IT-YOLOX: Object Detection Algorithm in UAV Perspective

Shuo Xu

Abstract—**Recently, drone scenes have been widely used in rescue, agriculture and other industries, and drones can fly at different altitudes to obtain multiple fields of view and viewing angles. However, since most of the drones are shot at higher altitudes, there are many small targets and objects in the shooting images. In view of these difficulties, this paper proposes a new network model :IT-YOLOX, which is based on YOLOX-tiny. First, an additional detection layer is added to the original three detection layers of YOLOX-tiny to detect small target objects. Secondly, iAFF module is used in YOLOX-tiny neck to fuse the feature information of backbone network. Finally, transformer structure is used in the last part of the backbone network. This paper conducted experiments on VisDrone2019, and the experimental results proved that the improved network improved by 5.17% compared with the baseline model (YOLOX-tiny) mAP, especially strengthening the recognition ability of small targets.**

Index Terms—**UAV image, object detection, YOLOX.**

I. Introduction

The main task of object detection[1] is to determine the category and location of the target, which is often used in computer vision fields such as human detection, pedestrian detection, vehicle detection and remote sensing monitoring. At present, the classical target detection methods mainly include one-stage and two-stage. The two-stage target detection method divides the detection problem into two stages. Firstly, the candidate region is generated, and then the candidate region is classified after location refinement. The two-stage detection algorithm has a low identification error rate but a slow speed, which cannot meet the real-time detection scenario. The common two-stage target detection algorithm is RCNN[2] series. one-stage object detection method does not need to produce candidate region stage, but directly generates the category probability and position coordinate value of the object, and can directly obtain the final detection result after a single detection. Therefore, it has a faster detection speed and can adapt to more application scenarios. The common one-stage object detection algorithm is YOLO[3] series.

In recent years, with the rapid development of deep learning and computer vision, object detection algorithms have been widely used in many fields. In particular, the target detection algorithm from the perspective of UAV is widely used in practical scenarios such as agriculture, traffic, military and rescue.

In agriculture, the introduction of drones can carry out pesticide spraying operations, because of its small size, good mobility, flexible adjustment of flight speed, spraying height and other advantages, not only can avoid damage to crops by large machinery, ensure crop production, but also improve the uniformity of pesticide spraying, reduce the waste of pesticides, and greatly improve the spraying efficiency of pesticides. In terms of traffic, when the UAV flies in the air and detects the ground, it can not be restricted by the road on the ground, and has the advantages of wide field of vision and high degree of freedom. When a traffic accident occurs, the image can be collected at the scene in the first time, and can be analyzed in time, and the results can be sent back to the traffic management department.

In the military aspect, UAVs, with their advantages of low cost, strong mobility and effective reduction of casualties of combatants, are widely used in reconnaissance and early warning, communication relay, military strikes and other fields, and can perform a series of tasks such as ultra-long distance reconnaissance, strike and damage, which has become the focus of weapons and equipment development of the world's military powers. In terms of rescue, the UAV is equipped with zoomable high-definition camera equipment, adjusts the hovering height and shooting Angle, realizes multi-dimensional shooting and all-round scene detail display, and can also carry out emergency mapping of the scene, and send the accident scene situation back to the scene headquarters through the picture transmission equipment, and grasp the information of the disaster accident scene in the first time.

Although UAVs can capture richer images, they also bring great challenges to target detection tasks. Due to the flexible flight height and different shooting angles of UAV, the target scale is unbalanced, there are many small target objects, and there are problems of dense occlusion and blurring. There is the problem of class imbalance in UAV image. And the brightness difference between the daytime background and the night background is large. For example, in the VisDrone-2019 dataset, "cars" and pedestrians account for a large part of the dataset, while "buses" and "awning tricycles" account for a small proportion. In addition, the distribution of pixels in this dataset shows that about 60% of the pixels are less than 1000, and the remaining 40% are more than 1000; Moreover, there are a large number of blocked objects in this dataset, which account for about 60% of the whole dataset.

