIoU}^{\mathrm{groundtruthpredict}}\right), \tag{14}$$

$$L'_2 = \sum_{0}^{\mathrm{cellnumber} * B} m \times \mathrm{focal\_loss}\left(\mathrm{CE}\left(p_{0,} q_0\right)\right), \tag{15}$$

$$L'_3 = \sum_{0}^{\mathrm{cellnumber} * B} I^{\mathrm{object}} \times \sum_{c=0}^{C} \mathrm{CE}(p(c), q(c)), \tag{16}$$

where L' is the loss function of SEDNet, $L_1{}'$ represents the bounding box loss function, $L_2{}'$ represents the bounding box object confidence loss function, and $L_3{}'$ is the bounding box regression loss function.

4. Experimental and Analysis

4.1. The Experimental Configuration. All experiments are based on a PC with an Intel(R) Core (TM) i7-8750H CPU @ 2.20 GHz, NVIDIA RTX 2070 GPU, the PyTorch framework.

We utilize Adam [12] to update the network weight. The initial learning rate of the network was set to 0.001, and the learning rate was reduced by 5% for each epoch experienced. Meanwhile, to further reduce the training time, the SEDNet network is trained to utilize migration learning in this paper. At the beginning of training, the parameters of some convolutional layers in the backbone network are initialized with the parameters in the pretraining weight file, and the parameters of some convolutional layers at the end are fine-tuned for training.

4.2. SA Dataset. In this paper, we used the labeling tool to manually label the collected target images to create a homemade SA dataset in VOC format for training. The situational element dataset, which is specially collected for this piece, contains ten types of targets, soldier, dog, gun, fighter, helicopter, UAV, tank, truck, warship, and submarine. In addition, to solve the overfitting problem, this paper performs data enhancement by scaling the collected images, mirror flipping, changing the brightness and contrast, and adding Gaussian noise [13], which enlarges the data volume by ten times and greatly improves the robustness and generalization ability of the network, allowing the limited data to produce more data value [14].

Figure 8 shows the schematic diagram of using SEDNet for the detection of situational elements in images with strong background interference and difficult situational target localization. From Figure 8, it can be seen that SEDNet can well overcome the problems of difficult target localization, difficult classification, and strong background interference in the process of posture element detection.

4.3. Comparison with Other Methods. To verify that the method proposed in this paper can accomplish situational element detection more effectively, experiments were conducted on the SA dataset with YOLOv3, YOLOv3 with the introduction of SENet (hereafter referred to as SE-YOLO), and the SEDNet model proposed in this paper, respectively. Randomly, 4680 images from 5850 images of the SA dataset were selected as the training set, and the remaining 1170 images were used as the test set for the experiments. Table 1 and Figure 9 show the comparison of the mean accuracy AP and the mean accuracy mAP for all classes of target detection by the conventional YOLOv3, SE-YOLOv3, and the method in this paper.

The experimental results on the SA dataset show that the mAP of SEDNet reaches a high level, improving 16.61% compared to the traditional YOLOv3 network and 6.85% compared to SE-YOLO, which proves that the SEDNet model proposed in this paper can significantly improve the performance of the deep learning network for situational element perception. In addition, both SEDNet and SE-YOLO have more significant improvements in the average accuracy rate compared with the traditional YOLOv3 network, which indicates that adding the attention module can effectively improve the model's ability to detect the situational elements.

The above results confirm the effectiveness and superiority of the work in this paper. Compared with the traditional YOLOv3 target detection algorithm, SEDNet achieves a large average accuracy improvement with only a small number of additional parameters and almost negligible computational effort and can detect targets more accurately in complex obscured overlapping scenarios with high target overlap and strong target camouflage capability and accurately obtain the location and class of most of the classes to be detected.

5. Conclusions

In order to explore a target detection algorithm suitable for situational element detection for the technical requirements of situational element detection, this paper proposes an efficient and accurate target detection network SEDNet based on the combination of YOLOv3 and CBAM. The main work of this paper has the following points.

