Hindawi
Advances in Multimedia
Volume 2022, Article ID 8703380, 10 pages
https://doi.org/10.1155/2022/8703380

Research Article

Gated Channel Attention Mechanism YOLOv3 Network for Small Target Detection

Xi Yang, Jin Shi, and Juan Zhang

Physical Education College of Zhengzhou University, Zhengzhou 450044, Henna, China

Correspondence should be addressed to Xi Yang; yangxi@peczzu.edu.cn

Received 2 June 2022; Revised 12 July 2022; Accepted 16 July 2022; Published 12 August 2022

Academic Editor: Qiang Li

In order to solve the problem of low recognition rate and high missed rate in current target detection task, this paper proposes an improved YOLOv3 algorithm based on a gated channel attention mechanism (GCAM) and adaptive up-sampling module. Firstly, darknet-53 is used as the backbone network to extract image basic features. Secondly, an adaptive up-sampling module is introduced to expand the low-resolution convolutional feature images, which effectively enhances the fusion efficiency of the convolutional feature images at different scales. Finally, GCAM is added to improve the network's feature expression and detection capability for small targets before the three-scale channels output the prediction results. The results show that the improved method can adapt to multiscale target detection tasks in complex scenes and reduce the missing rate of a small target.

1. Introduction

Small target detection is an object detection technology that can find and judge the category of the object in the image with the help of computer vision [1]. At present, the technology has been widely used in national defense, military, transportation, industry, virtual reality, and other fields [2]. In complex realistic scenes, small targets are difficult to locate and identify due to different shooting angles, nontarget object occlusion, imaging weather, and lighting conditions. At the same time, the small-size target lacks the appearance information to distinguish itself from the background or similar categories, and it is easy to lose feature information in the deep convolutional network, and the detection is prone to miss and misdetection.

There are two ways to define small targets in target detection, namely, the definition of relative size and the definition of absolute size [3]. Relative size is defined by the society of Photo-optical Instrumentation Engineers (SPIE). A small target is defined as a target area less than 80 pixels in a 256×256 pixel image. That is, less than 0.12% of 256×256 pixels is a small target. The other is the definition of absolute size. In the MS COCO data set, targets with sizes less than 32×32 pixels are considered small targets. In 2016, a scholar defined small targets as targets ranging from 16×16 pixels to 42×42 pixels in an image of 640×480 pixels. According to the data of pedestrians and nonmotor vehicle drivers in traffic scenes, some scholars believe that objects with 30 to 60 pixels and less than 40% occlusion are small target objects. Objects with pixel values ranging from 10 to 50 pixels were defined as small objects in the aerial image dataset DOTA [4] and Wider Face dataset [5]. In the pedestrian recognition dataset City Persons [6], targets with a height of less than 75 pixels are defined as small targets. In general, there is no precise and unique definition of small goals, which need to be determined according to the application scenario.

In recent years, small target detection in the large image has become a research hotspot of domestic and foreign scholars, playing an important role in industrial production, satellite remote sensing, target tracking, and other fields [7]. Small target detection is generally used to accurately locate small targets (generally smaller than 32 pixel $\times$ 32 pixel) in large-size images, such as finding the designated targets (cars, planes, etc.) in remote-sensing images and the positions of defects (scratches and black spots on ceramic tiles, etc.) in industrial products and then marking the categories of targets. The difficulty of small target detection lies in the

small proportion of target in the original image, and the detector cannot extract sufficient and effective features, resulting in unsatisfactory small target detection results [8].

In recent years, the deep convolutional neural network has been widely used in many methods of target detection. According to the processing of candidate boxes, such detection methods are divided into two categories [9]. (1) Based on the one-stage target detection method, this method takes the whole image as the input. Its purpose is to increase the receptive field of the target on the image and return the position and category information of the target at different positions on the image. The most representative methods are [10] and [11]. (2) Based on the two-stage target detection method, target candidate boxes that may exist in the image are firstly extracted, and then each region candidate box is classified and location regression is performed. The representative methods mainly include [12–14]. The first method has fast detection speed and good adaptability to large targets, but small targets are easy to miss detection. The second method has relatively high accuracy in small target detection, but the speed of feature extraction, detection, and classification is relatively slow. Since each stage is separated, it can be improved and optimized separately, which has a lot of room for improvement.

Recently, literature [15–17] have proposed several methods for small target detection. In [15], deconvolution technology is applied to all feature images to obtain magnified feature images. However, the application of the deconvolution module to all feature graphs has the limitation of increasing model complexity and slowing down detection speed. Tong et al. [16] obtained high accuracy and speed by combining characteristic information and deconvolution operations at different scales. Yan et al. [17] use generative adversarial networks to generate high-resolution features by using low-resolution features as the input of GAN. However, these methods have some limitations, and there is still a lot of room for improvement. Therefore, this paper proposes a small target detection algorithm based on an improved YOLOv3 model and attention mechanism. On the basis of YOLOv3, darknet-53 was used as the main dry extraction network, and the original algorithm was improved by introducing an adaptive up-sampling module that could learn weight parameters and GCAM channel attention mechanism. Finally, the improved network was tested on the data set collected and annotated by ourselves. Experimental results show that the improved algorithm has better prediction results when the background interference greatly affects the target detection. At the same time, it is proved that the improved algorithm has good robustness and strong antienvironmental interference ability and effectively improves the ability of target detection.