In order to solve the above problems, the model proposed in this paper first increases the three detection layers of the YOLO-Head part of YOLO-tiny to four detection layers, then uses three iAFF[4] modules in the neck to fuse the scale information, and finally adds transformer structure to the

Manuscript received February 22, 2024

Shuo Xu, School of Computer Science and Technology, Tiangong University, Tianjin, China

backbone network to enhance the recognition ability of small targets and obscured objects. The main contributions of this work are as follows:

1) On the basis of YOLOX-tiny, a detection layer is added from the original three detection layers to enhance the acquisition ability of small targets and improve the detection accuracy.

2) In the neck of the network, three iAFF modules are used to fuse the features of the backbone network and the neck features, and carry out the feature fusion with inconsistent scales, and obtain the multi-scale fusion feature map.

3) Use transformer structure in the backbone part of the network to understand the relative position and arrangement of objects in the image and better capture location information.

4) This paper was evaluated on the VisDrone2019-DET-test-dev dataset with an image resolution of 640×640. The IT-YOLOX proposed by us reaches 36.2% at mAP, which is about 5.17% higher than mAP when compared with the baseline model (YOLOX-tiny).

II. Related Work

A. *YOLOX[5]*

On the basis of inheriting the advantages of the original YOLO model, YOLOX has better detection accuracy and is more conducive to environment deployment due to its more advanced network structure, Mosaic data enhancement, label matching strategy and Anchor-Free mechanism. YOLOX offers six different network width and depth options, including YOLOX-Nano, YOLOX-Tiny, YOLOX-S, YOLOX -M, YOLOX-L, and YOLOX -X. From left to right, the higher the accuracy of the network model, but the higher the parameter two, the slower the reasoning speed.

The backbone feature extraction network of YOLOX adopts CSP-Darknet and uses the Focus structure. The specific operation of this structure is to obtain a value every other pixel in the special map, so as to obtain four independent feature maps, and then stack the four independent feature maps to expand the width and height information of the feature map into the channel, so that the channel is expanded by four times.

YOLOX enhanced feature extraction network adopts FPN[6] to perform feature fusion on the three effective feature layers obtained from the backbone. Features are not only up-sampled to achieve feature fusion, but also down-sampled again to achieve feature fusion. Finally, YOLO-Head is used to classify and regression the result after feature fusion. The feature map is regarded as a set of feature points, and the feature points are judged to determine whether there are objects corresponding to the feature points.

YOLOX uses the SimOTA label matching policy. The core idea of the strategy is to determine the threshold of positive and negative samples according to the similarity of targets, so as to realize the adaptive matching of samples and set different numbers of positive samples for different targets. SimOTA matching strategy not only shortens the training time, but also reduces the extra parameters in the algorithm. In traditional object detection algorithms, anchor boxes are commonly used to locate and identify objects in images. The so-called anchor boxes are pre-set boxes representing objects of different scales and aspect ratios, which are used for prediction and positioning. However, this method requires manual selection of prior boxes according to different data sets. And this method is easy to be affected by the shape change of the object, especially the irregular shape of the object. Therefore, YOLOX adopts an anchor free method, which no longer needs to design prior boxes in advance to detect objects, but only needs to perform regression on the target center point and width and height of feature maps of different scales. The anchor-less method is not only flexible and general, it can adapt to objects of various sizes and shapes, especially irregular objects, but also reduce the workload of designing prior frames in advance, and also reduce the computational amount and complexity of the model, making it perform better in the real-time scene environment.

B. *VisDrone2019[7]*

The VisDrone2019 dataset is a large-scale, high-quality drone view object detection dataset derived from video footage captured by drones, each consisting of multiple continuous image frames and provided in high-resolution image form. This brings the data set closer to practical applications and can better simulate the visual perception of drones in real environments.

