Small Target Detection Based on YOLOX

Keying Ren
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China

Xiaoyan Chen
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China

Zhihui Chen
School of Electronic Information and Automation, Tianjin University of Science and Technology,
1038 Dagu Nanlu, Hexi District, Tianjin, China

E-mail: renkeying@mail.tust.edu.cn, cxywxr@tust.edu.cn, 1594838831@qq.com

www.tust.edu.cn

Abstract

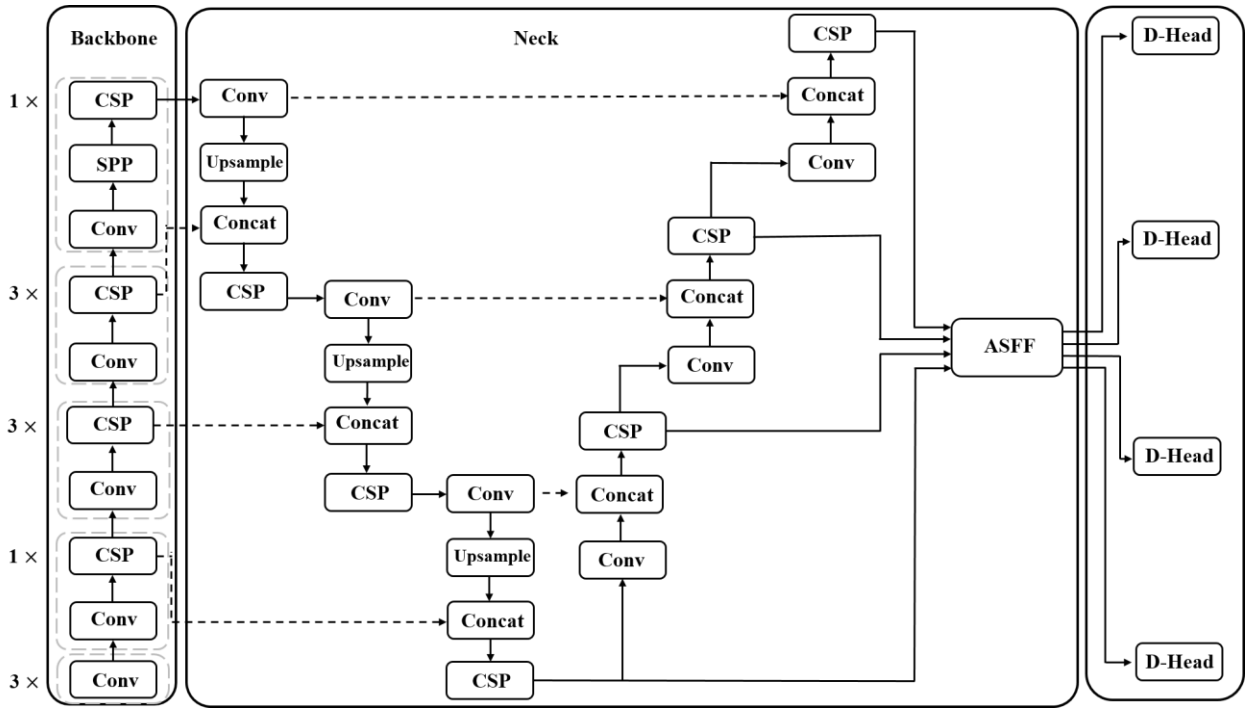
With the development of drone, small target detection has become a hotspot of current research. In this paper, the network of small target detection is based on YOLOX is studied. There are a lot of small targets in the images taken by UAV, which brings great difficulty to the detection task. The main improvement of the proposed network is that, based on the original YOLOX network, four-scale adaptive spatial feature fusion pyramid is added to filter the conflicting information between different scales and improve the expressive accuracy of the small target features. Experiments show that the proposed method has average accuracy of 32.83% in the self-built dataset, which is 2.53% higher than that of YOLOX-s, 1.53% higher than that of Yolov6-s, and 5.73% higher than that of YOLOv5-s.

