Advances in Artificial Intelligence, Big Data and Algorithms
G. Grigoras and P. Lorenz (Eds.)

doi:10.3233/FAIA230829

Based on Improved YOLOv7 Small Target Detection

Ruohan WANG[1] and Yu LI
Wuhan Textile University, WuHan 430200, China

Abstract: In recent years, deep learning-based object detection has developed rapidly, but its performance in the small object detection field is not ideal compared to the natural scene image domain. Hence, this paper proposes an improved small object detection algorithm based on YOLOV7 algorithm. Firstly, based on Ghost model, the ELAN module in backbone structure is improved to Gmodel which effectively reduces computation and improves accuracy. Secondly, this paper introduces a Triplet Attention-improved small object attention module Amodel in YOLOV7's head structure; through Amodel's cross-latitude interaction function, it enhances the feature detection performance for small objects. Experiments were conducted on RSOD dataset and our method increased yolov7's AP50 by 1.65mAP and AP50-95 by 1.88mAP while also reducing FLOPs by 0.2G, making it more suitable for dense small target scenes for object detection.

Keywords. small object detection, attention mechanism, remote sensing images, YOLOV7

1. Introduction

In recent years, with the rapid development of deep learning, remarkable progress has been made in the field of computer vision [1,2], and some neural network models have made huge strides in detection speed [3,4], making object detection technology more widely used in fields that require real-time detection such as autonomous driving.

However, the accuracy of small target detection remains a challenging problem in the field of object detection. For example, YOLOv7 [5], which was proposed in 2022 as the fastest detector yet still only achieved 29.6mAP on small target detection - half that of medium and large targets - on COCO [6] , reaching 69.7mAP and 55.9mAP respectively for medium and large targets, but only 29.6mAP in small target detection, which is only half of the accuracy of medium and large targets.This kind of situation in the present is the most advanced detector, this paper argues that this kind of situation appears the reason has 3 points:1.Convolutional Neural Networks, CNN as a backbone network, will be the sampling operation to compress feature information, this operation will lead to the characteristics of the small target is inherently less information by deleting a lot Minus, may even be cut off as a noise screen in the image.2.On the high-compression low-feature map, the feature information of small targets is likely to be misunderstood, resulting in inconsistency with the features of the original image [16]. 3.The location information of small target is difficult to find accurately, and small

[1] Corresponding Author, Ruohan WANG, Wuhan Textile University, China;
Email: 2115363019@mail.wtu.edu.cn

factors such as external noise will have a huge impact on it.. This situation is also present in today's most advanced detectors.

After referring to a large number of attention mechanisms, this paper proposes an improvement for small target detection, which aims to improve the detection accuracy of the yolov7 network against small targets while reducing the computation amount (FLOPs) of the original network detection head.

First of all, our network can detect the feature information of small targets in the low-feature graph, but the recognition ability of the feature information is insufficient, so it cannot accurately detect the target. Therefore, we use some attention mechanism methods to strengthen the detector's recognition ability for the feature information of small targets. In the detection header of yolov7, we replace part of CNN convolution with our improvement The attention module Amodel.

However, in most cases, the features of small targets are less and the distribution is scattered, which makes the calculation amount of low feature graphs very high. Our method is to improve the ELAN structure of the head in the trunk into the convolution module Gmodel with smaller calculation amount.

Our experiments were mainly conducted on the RSOD dataset [8], a drone remote sensing image small target dataset created by Wuhan University, which contains a large number of small targets. The final test results showed that our method can significantly improve the detection performance while reducing the computation cost. In summary, our contributions are mainly twofold:

- We improved the ELAN module of YOLOv7's main network and introduced our smaller convolutional module Gmodel, which reduced the computational cost of the detector while maintaining accuracy.
- We modified the CNN structure of YOLOv7's detection head and enhanced its ability to capture small target features by introducing our attention module Amodel, thereby improving its accuracy for small targets.

2. Methods

2.1. YOLOV7 Methods

YOLOV7 is an improvement on the structure of yolov5 [17] , so the structure of the two is similar. The only difference is that the neck layer and the head layer in yolov5 are integrated into a head layer, but there is no difference in function.

