FEATURE-ENHANCED SMALL-TARGET DETECTION

Gao YuHang, Guo Hanasi
Belarusian State University Department of Applied Mathematics
Department of Business Administration, Business School, Belarusian State University
Minsk, Republic of Belarus
E-mail: g15535458181@gmail.com, 3532803499@qq.com

Detecting small targets from images is still a challenging problem in computer vision due to the limited size, few appearance and geometric cues, and the lack of large-scale small target datasets. To address this problem, an adaptive feature-enhanced target detection network (YOLO-FENet) is proposed to improve the detection accuracy of small targets. Firstly, an improved adaptive two-way feature fusion module is designed by introducing a feature fusion factor to make full use of the feature maps of various scales to improve the feature expression ability of the network; secondly, a spatial attention generation module is proposed by combining the characteristics of the network, which improves the feature localization ability of the network by learning the positional information of the region of interest in the image. The experimental results on the UAVDT dataset show that the average precision (AP) of the proposed YOLO-FENet is 6.3 percentage points higher than that of the pre-improvement YOLOv5, and it is also better than other target detection networks.

Introduction

Target detection techniques are the basis of many computer vision tasks such as instance segmentation, image captioning, target tracking, etc., and have been widely used in the fields of video surveillance, automatic driving, medical diagnosis, etc [1]. Deep learning has made significant contributions to the development of target detection technology, the existing target detection technology is mostly based on deep learning technology [2], which can be divided into two categories: 1) two-stage target detection algorithms (two-stage), such as the RCNN series [3-5] has the advantage of high accuracy; 2) single-stage target detection algorithms (one-stage), The detection speed is relatively faster, such as YOLO series [6-9], SSD series [10], and so on.

I. Related Work

YOLOv5 is the 5th generation target detection algorithm of YOLO series: it includes YOLOv5s, YOLOv5m, YOLOv5l, YOLOv5x and other networks of different sizes. Compared with YOLOv4, YOLOv5 introduces the focus module and the adaptive anchor frame calculation before training the network, which has the features of small weight file, short training time and faster inference, and its overall structure is shown in Figure 1.

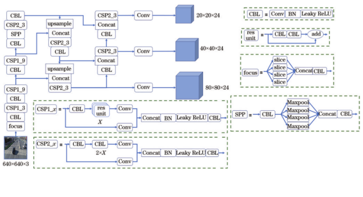

Figure 1 – Details ofYOLOv5 network

YOLOv5 consists of four parts: input, backbone, neck and head.The input part of YOLOv5 consists of two main parts: data enhancement and adaptive anchor frame calculation. Data enhancement includes input image resizing, color space conversion, and mosaic data enhancement: adaptive anchor frame computation first computes the maximum possible recall (BPR) of the default anchors, and if the BPR is less than 98%, the anchors are updated using K-means and genetic algorithms. In the feature extraction network, YOLOv5 adds a new focus module to reduce the number of parameters and computational complexity of the network.YOLOv5 adopts the path aggregation network (PANet) structure as the multiscale feature fusion module with top-down and bottom-up features.The convolution part adopts the CSP structure which is different from that of the backbone network for better diversity and robustness. The detection head is responsible for localizing and identifying the target and outputting the final detection result.

II. YOLO-FENet Model Analysis

The YOLOv5 target detection algorithm has received a lot of attention for its excellent performance in terms of model size and detection speed. At the same time, YOLOv5 experiments on the PASCAL VOC and COCO datasets demonstrate that the network PASCAL VOC and COCO datasets demonstrated the advantages of the network in terms of detection accuracy.YOLOv5 has demonstrated the advantages of the network in terms of detection accuracy. However, the research content of YOLOv5 However, the research of YOLOv5 is on generalized target detection, and the network design has not taken much consideration of the characteristics of small targets, so the network needs to be improved.Therefore, the network needs to be improved and optimized to be suitable for small target scenarios. The proposed YOLO - AFENet consists of a feature extraction network, an improved self adaptive multi-scale fusion network, and a spatial attention generation network (SAGN),The overall structure of the network is shown in Figure 2.