The innovations and contributions of this paper are listed below:

(1) The algorithm uses darknet-53 as the backbone network for image basic feature extraction

(2) The adaptive up-sampling module is introduced to expand the low-resolution convolution feature map, which effectively enhances the fusion effect of convolution feature maps with different scales

(3) GCAM is added before the three scale channels output the prediction results to improve the feature expression and detection ability of the network to small targets

This paper consists of five main sections as follows. Section 1 is the introduction, Section 2 is state of the art, Section 3 is a methodology, Section 4 is result analysis and discussion, and Section 5 is the conclusion.

2. State of the Art

2.1. Introduce YOLOv3. Yolov3 is improved on the basis of yolov1 and yolov2. It is a single-stage target detection algorithm. Unlike the R-CNN series, which divides the target detection task into two steps of generating a candidate frame and identifying objects in the frame, it merges the whole process to directly generate prediction results. Compared with the two-stage target detection algorithm, the single-stage target detection algorithm has the characteristics of fast detection speed but low accuracy. YOLOv3 adopts darknet-53 with residual connection as the backbone feature extraction network. In addition, referring to the Feature Pyramid Network (FPN) structure in [18], it uses feature maps of three different scales to perform multiscale feature fusion and output prediction results, so as to achieve a balance between speed and accuracy in the target detection tasks. The overall architecture of yolov3 is shown below (see Figure 1). For 416×416 input images, basic feature extraction is firstly carried out through darknet-53 (full connection layer removed) backbone feature extraction network, which contains 1 DBL module and 5 residual modules. Then, the outputs of the last three residual modules in the backbone network are input into the feature pyramid structure as features of three different scales for feature fusion. Finally, convolution operations were performed on the fused feature layers in the three channels to output the prediction results of 13×13, 26×26, and 52×52 scales.

2.2. Prediction Target Box. YOLOv3 algorithm segmented the image to be detected into $S \times S$ grid cells of different scales U ($U = 3$) (e.g., 13×13, 26×26, and 52×52). They correspond to the outputs of the three parallel network branches on the right side of Figure 1. If the target object center falls into a grid cell, the grid cell needs to predict the target object. Figure 2 shows the relationship between the input image and the $S \times S$ feature layer. For any of the above grid cells, three prior boxes with different aspect ratios should be predicted, and each prior box contains the confidence Conf, category H, and location information cls of the current grid.

The confidence represents the likelihood that the current grid cell contains objects as

$$\mathrm{Conf}_w^t = U_{w,t}\,(\mathrm{Object}) \times \mathrm{IOU}_{\mathrm{pred}}^{\mathrm{truth}}, \tag{1}$$

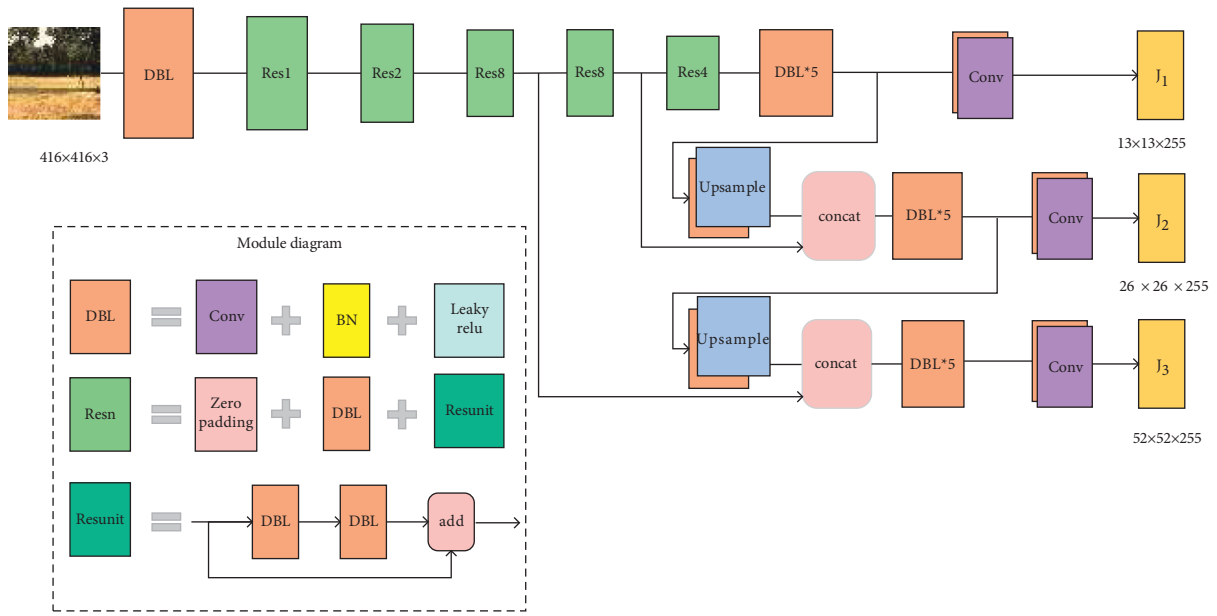

Figure 1: YOLOv3 structure diagram.