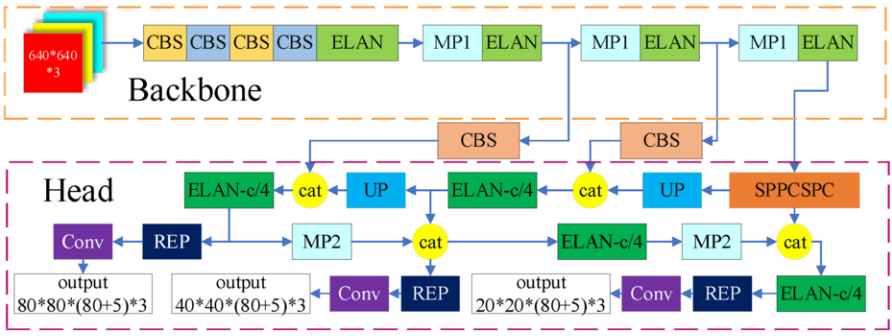

Figure 1. YOLOV7 Network Structure Diagram

The overall network structure is shown in Figure 1. It can be seen from the figure that the yolov7 network is composed of three parts :input,backbone and head.The functions of each part are exactly the same as those of yolov5. For example, backbone is the backbone feature extraction network, and head is the prediction head used for prediction.

As shown in Figure 2, the shows ELAN layer is a module composed of CBS layers of several kernels with the same input and output channels.

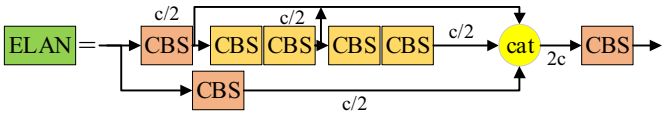

Figure 2. ELAN layer structure diagram

2.2. Feature Extraction Network Based on Lightweight Method

In this section, we will first introduce the Ghost module, which is a very efficient architecture and high performance GhostNet [11] proposed by Kai Han,Yunhe Wang et al.In this modular method, some small filters are used to generate more feature mappings in the original convolution layer, so as to reduce the amount of network computation.

Deep convolutional neural network [12] is a network formed by sequential accumulation of convolutional layers. Too many convolutional layers and too deep network depth lead to a large increase in the amount of computation.In recent years, lightweight networks, such as MobileNet[14] and ShuffleNet[15], have made the original CNN structure more efficient by adding deep convolution and shuffle operation, combined with the application of smaller convolutional filter (floating-point number operation), thus reducing the amount of computation.However, it does not solve the problem that the 1X1 convolution layer consumes a lot of computation.

As shown in Figure 3, we replaced the first two CBSs of ELAN module with Ghost model; this operation can effectively compress input information. Subsequently, both the third and fifth layer CBSs on the main branch were replaced with Ghost model modules with same kernel size to form two CBS+Ghost model structures; this way can reduce computation while maintaining a certain degree of accuracy.

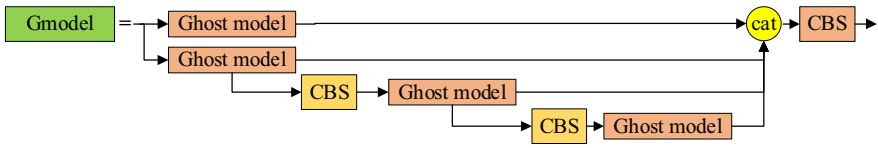

Figure 3. Gmodel module structure based on ELAN module improvement

2.3. Multi-scale Analysis Head Based on Attention Mechanism

In this section, our main research goal is to establish an inexpensive yet effective small object attention module without significantly increasing the network's computation. Common attention mechanisms such as CBAM [9] and SENet [7] rely on a sufficient number of learnable parameters to create dependencies between channels within the network to improve its attention mechanism for targets. In our study, such an

improvement clearly does not meet our requirements; therefore, we referred to Triplet Attention [10], a nearly parameter-free attention mechanism proposed by Diganta Misra et al.

We improved the CBS module to obtain the Amodel module based on the structure design and small target detection characteristics of yolov7 head, as shown in Figure 4.In this paper, the activation function in the CBS module is improved into a new structure of TripletAttention + FReLu [13] , We choose to use the FReLu method because the area and size of the Receptive field are crucial in small target visual recognition tasks. Therefore, compared with convolution or attention mechanism, the FReLu method introduces the Receptive field into nonlinear activation in a more concise and effective way.

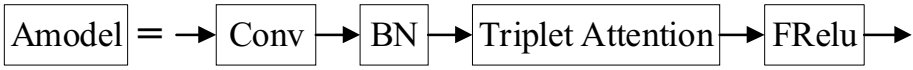

Figure 4. Amodel module structure based on CBS module improvement

In previous studies on effective receptive field, people found that each pixel of an image makes different contributions to the detection effect, and the central pixel makes the most contribution. The cross-latitude interaction of Big TripletAttention was also one of the main reasons we chose it as an improved yolov7 head.Common attention, such as CBAM and SENet, will calculate singular weights when processing features. The main function of singular weights is to uniformly shrink the input feature mapping. In traditional algorithms, every channel scalar of the input tensor will be shrunk.