(1) Introduces the research background and significance of this paper, analyzes the shortcomings of the current research work on situational element detection technology, and proposes solutions

(2) We propose a Situation Element Detection Network based on the combination of YOLOv3 and CBAM, which embeds CBAM into the downsampling process of YOLOv3, builds the network structure of SEDNet, and makes full use of the ability of CBAM to extract key local information from the global information. The network structure of SEDNet is built to take full advantage of CBAM's ability to extract key local information from global information, which overcomes the problems of difficult target localization, difficult classification, and strong background interference in the process of situational elements detection and improves the accuracy of the model in detecting situational elements

(3) We reconstructed the loss function based on traditional cross-entropy and focal loss and optimized the training strategy based on the K-means method, Adam algorithm, and migration learning method to improve the detection accuracy and training speed of the model

(4) We constructed the SA dataset and used data augmentation techniques to expand the training set data by a factor of 10 to avoid generating overfitting of the training results

(5) We performed the localization and classification of 10 common categories of situational elements, namely, soldiers, military dogs, guns, fighter jets, helicopter gunships, unmanned reconnaissance aircraft, tanks, military trucks, warships, and submarines, on the SA dataset. The experimental results show that SEDNet achieves an average accuracy of 83.83% on the SA dataset, and the mAP is improved by 16.61% compared with the traditional YOLOv3 target detection model, which verifies the effectiveness and superiority of the method in this paper

The SEDNet model can effectively solve the problems of multiple overlapping targets and strong target camouflage ability in complex environments that traditional target detection networks can hardly cope with and is applicable to the field of situational element perception. The research work in this paper lays the foundation for further research work on situational understanding, assisted decision-making, and assessment and provides important technical support.

Data Availability

The data used to support the findings of this study have been deposited in the GITHUB repository (https://github.com/chengkangda/1).

Conflicts of Interest

The authors declare that they have no competing interests.

References

[1] M. R. Endsley, "Situation awareness in aviation systems," *Handbook of Aviation Human Factors: Second Edition*, vol. 12, no. 1-12, p. 22, 1999.

[2] M. R. Endsley, "Toward a theory of situation awareness in dynamic systems," *Human Factors: The Journal of the Human Factors and Ergonomics Society*, vol. 37, no. 1, pp. 32–64, 1995.

[3] H. Peng, Y. Zhang, S. Yang, and B. Song, "Battlefield image situational awareness application based on deep learning," *IEEE Intelligent Systems*, vol. 35, no. 1, pp. 36–43, 2020.

[4] M. Liu, B. Li, Y. Chen et al., "Location parameter estimation of moving aerial target in space-air-ground integrated networks-based IoV," *IEEE Internet of Things Journal*, vol. 99, 2021.

[5] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: convolutional block attention module," in *Computer Vision – ECCV 2018*, pp. 3–19, Springer, Cham, 2018.

[6] X. Chengji, X. Wang, and Y. Yang, "Attention-YOLO: YOLO detection algorithm with the introduction of attention mechanism," *Computer Engineering and Applications*, vol. 55, no. 6, pp. 13–125, 2019.

[7] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, Las Vegas, NV, USA, 2015.

[8] J. Redmon and A. Farhadi, "YOLOv3: An Incremental Improvement," 2018, https://arxiv.org/1804.02767.

[9] B. Ali, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[10] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Salt Lake City, UT, USA, 2018.

[11] M. Liu, Z. Liu, W. Lu, Y. Chen, X. Gao, and N. Zhao, "Distributed few-shot learning for intelligent recognition of communication jamming," in *IEEE Journal of Selected Topics in Signal Processing*, 2021.

[12] D. Kingma and J. Ba, "A method for stochastic optimization," *Computer Science*, vol. 1412.6980, 2014.

[13] M. Liu, J. Wang, N. Zhao, Y. Chen, H. Song, and R. Yu, "Radio frequency fingerprint collaborative intelligent identification using incremental learning," *IEEE Transactions on Network Science and Engineering*, vol. 99, 2021.

[14] M. Liu, C. Liu, M. Li, Y. Chen, S. Zheng, and N. Zhao, "Intelligent passive detection of aerial target in space-air-ground integrated networks," *China Communications*, vol. 19, no. 1, pp. 52–63, 2022.