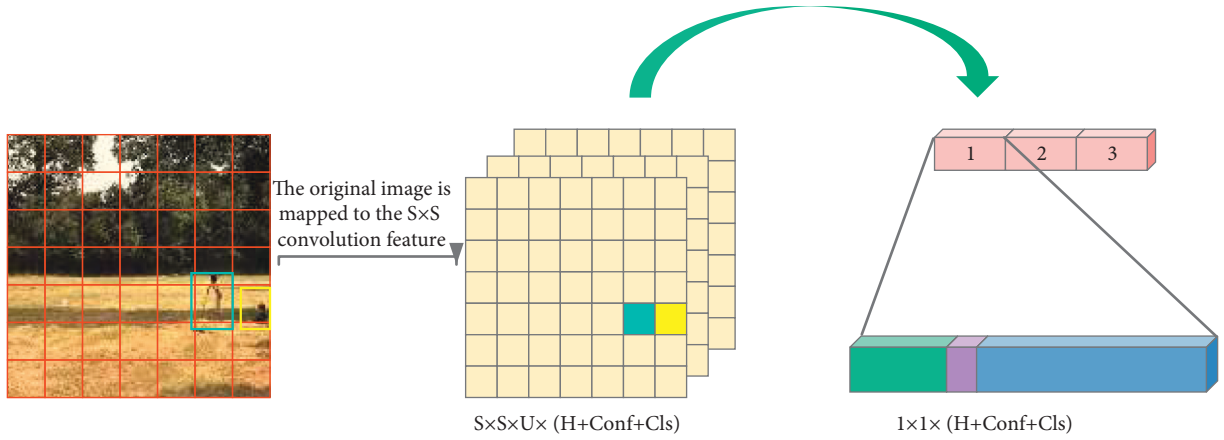

Figure 2: Mapping relationship between the input image and $S \times S$ feature layer.

where Conf_w^t represents the confidence of the tth prior box in the wth grid cell. $\text{IOU}_{\text{pred}}^{\text{truth}}$ represents the prediction box of the object that the current grid is responsible for. $U_{w,t}$ (Object) indicates that if an Object's center falls into the current grid, its value is 1. The value is 0 if no object center falls into the current grid, i.e., the current grid is in the background:

$$U_{w,t}\text{Object} = \begin{cases} 1, & \text{Objects exist in the grid,} \\ 0, & \text{No object exists in grid.} \end{cases} \tag{2}$$

The result of cls prediction of location information contains four values, n_i, n_j, n_m, and n_b. The coordinates, width, and height of the center point of the prediction box are obtained by transforming the four predicted values through the following formula:

$$\begin{cases} h_i = \sigma(n_i) + c_i, \\ h_j = \sigma(n_j) \\ h_m = U_m * e^{n_m}, \\ h_b = U_b * e^{n_b}. \end{cases} \tag{3}$$

The position relationship of predicted values in the cell grid is shown in Figure 3.

U_m and U_b are the width and height of the prior box. h_m and h_b are the actual width and height predicted after the conversion. h_i and h_j are the actual center coordinates predicted after the conversion. C_i and C_j are the coordinates of the upper left corner of the cell relative to the whole picture.

YOLOv3 uses a nonmaximum suppression algorithm to filter the prediction box. For a certain target on the image to

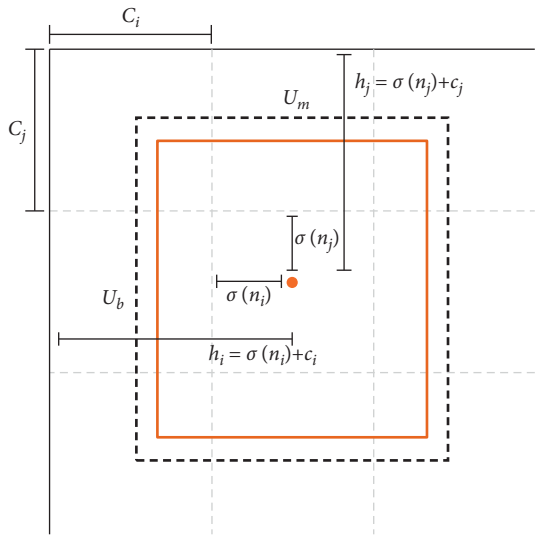

Figure 3: Position relationship diagram of the predicted value in cell grid.

be detected, the detection frame C with the highest score is selected first, and then the IOU value of the remaining frame and C is calculated, respectively. When the IOU value exceeds the threshold, the enclosure that exceeds the threshold is suppressed. Then, select the detection frame with the highest score from the remaining detection frames and repeat the above process until finally ensuring that only one detection frame exists in each target.

3. Methodology

In the proposed algorithm, the up-sampling module in the original network structure (Figure 1) is replaced by the adaptive up-sampling module, and GCAM attentional mechanism is added before the output of three-scale prediction results, j_1, j_2, and j_3. Compared with the original network, it has the following two improvements.

The GCAM (gated channel attention mechanism) is introduced to realize the interaction between the feature layer channels. The correlation and importance of the information of the feature layer of different channels are learned by assigning the weight of the features of each channel. In addition, the attention mechanism also learns the important relationship between the two feature layers before and after filtering channel information. It effectively improves the feature extraction ability of the network for small targets and reduces the false detection and missed detection caused by the complex background of remote-sensing images.

Meanwhile, an adaptive up-sampling module is introduced to replace the original up-sampling operation. This method can find the most suitable up sampling method for training tasks by learning weight parameters autonomously, effectively reducing the semantic loss of up sampling in a low-resolution feature layer, and enhancing the fusion effect of convolution at different scales.

3.1. Gated Channel Attention Mechanism. The mechanism of attention originated from the study of the human thinking mode. When humans deal with a large amount of information with varying degrees of importance, they always pay selective attention to a part of all information, namely, the important information, while ignoring the rest. Since human beings have a limited capacity to process information resources, to allocate these resources properly, we need to select the most important part of the information and focus on it. Similarly, the attentional mechanism in deep learning is to select the most important part of the input information and give it a higher weight so that the network can pay attention to this information.