Although the computation amount of the attention module using this algorithm is much less than that of the convolution module with the same improvement, it can be said that it can already play a role in lightweight and improving network performance.However, this method has a major defect, in order to calculate the singular weight of the scaled feature map, the input tensor needs to be decomposed, and then the global average is obtained to obtain a scalar of each channel.The operation of decomposition and global averaging causes the loss of a large amount of spatial information in the feature input, which leads to the failure in the subsequent integration of spatial and channel latitude information.

2.4. The Improvement of YOLO7 Structure

As shown in Figure 5, the improvements of YOLO7 are highlighted by the dashed circles, which are the parts we have modified.

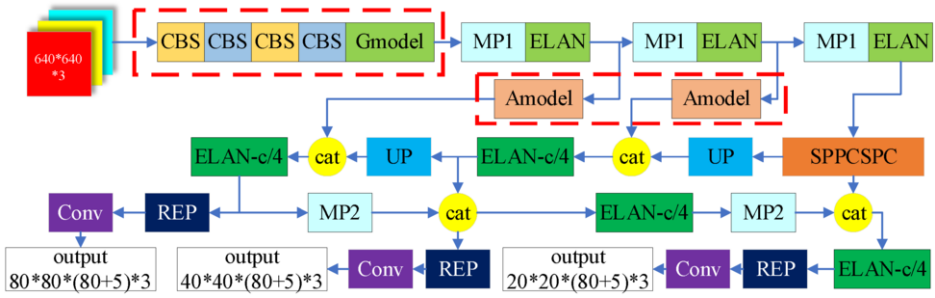

Figure 5. The overall structure Figure of YOLOv7 improvement

In order to distinguish complex background from small targets and reduce the computational amount of feature extraction network in extracting small target feature information to a certain extent, we replace our improved GModel module with the first ELAN structure in backbone structure. Here, we mainly want to improve the small target feature information on low-level feature map, because there are a lot of small items in the dataset. The extraction of low-level features will occupy a large amount of computation, so it can also reduce the computation amount of feature extraction network.

In the head structure of yolov7, we replaced the CBS module that received the input of low - and medium-level feature graphs with the improved AModel module. This is because in the small target detection, the structure of FPN class will be detected from the high - resolution low - level feature graphs, and the sparse distribution of small targets in the pictures will result in the calculation effect of low - resolution high - level feature graphs. It is very low, so that the localization and recognition of small targets are basically completed in the low-level feature map.Such an improvement can improve the detection of small targets and control the increase of computation caused by the introduction of attention.

3. Experiments

This paper presents a quantitative experiment on the Wuhan University UAV Small Target RSOD dataset. RSOD is a large-scale public dataset for small target detection tasks in urban environments. This dataset contains 22,091 manually annotated instances and 5 classes, and due to its geographic information and annotation orientation, it will provide researchers with a challenging detection task.

3.1. Implementation Details

This paper implements our method using the PyTorch framework, just like YOLOv7. All models were trained on an RTX 3090 Ti GPU, so this paper adds YOLO and COCO file formats to the RSOD dataset.

The same approach as YOLOv7 was used to train the improved model on the RSOD dataset with a standard 1x schedule, Detectron2 default data augmentation mode, batch size set to 16 and an initial learning rate of 0.01 for a total of 300 training rounds.

3.2. Effectiveness of Our Approach

In order to verify the superiority of the improved object detection algorithm compared with other algorithms,In Table 1, this paper mainly records the mean average precision (mAP) of the algorithm, takes the test results of yolov7 as the baseline, and compares the mainstream network model under the current single-phase target detection system with our improved yolov7 model.

Table 1. To verify the superiority of the improved target detection algorithm in this paper compared to other algorithms, we records the mean average precision (mAP) of the algorithm. Taking Yolov7's test results as a baseline, we compared mainstream network models under the current single-stage target detection system with our improved Yolov7 model.

Method	mAP50/%	mAP50-95/%
SSD	86.90	59.91

Retinanet	92.75	64,14
Yolov5	94.63	66.52
Yolox	94.74	67.04
Yolov7	94.92	66.97
Ours	96.57	68.85

It can be concluded from Table 1 that the improved algorithm proposed in this paper is superior to other popular algorithms in terms of detection accuracy under the premise of reducing the calculation amount of the model. After the original yolov7 network is trained on the small target-oriented data set of RSOD,AP50 and AP50-95 obtain 94.92mAP and 66.97 map, respectively mAP results.However, our improved model has the best performance in the yolo system, achieving 96.57mAP and 68.85mAP in AP50 and AP50-95 respectively, which are 1.65mAP and 1.88mAP higher than yolov7, because our improvement focuses on the location detection of small targets in the AP50In terms of mark detection, the detection accuracy of location information has achieved a very high achievement, but the improvement is less than that of AP50-95.