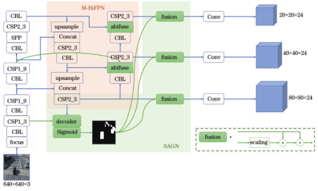

Figure 2 – Details ofYOLOv5 network

YOLOv5 implements a multi-scale feature fusion module using the PANet structure, as shown in Fig. 3(a). This structure improves the detection accuracy by bi-directionally fusing the semantic information of the deep feature maps and the localization information of the shallow feature maps, but it introduces more parameters and computations, which reduces the efficiency of the model, and PANet simply assumes that feature maps at different stages have the same contribution to the final fusion result, so that the effect of feature fusion needs to be further improved. In order to improve the model efficiency and fusion effect, this study combines the idea of BiFPN and designs an improved MBiFPN based on PANet, the specific structure is shown in Fig. 3(b). Among them, Pi represents the feature map of layer i in the backbone layer, Oi represents the output feature map of the neck module, and Fi and Ni represent the intermediate feature maps generated by the neck layer.

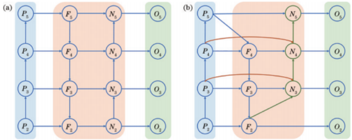

Figure 3 – Structures of PANet and MBiFPN. (a) PANet; (b) MBiFPN

III. Experimental results and analysis

In order to verify the validity and reliability of the proposed method, this experiment adopts the dataset UAVDT proposed by Du et al. at the 2018 European Conference on Computer Vision, which can be divided into three categories of detection targets, namely, cars, trucks, and buses, and is composed of UAVs in a variety of complex scenes The dataset contains more common scenes such as plazas, highways, and intersections. The dataset contains 50 video clips, of which 30 video sequences are used for training (containing 23829 images), and the remaining 20 video sequences are used for testing (containing 16580 images), with a resolution of 1024×540 per frame.In addition, the dataset provides attributes such as weather, view angle, and altitude.In addition, the dataset is labeled with attributes such as weather, viewing angle and altitude. The relationship between the number of small, medium and large targets in the dataset is shown in Table 1

Table 1 – Size Distribution of object in dataset

object	Small object	Medium object	Large object	Total
number	61.9%	36.4%	1.7%	100%

IV. Enhancing Model Performance

The loss function of YOLOFENet consists of 4 parts, which are location loss, confidence loss, classification loss and attention loss:

$$L_{\text{total}} = \alpha L_{\text{CIoU}} + \beta\Big(L_{\text{obj}} + L_{\text{SAGN}}\Big) + \gamma L_{\text{class}}$$

where: α,β,γ are the weights of the corresponding loss functions, taking the values of 0. 05, 0. 07 and 0. 03. LSAGN denotes the loss of attention and is used to optimize the network The SAGN module is used to generate foreground/background spatial attention maps, which can be regarded as a two-class image segmentation problem, so the loss function is a binary cross-entropy loss function.

$$L_{\text{SAGN}} = -[y_i \log \hat{y}_i + (l - y_i) \times \log(l - \hat{y}_i)],$$

yi denotes the true label of sample i; yi is its corresponding network predictive label.is its corresponding network prediction label.

V. Conclusion

The proposed small target detection algorithm improves the detection accuracy while ensuring that the detection speed is not significantly reduced. However, the algorithm still has some problems, such as the loss function does not consider The loss function does not consider the characteristics of the long-tailed distribution of the dataset, The algorithm's detection performance for large targets is not obvious, etc. Therefore, the network design can be further optimized to improve the detection performance of the network.

1. Zou Z X, Shi Z W, Guo Y H, et al. Object detection in 20 years: a survey[EB/OL]. (2019-05-16) [2021-11-15]. https://arxiv.org/abs/1905.05055.

2. Jiao L C, Zhang F, Liu F, et al. A survey of deep learning-based object detection[J]. IEEE Access, 2019,7: 128837-128868.

3. Ren S Q, He K M, Girshick R, et al. Faster R-CNN:towards real-time object detection with region proposal networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149

4. Girshick, R., Donahue, J., Darrell, T., et al. (2014) Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. 2014 IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 23-28 June 2014,580-587. https://doi.org/10.1109/CVPR.2014.81

5. He, K., Zhang, X., Ren, S., et al. (2014) Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. IEEE Transactions on PatternAnalysis and Machine Intelligence, 37,1904-1916. https://doi.org/10.1109/TPAMI.2015.2389824.

6. Ren, S., He, K., Girshick, R., et al. (2017) Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. IEEE Transactions on Pattern Analysis & Machine Intelligence, 39, 1137-1149. https://doi.org/10.1109/TPAMI.2016.2577031