Some scholars began to explore ways to improve the performance of the convolutional neural networks in computer vision by using an attention mechanism. Currently, there are mainly two attention mechanisms commonly used in computer vision, channel attention mechanism and spatial attention mechanism. The channel attention mechanism considers that the importance of each channel in the convolutional layer is different, and the weight of each channel is adjusted to enhance network feature extraction ability. The spatial attention mechanism uses the idea of the channel attention mechanism to think that the importance of each pixel in different channels is different, and the weight of all pixels in different channels can be adjusted to enhance the ability of network feature extraction.

ECA (Efficient Channel Attention) is a classical structure of the channel attention mechanism. As shown in Figure 4, for a $B \times M \times C$ input convolution layer, features are compressed from the spatial dimension through global average pooling operation. Thus, a $1 \times 1 \times C$ convolution layer with a global receptive field matching the number of input channels is obtained. Then, a 1×1 convolution ensures cross-channel information interaction without dimensionality reduction. Finally, the weight value is compressed to 0-1 by the sigmoid function and then multiplied with the input convolution layer channel by channel to complete channel importance weight allocation.

GCAM is an improved gated channel attention mechanism based on ECA. The standard ECA structure directly uses channel importance weights to perform subsequent operations on the feature layer after input convolution filtering. However, some important information may be filtered in the original input feature layer. Therefore, GCAM learns another set of weights to determine whether certain channels of the original input feature layer should be retained.

The structure of GCAM is shown in Figure 5. The upper channel learns the importance weight j_1 of the input features, and the lower channel learns the importance weight j_2 of the original feature input layer and the filtered channel layer. The calculation process is as follows:

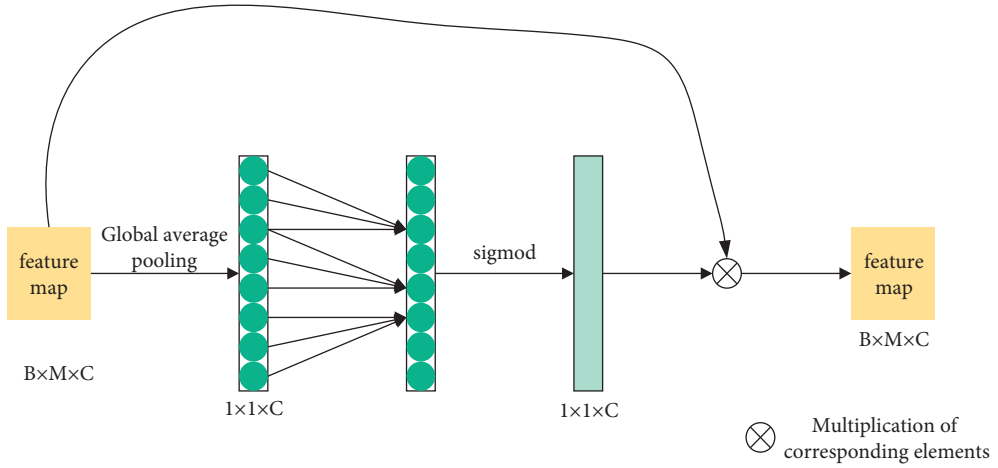

Figure 4: ECA structure.

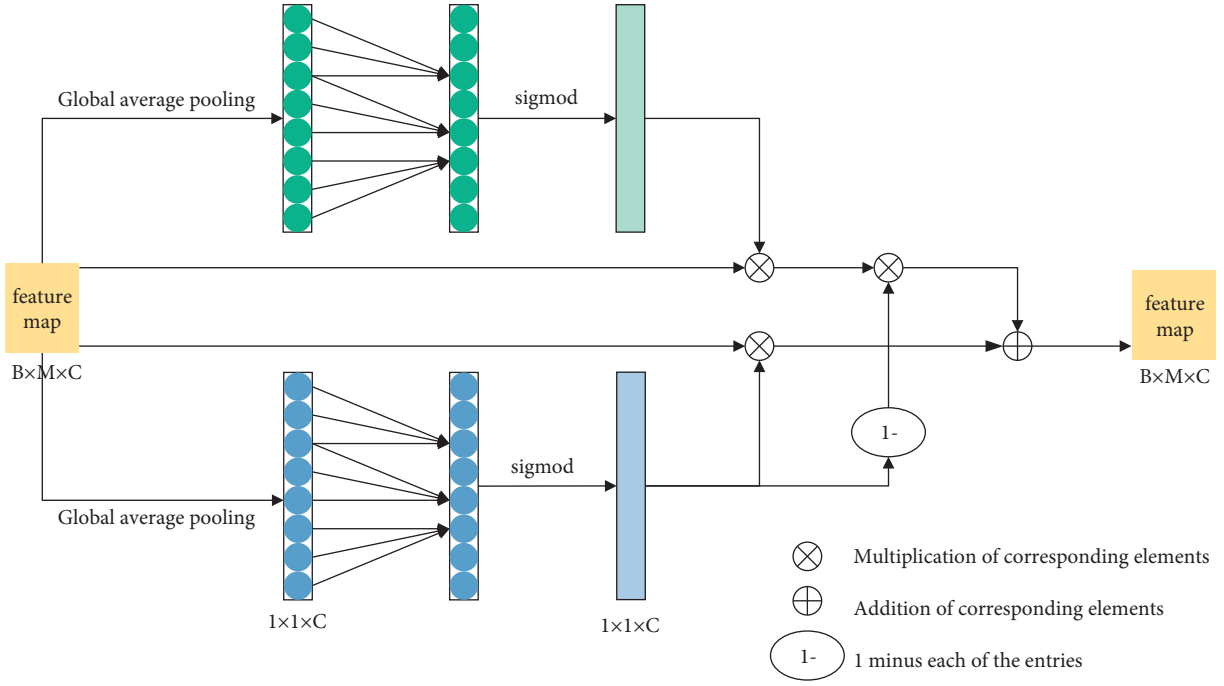

Figure 5: GCAM structure diagram.