3.3. Ablation Studies

As shown in Table 2, the influence of each improvement part on small target detection accuracy and network computation is analyzed.The yolov7 model separately added to Gmodel was retrained in this paper. In the test set, the AP50 and AP50-95 of the training results reached 96.45mAP and 67.28mAP respectively, while FLOPs decreased from 11.3G to 11.1G.Overall, there is a slight AP enhancement due to the ELAN structure improvements in Section 2.1.

Table 2. In this paper, an ablation study was conducted on the RSOD test set. the effects of each improvement part on small target detection accuracy and network computation are analyzed.

Method	Amodel	Gmodel	mAP50-95/%	FLOPs
YOLOv7	×	×	66.97	11.3G
Improvement of 1	√	×	67.38	11.9G
Improvement of 2	×	√	67.28	11.1G
Improvement of 3	√	√	68.85	11.1G

Then, the yolov7 model added by Amodel was retrained in this paper. The AP50 and AP50-95 produced by the training reached 96.34mAP and 67.38 mAP respectively, which increased by 1.42mAP and 0.41mAP respectively compared with yolov7, indicating that the attention mechanism was in The improvement of small target detection capability in head structure is very important

4. Conclusion

This paper has improved the YOLOv7 algorithm for small object datasets. The Ghost module was introduced into the backbone to enhance feature extraction capabilities while reducing computational costs; subsequently, an attention mechanism was introduced into the head structure, incorporating our attention-based Amodel to focus

more attention on densely populated small object areas and improve small object feature detection capabilities.

Through experiments, compared to existing object detection algorithms, this paper's improved algorithm effectively enhances YOLOv7's ability to detect small objects and reduces the computational cost of feature extraction networks, making our modified model more advantageous in processing small object image target detection tasks.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun.Deep residual learning for image recognition. In CVPR,2016.
[2] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. Aggregated Residual Transformations for Deep Neural Networks. In CVPR, 2017.
[3] Joseph Redmon and Ali Farhadi. Y olov3: An incremental improvement. arXiv preprint arXiv:1804.02767, 2018.
[4] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun.Faster r-cnn: Towards real-time object detection with region proposal networks. In NeurIPS, 2015.
[5] Wang C Y, Bochkovskiy A, Liao H Y M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors[J]. arXiv preprint arXiv:2207.02696, 2022.
[6] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays,Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In ECCV. Springer, 2014.
[7] Mingxing Tan, Ruoming Pang, and Quoc V Le. EfficientDet:Scalable and efficient object detection. In CVPR, 2020.
[8] Y. Long, Y. Gong, Z. Xiao and Q. Liu, "Accurate Object Localization in Remote Sensing Images Based on Convolutional Neural Networks," in IEEE Transactions on Geoscience and Remote Sensing, vol. 55, no. 5, pp. 2486-2498, May 2017. doi: 10.1109/TGRS.2016.2645610
[9] Jingdong Wang, Ke Sun, Tianheng Cheng, Borui Jiang,Chaorui Deng, Yang Zhao, Dong Liu, Yadong Mu, Mingkui Tan, Xinggang Wang, et al. Deep high-resolution representation learning for visual recognition. IEEE transactions on pattern analysis and machine intelligence, 2020.
[10] Misra D , Nalamada T , Arasanipalai A U , et al. Rotate to Attend: Convolutional Triplet Attention Module[J]. 2020.
[11] Kai Han, Yunhe Wang, Qi Tian, Jianyuan Guo, Chunjing Xu, and Chang Xu. Ghostnet: More features from cheap operations. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), June 2020.
[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In NeurIPS, pages 1097–1105, 2012.
[13] Ningning Ma, X. Zhang, and J. Sun. Funnel activation for visual recognition. 2020.
[14] Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv preprint arXiv:1704.04861, 2017.
[15] Xiangyu Zhang, Xinyu Zhou, Mengxiao Lin, and Jian Sun. Shufflenet: An extremely efficient convolutional neural network for mobile devices. CVPR, 2018.
[16] Yang hao Li, Y untao Chen, Naiyan Wang, and Zhaoxiang Zhang. Scale-aware trident networks for object detection.In ICCV, 2019.
[17] Jocher Glenn. YOLOv5 release v6.1. https://github.com/ultralytics/yolov5/releases/tag/v6.1, 2022.