Keywords: Drone shooting, Small target detection, YOLOX, Adaptively Spatial Feature Fusion

1. Introduction

With the development of artificial intelligence technology, the field of computer vision has made a great breakthrough. Object detection is one of the main tasks of computer vision, which has been applied to pedestrian detection, face detection and other tasks. Object detection algorithms include SSD (Single Shot MultiBox Detector)[1], CornerNet[2], YOLO (You Only Look Once)[3]series, etc. These algorithms can directly classify and locate objects. The one-stage algorithm has faster detection speed than the two-stage algorithm, but the error rate and missing rate are relatively high. In MS COCO[4], the accuracy of small-target detection is less than half in the large/medium target detection. Therefore, it is an urgent problem to improve the accuracy of small targets in UAV aerial images.

In 2020, Nayan et al. proposed a small target detection algorithm based on YOLOv3, which used upsampling and residual joint to extract multi-scale features of different convolution levels in learning tasks[5]. This method significantly improves the ability of small target detection. In the same year, an Enhanced Context Model (ECM) was proposed, which used a double-dilation convolution structure to reduce the number of parameters, expand the effective receptive field, strengthen the context information, and apply to the prediction layer of the network. However, this method depends on the



design of the context window or the size of the receptive field, which may result in the loss of important context information. Wang Jianjun et al. studied an improved YOLOv3 small target detection algorithm, It increases the depth of shallow convolution in backbone networks to enhance the feature extraction capability of backbone. The RFB (Receptive Field Block) structure is used to enlarge the receptive field of the shallow feature map. Megvii Technology's YOLOX algorithm uses anchor free method and does not need to set anchors in target detection tasks. It combines data enhancement, simOTA label allocation strategy and so on. YOLOX has become a mainstay in recent years.

In order to improve the accuracy of small target detection, this paper proposes a small target detection algorithm based on YOLOX, which has the following characteristics:

1.Add a small target detection layer based on the original YOLOX structure.

Fig.1 Structure comparison of YOLOX

2.A four-scale adaptive spatial feature fusion pyramid is used to filter the conflicting information between different scales.

2. Method

The overall structure of this method is shown in Figure 1, which mainly includes backbone, Neck and decoupling detection head.

2.1. *backbone*

The backbone is used to take the input image and extract useful features from it. This is a very important step in the object detector because it is the main structure for extracting context information. Most of the backbone of traditional YOLO algorithm adopts residual connection [6]. Identity mapping is used to solve the problem of network training difficulty

Our backbone uses the structure of CSPNet. its small model uses a basic building block ratio of 1,3,3,1. Four stages were used in the backbone to obtain different scale features. At the end, the SPP acquires features at different scales and merges them.

2.2. *Neck*

In convolutional networks, deep networks are sensitive to semantic features, while shallow networks are sensitive image features. This feature of convolutional networks often brings a lot of trouble in object detection.

The high-level network can respond to semantic features. However, because the size of feature map is too small to get detailed geometric information, which is not conducive to target detection.

To solve this problem, YOLOX introduced PANet[7], which uses top-down feature fusion followed by bottom-up feature fusion. Through this operation, we can achieve better detection results.

2.3. *Decoupled head*

The existing target detector still uses coupling head to realize object detection. YOLOX introduced the

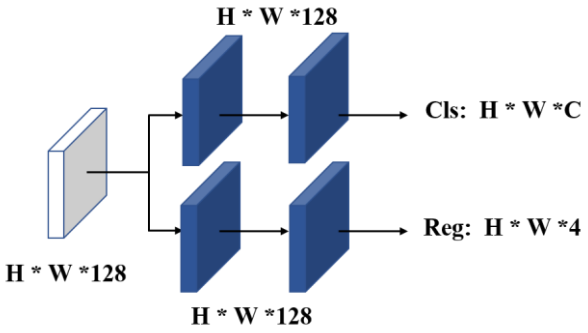


Fig.2 decoupling Head

decoupling detection head, whose structure is shown in Figure 2. In this structure, 1×1 convolution was used to compress the feature channels to 128, and two 3×3 convolution were used respectively to extract the features required for classification and regression tasks.

2.4. *Adaptively spatial feature fusion*

In object detection, feature pyramid is used for feature fusion of different scales. However, this operation results in inconsistent information between different scale features. This phenomenon increases the difficulty of small target detection. Therefore, adaptive spatial feature fusion pyramid (ASFF) is adopted in this paper to filter the spatial conflict information between different scale features, so as to improve the inconsistency between different scale features.