$$j_x = \sigma(\mathrm{Conv}(a(i))),$$

$$a(i) = \frac{1}{MB}\sum_{x=1}^{M}\sum_{y=1}^{B} i_{x,y}, \tag{4}$$

$$\sigma(i) = \frac{1}{1+e^{-i}},$$

where i represents an original feature input whose length, width, and number of channels are m, b, and c, respectively, and Conv(i) represents the convolution of the feature layer with 1×1.

Finally, the two weight parameters learned and feature input are integrated, as shown in formula (5), to obtain the final output feature layer j:

$$j = i * j_2 + (i * j_1) * (1 - j_2). \tag{5}$$

3.2. Adaptive Up-Sampling Module. Due to the small number of pixels in the low-resolution image, many details will be lost when it is up sampled. Therefore, it is one of the core problems of up sampling to reduce the loss of detail features as much as possible and improve the feature recovery ability of low-resolution images.

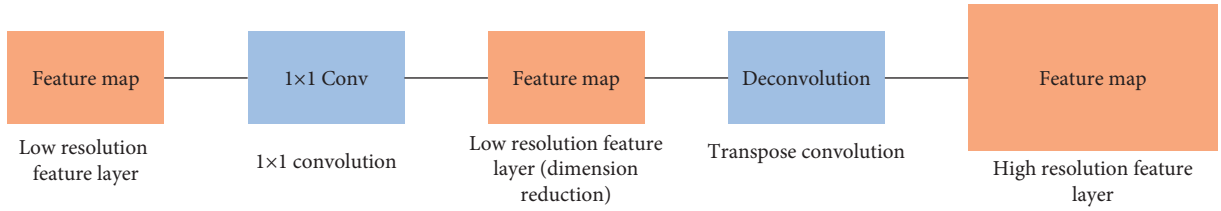

Figure 6: Adaptive up-sampling module.

Traditional up-sampling methods include linear interpolation and inverse pooling. Linear interpolation uses geometric relations to estimate the newly added pixels through the known pixels. Taking nearest neighbor interpolation as an example, when the picture is enlarged, the new pixel is directly generated using the color of the nearest original pixel. The antipooling is to do some simple zero-filling and expansion operations on the image. Firstly, the location information of the maximum value in the process of pooling is recorded. Then, when the image size is enlarged by antipooling, only the location of the maximum value is restored, and other values are directly set to 0. Although the traditional sampling method is simple and fast, it will produce obvious serration, which leads to the loss of the original image details.

New up-sampling methods based on deep learning include transpose convolution. Transpose convolution is a special convolution operation. This is an adaptive up-sampling method, which uses the fitting of weight parameters to keep the details of the up-sampled image consistent with the original image as much as possible.

The traditional sampling method is low in computation and relatively simple in implementation, but it will inevitably lead to the loss of original detail features due to the limited number of pixels in low-resolution images. Transpose convolution can learn weight parameters to better fit the original image, so it can ensure that the details of low-resolution images can be restored as much as possible. However, its implementation is complicated and requires a lot of calculation. Therefore, this paper proposes an adaptive up-sampling module (see Figure 6). Using 1×1 convolution, we can keep the width and height of the input feature layer unchanged while reducing the channel dimension to reduce network parameters and reduce the amount of computation. In the design of transpose convolution, due to the uneven overlap in the process of transpose convolution operation, some parts of the image will be darker than other colour checkerboard effect. When the convolution kernel size of transpose convolution can be divisible by the step size, this effect will be relieved.

4. Result Analysis and Discussion

Multiscale target detection experiments were carried out on the dataset collected and annotated by ourselves. The dataset contains 420 optical images with an average scale of about 1300×900. There are 3,324 labeled targets with a resolution of 0.8~2.0 m, and each image contains at least one target.

Table 1: Definition of boundary area and number of targets based on instance size distribution.

Target	Scale	Quantity
Small target	$(0, 60^2)$	1347
Medium target	$(60^2, 120^2)$	1975
Large target	$(120^2, +\infty)$	332

In order to avoid overfitting, data expansion is carried out by rotation and inversion. During training, the experiment randomly assigned 70% of the dataset to train and the remaining 30% to testing. According to the scale of the dataset (1300×900) and the target distribution information, the corresponding boundary box scale (small target: $S \le 60^2$, medium target: $60^2 < S \le 120^2$, and large target $S > 120^2$) was defined. Statistics are made on target scales in the collected data set, and the results are shown in Table 1. The number of small targets in the collected data set accounts for 36.8% of the total number of targets.

The experimental hardware environment was Inter E5-2680 CPU, 256G memory, NVIDIA TITAN RTX GPU, and Ubuntu 16.04 operating system. PyTorch was used as a deep learning framework for training and testing. In this paper, the end-to-end training method is adopted.

Average Precision (AP) is used as the evaluation index of target detection results. AP calculates the average accuracy of recall rate between 0 and 1, that is, the area enveloped by the accuracy-recall curve. Therefore, the higher AP value represents better detection performance. Precision P and recall rate R can be expressed as follows:

$$u = \frac{N_u}{N_u + F_u}, \qquad r = \frac{N_u}{N_u + F_t}, \tag{6}$$

where N_u is a true example, F_u is a false positive example, and F_t is a false negative example.

Compared with other algorithms, the detection results are shown in Figure 7.

The detection accuracy of [19] is 86.2%. Compared with [19], the accuracy of [20] as a single-stage target is insufficient. In [21], 84.6% detection accuracy was achieved while considering detection speed. Compared with [21], deconvolution feature fusion is adopted in [22] to improve the detection ability of multiscale targets. On the basis of [21], the pyramid feature fusion method was introduced in [23]. The improved YOLOv3 algorithm achieves the optimal

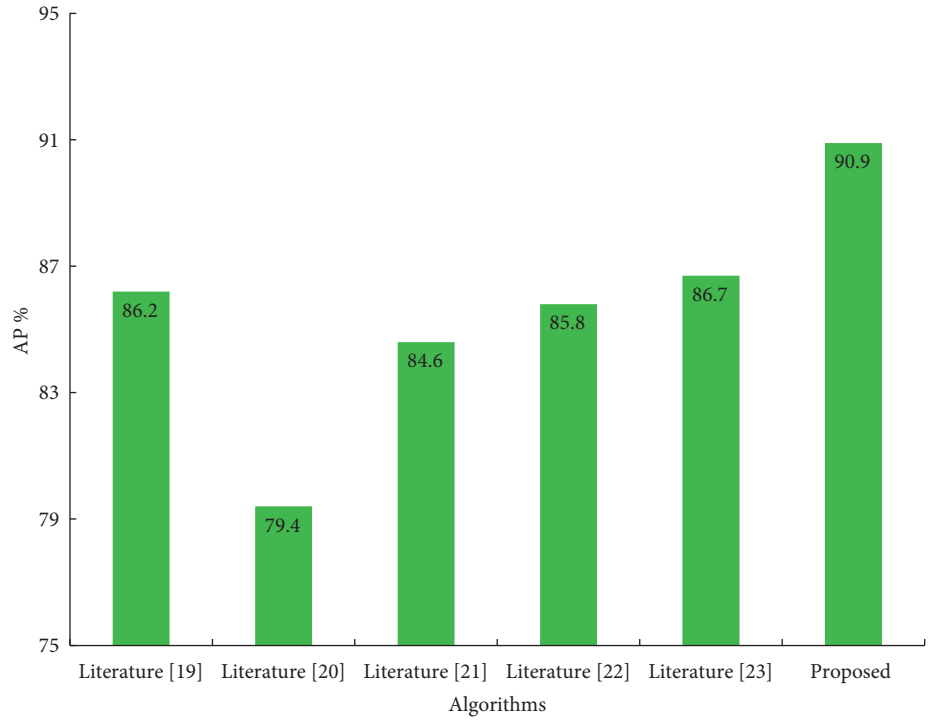

Figure 7: Comparison of average precision of different algorithms.