3. Experiments

3.1. *Dataset*

The images used for the experiment are taken by drones at different locations and altitudes. There are 9,563 images, 11 categories, and more than 560,000 pre-labeled anchors.

3.2. *Experiments and Results*

Stochastic gradient descent (SGD) was used in training, and the parameters were set as follows: momentum=0.9, weight decay=5e-4. Cosine annealing learning rate was used, initial learning rate was set to zero, warm-up training for 5 epoch, the basic learning rate was 0.01, and the learning rate was adjusted by linear scaling. In the training process, the exponential moving average (EMA) strategy was used, and its attenuation was 0.9998.

Common Objects in Context (COCO) evaluation criteria are most commonly used in target detection, which can best reflect the performance of a target detection algorithm. COCO evaluation criteria are divided into average precision (AP) and average recall (AR). AP and AR are subdivided into indexes under different parameters. Among them, AP is divided into mean average precision (mAP), AP50, AP75. mAP is the most commonly used index among these evaluation indexes and can best reflect the performance of the detector.

In order to compare the effectiveness of the proposed algorithm, a comparative experiment is conducted in this paper. COCO evaluation index is used to evaluate. The results are shown in Table 1.

Table 1 Detection results of different algorithm

Method	mAP（%）	AP50（%）	AP75（%）
YOLOv5-s	27.1	47.6	28.4
YOLOX-s	30.3	59.5	32.5
YOLOV6-s	31.3	60.7	33.6
our	**32.83**	**62.36**	**34.54**

Experiment results reported in Table 1 is the average of multiple experiments. Experiments show that the proposed method has a mAP average accuracy of 32.83% in the self-built dataset, which is 2.53% higher than that of YOLOX-s, 1.53% higher than that of Yolov6-s, and 5.73% higher than that of YOLOv5-s.

4. Discussion

In order to effectively detect small targets in aerial photography, this paper proposes a small target detection algorithm based on YOLOX. A four-scale adaptive feature fusion module (ASFF) was used to filter the conflicting information between different scale features, so as to improve the efficiency of the detector. Although the method proposed in this paper has improved the performance index, there is still room for improvement in practical application. For example, optimize the detection effect on complex scene data set.

Acknowledgment

This work was supported by "Tianjin University of Science and Technology – softsz Intelligent edge computing Joint Laboratory"

References

1. Liu W , Anguelov D , Erhan D , et al. SSD: Single Shot MultiBox Detector, *European Conference on Computer Vision. Springer, Cham*, 2016: pp.21-37.
2. Law H, Deng J. Cornernet: Detecting objects as paired keypoint, *Proceedings of the European conference on computer vision (ECCV)*, 2018: pp 734-750.
3. Redmon J , Divvala S , Girshick R , et al. You Only Look Once: Unified, Real-Time Object Detection, *IEEE conference on computer vision and pattern recognition*, 2016: pp. 779-788.
4. Lin T Y, Maire M, Belongie S, et al. Microsoft coco: Common objects in context, *European conference on computer vision*, 2014: pp 740-755.
5. Nayan A A, Saha J, Mozumder A N, et al. Real time detection of small objects[J]. International Journal of Innovative Technology and Exploring Engineering, 2020, 29(5): pp 14070-14083.
6. He K, Zhang X, Ren S, et al. Deep residual learning for image recognition, *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016: pp 770-778.
7. Liu S, Qi L, Qin H, et al. Path aggregation network for instance segmentation, *Proceedings of the IEEE conference on computer vision and pattern recognition.* 2018: pp 8759-8768.

Authors Introduction

Mr. Keying Ren

He is an Master of Control Science and Engineering, Tianjin University of Science and Technolog. The research topic is target detection and tracking based on deep learning.

Prof Xiaoyan Chen

She has been working on the application of deep learning network models to the field of computer vision for many years. She is currently a Professor with the Tianjin University of Science and Technology.

Mr. Zhihui Chen

He received his bachelor's degree from the school of electronic information and automation of Tianjin University of science and technology in 2021. He is acquiring for his master's degree at Tianjin University of science and technology.