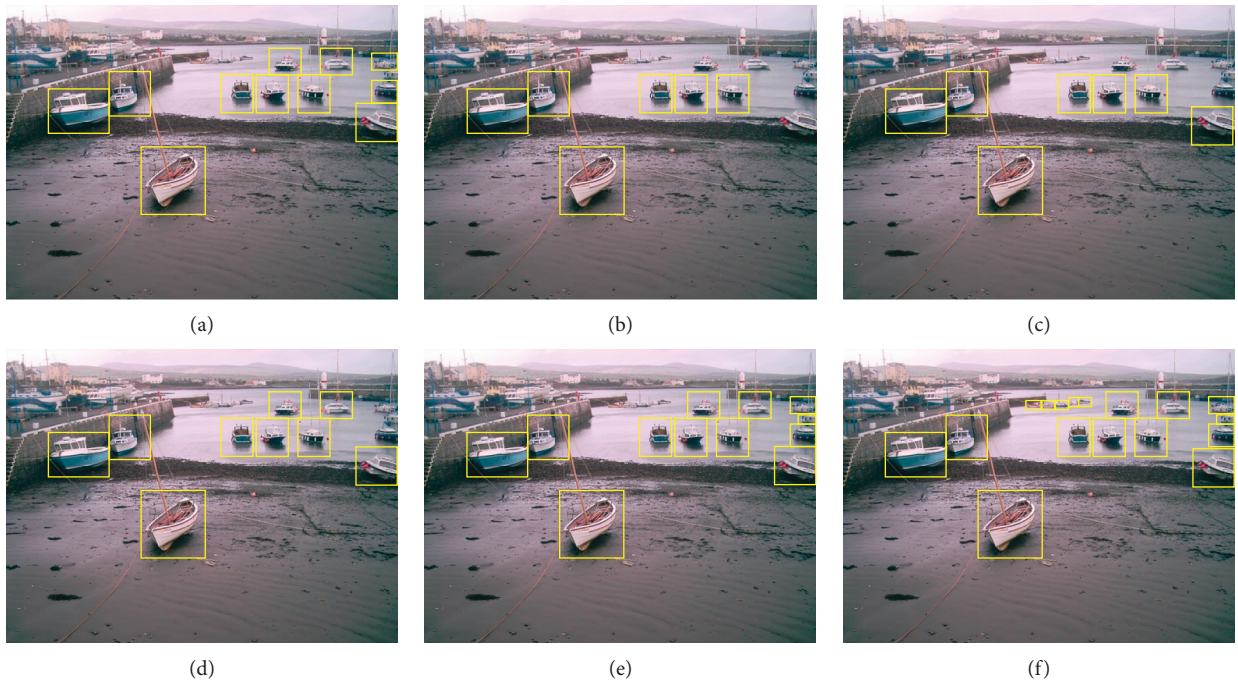

Figure 8: Detection results of different algorithms. (a) Literature [19]. (b) Literature [20]. (c) Literature [21]. (d) Literature [22]. (e) Literature [23]. (f) Proposed.

average detection accuracy, which is 4.7% higher than the algorithm in [21]. It shows that the improved YOLOv3 algorithm can efficiently detect small targets in the complex backgrounds on the basis of satisfying real-time detection. As can be seen from Figure 8, compared with [19–23], the method in this paper has a stronger detection ability for small targets and can effectively reduce the missed detection rate of small targets.

4.1. Ablation Experiment. In order to quantitatively analyze the influence of the adaptive up-sampling module, ECA

Table 2: Results of ablation experiments for each module.

Modules	AP/% (IOU = 0.5)	Frame frequency/s
YOLOv3	86.50	31.3
Add adaptive up sampling	89.20	26.4
Add ECA	88.80	29.7
Add GCAM	89.60	31.2
Proposed	91.60	25.4

(a)

(b)

Figure 9: Detection results of different types of targets. (a) YOLOv3 algorithm. (b) Proposed algorithm.

channel attention, and GCAM channel attention on the target detection accuracy, an ablation experiment was designed [24]. And the average detection accuracy and frame rate at IOU = 0.5 were used as evaluation indexes to characterize the algorithm performance [25]. The ablation results are shown in Table 2.

As can be seen from Table 2, the adaptive up-sampling module is added on the basis of the YOLOv3 algorithm, and the average accuracy is improved from 86.5% to 89.2%, with an overall improvement of 2.7%. Due to the increase of the adaptive up-sampling module, the detection speed slows down, and the frame rate decreases from 31.3 frame/s to 26.4 frame/s. ECA channel attention mechanism was added on the basis of the YOLOv3 algorithm, the accuracy was improved by 2.3%, and the frame rate decreased by 1.6 frame/s. GCAM channel attention mechanism was added on the basis of the YOLOv3 algorithm, and the accuracy was improved by 3.1% from 86.5% to 89.6%, with little change in frame frequency. The GCAM channel attention module (the proposed algorithm) was added on the basis of the addition of the adaptive up-sampling module. The accuracy was improved from 86.5% to 91.6%, and the detection speed was 25.4 frame/s.

4.2. Migration Experiment. In addition to the performance comparison of the algorithm in the data set, the detection model trained on the above data set is also migrated to different types of target detection. Thirty five different types of pictures were collected. The average detection accuracy of the improved YOLOv3 algorithm and YOLOv3 algorithm is 0.561 and 0.428, respectively, and the migration accuracy of the improved YOLOv3 algorithm is 0.133 higher than that of the YOLOv3 algorithm. The detection results are shown in Figure 9.

Figure 9 shows the detection effect of the detection model of the YOLOv3 algorithm and the improved method on the target. According to the detection results, compared with the YOLOv3 algorithm, the detection model trained by the improved method can detect different types of targets more effectively. This is because the feature fusion process not only improves the feature extraction ability but also enhances the feature generalization ability to a certain extent, making the model have better migration ability. The migration experiment results prove that the detection model learned by the improved algorithm has certain portability and versatility.

5. Conclusion

Aiming at the problem that the feature recovery ability of sampling on low-resolution convolution feature map is weak for small targets in the process of multiscale feature fusion, an adaptive up-sampling module is designed to replace the traditional interpolation operation so that the network can independently select the interpolation method suitable for the task of the target training set, so as to enhance the effect of feature fusion at different scales. To solve the problem that small targets contain little feature information and high positioning accuracy, this paper proposes a gated channel attention mechanism, which realizes the interaction between feature layer channels, and learns the correlation and importance of feature layer information of different channels by assigning weights to the features of each channel. Ablation experiments and migration experiments demonstrate the completeness and universality of the method. Results show that the proposed method can effectively enhance the ability of small target representation and reduce the rate of small target detection. On the basis of satisfying the real-time detection, it can realize the efficient detection of small and medium targets in complex background images. It is proved that the improved algorithm has good robustness and strong antienvironmental interference ability. However, the algorithm improved by the YOLOv3 model and attention mechanism still has error detection and missing detection phenomenon in some extreme cases, and the follow-up work will continue to optimize the detection effect of the network in extreme cases.

Data Availability

The labeled dataset used to support the findings of this study is available from the corresponding author upon request.

Conflicts of Interest

The authors declare no conflicts of interest.

Acknowledgments

This work was supported by the Physical Education College of Zhengzhou University.

References

[1] L. I. N. Liangkui, W. Shaoyou, and T. Zhongxing, "Using deep learning to detect small targets in infrared oversampling images," *Journal of Systems Engineering and Electronics*, vol. 29, no. 5, pp. 947–952, 2018.

[2] K. Zhao and X. Guo, "Analysis of the application of virtual reality technology in football training," *Journal of Sensors*, vol. 2022, Article ID 1339434, 8 pages, 2022.

[3] W. Zhang and W. Sun, "Research on small moving target detection algorithm based on complex scene," *Journal of Physics: Conference Series*, vol. 1738, no. 1, Article ID 012093, 2021.

[4] P. Zhao, Z. Qu, Y. Bu, W. Tan, and Q. Guan, "PolarDet: a fast, more precise detector for rotated target in aerial images," *International Journal of Remote Sensing*, vol. 42, no. 15, pp. 5831–5861, 2021.

[5] W. Chen, H. Huang, S. Peng, C. Zhou, and C. Zhang, "YOLO-face: a real-time face detector," *The Visual Computer*, vol. 37, no. 4, pp. 805–813, 2021.

[6] M. Braun, S. Krebs, F. Flohr, and D. M. Gavrila, "EuroCity Persons: a novel benchmark for person detection in traffic scenes," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 8, pp. 1844–1861, 2019.

[7] J. Guo, Y. Wu, and Y. Dai, "Small target detection based on reweighted infrared patch-image model," *IET Image Processing*, vol. 12, no. 1, pp. 70–79, 2018.

[8] F. S. Marvasti, M. R. Mosavi, and M. Nasiri, "Flying small target detection in IR images based on adaptive toggle operator," *IET Computer Vision*, vol. 12, no. 4, pp. 527–534, 2018.

[9] Y. Li, K. Fu, H. Sun, and X. Sun, "An aircraft detection framework based on reinforcement learning and convolutional neural networks in remote sensing images," *Remote Sensing*, vol. 10, no. 2, p. 243, 2018.

[10] M. C. Shiu, L. Y. Wei, and J. W. Liu, "A hybrid one-step-Ahead time series model based on GA-SVR and EMD for forecasting electricity loads," *Journal of Applied Science and Engineering*, vol. 20, no. 4, pp. 467–476, 2017.

[11] J. Gao, Y. Chen, Y. Wei, and J. Li, "Detection of specific building in remote sensing images using a novel YOLO-S-ciou model. Case: gas station identification," *Sensors*, vol. 21, no. 4, p. 1375, 2021.

[12] Y. Li, J. Xu, R. Xia, X. Wang, and W. Xie, "A two-stage framework of target detection in high-resolution hyperspectral images," *Signal, Image and Video Processing*, vol. 13, no. 7, pp. 1339–1346, 2019.

[13] J. Deng, X. Xuan, W. Wang, Z. Li, H. Yao, and Z. Wang, "A review of research on object detection based on deep learning," *Journal of Physics: Conference Series*, vol. 1684, no. 1, Article ID 012028, 2020.

[14] P. Du and A. Hamdulla, "Infrared small target detection using homogeneity-weighted local contrast measure," *IEEE Geoscience and Remote Sensing Letters*, vol. 17, no. 3, pp. 514–518, 2020.

[15] Y. Chen, J. Wang, X. Chen, A. K. Sangaiah, K. Yang, and Z. Cao, "Image super-resolution algorithm based on dual-channel convolutional neural networks," *Applied Sciences*, vol. 9, no. 11, p. 2316, 2019.

[16] X. Tong, B. Sun, J. Wei, Z. Zuo, and S. Su, "EAAU-net: enhanced asymmetric attention U-net for infrared small target detection," *Remote Sensing*, vol. 13, no. 16, p. 3200, 2021.

[17] X. Yan, B. Cui, Y. Xu, P. Shi, and Z. Wang, "A method of information protection for collaborative deep learning under GAN model attack," *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 18, no. 3, pp. 871–881, 2021.

[18] Q. Zhao, T. Sheng, Y. Wang et al., "M2Det: a single-shot object detector based on multi-level feature pyramid network," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 9259–9266, 2019.

[19] J. Li, X. Liang, and S. M. Shen, "Scale-aware fast R-CNN for pedestrian detection," *IEEE Transactions on Multimedia*, vol. 20, no. 4, pp. 985–996, 2017.

[20] D. T. Nguyen, T. N. Nguyen, H. Kim, and H. J. Lee, "A high-throughput and power-efficient FPGA implementation of YOLO CNN for object detection," *IEEE Transactions on Very Large Scale Integration Systems*, vol. 27, no. 8, pp. 1861–1873, 2019.

[21] S. Zhou and J. Qiu, "Enhanced SSD with interactive multi-scale attention features for object detection," *Multimedia Tools and Applications*, vol. 80, no. 8, Article ID 11539, 2021.
[22] H. Zhang and X. Hong, "Recent progresses on object detection: a brief review," *Multimedia Tools and Applications*, vol. 78, no. 19, Article ID 27809, 2019.
[23] X. Liang, J. Zhang, L. Zhuo, Y. Li, and Q. Tian, "Small object detection in unmanned aerial vehicle images using feature fusion and scaling-based single shot detector with spatial context analysis," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 6, pp. 1758–1770, 2020.
[24] Q. An, Z. Pan, L. Liu, and H. You, "DRBox-v2: an improved detector with rotatable boxes for target detection in SAR images," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 57, no. 11, pp. 8333–8349, 2019.
[25] S. H. I. Wen-xu, T. A. N. Dai-lun, and B. A. O. Sheng-li, "Feature enhancement SSD algorithm and its application in remote sensing images target detection," *Acta Photonica Sinica*, vol. 49, no. 1, Article ID 0128002, 2020.