The VisDrone2019 dataset includes 288 video clips, comprising a total of 261,908 frames and 10,209 still images. It was collected using drones in different scenarios, weather and lighting conditions. It covers the landscapes of more than a dozen Chinese cities and includes 10,000 images and 2.6 million annotated messages. This annotation information provides detailed target labeling information, with each target labeled with bounding box coordinates, including the location and dimensions of the target. The dataset includes ten sub-target categories: cars, pedestrians, buses, bicycles, tricycles, awning tricycles, vans, trucks, people and cars. There were 6471 images and their labels in the training dataset, 548 images and their labels in the verification dataset, and 1610 images in the test dataset. This partitioning method is convenient for researchers to train, tune and evaluate the model performance.

In addition to rich target categories and annotation information, the VisDrone2019 dataset also covers a range of complex scenarios, such as extensive occlusion, long-distance shooting, blurred light at night, and some targets being too small. These scenarios make it more challenging to identify data sets and evaluate the performance of algorithms in real-world, complex environments. In addition to this, the VisDrone2019 dataset provides a range of evaluation metrics to measure the performance of target detection and tracking algorithms. These include metrics such as Precision, Recall, and average accuracy mean (mAP), which comprehensively assess the performance of algorithms in different categories and levels of difficulty.

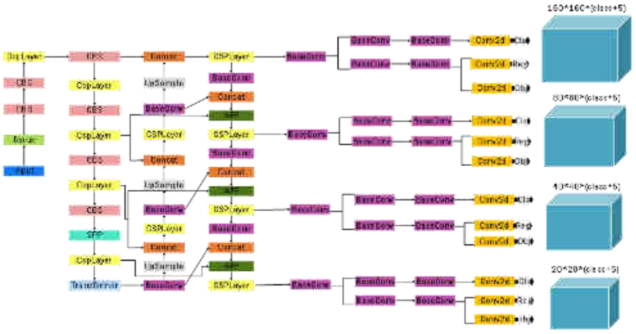
In general, the VisDrone2019 dataset is a UAV vision dataset with multi-target categories, diversified scenes, large-scale data, detailed standard information and multi-modal data, which has a wide range of application value in the research and application of UAV vision. It can be used for the research and evaluation of multiple tasks such as target detection, target tracking, behavior recognition, traffic flow analysis, etc., providing rich resources and benchmarks for the research and application of UAV target detection and tracking.

III. Methods

A. Global network structure

This section mainly introduces the improved overall model architecture of this paper, which is an improved algorithm based on the YOLOX-tiny version. The network structure mainly consists of input part, feature extraction, feature fusion and prediction part. First of all, due to the excessive number of small target objects captured by UAV, we added an additional detection layer for small target detection in the YOLO-Head part, increasing the original three detection layers to four detection layers, which enhanced the ability to improve the identification of small targets. Secondly, the feature layer of the backbone network is fused with three iAFF in the neck to obtain the multi-scale fusion feature map. Finally, transformer structure is added to the backbone network to realize multi-scale information processing. Figure 1 shows the structure of the network.

Figure 1. IT-YOLOX



B. Multi-scale fusion capability

Multi-scale fusion capability refers to the ability to effectively process information at multiple scales in a deep learning model. By using information at multiple scales, the model can better capture the features and context information of objects at different scales. Secondly, the detection and segmentation capabilities of the model for small-scale and large-scale targets can be improved through the detection of receptive fields and feature fusion at different scales.

The iAFF module is mainly composed of MS-CAM module. MS-CAM addresses the scaling problem of channels by using point-by-point convolution (1x1 convolution), rather than convolution cores of different sizes, using point convolution, in order to make MS-CAM as lightweight as possible, plus local and global feature context features in the channel attention module.

The main idea of iAFF is to enhance convolutional neural networks by improving feature fusion through more refined attention mechanisms. The iAFF alternately integrates the initial feature fusion with another attention module, using the attention machine to change the traditional simple feature fusion method.

C. Improvement of YOLO-Head

In summary 2.2, this paper analyzes the Visdrone2019 dataset. In the Visdrone2019 dataset, due to the high shooting, there are many small objects, such as people and cars, and even hundreds of small objects in one picture.

The traditional YOLOX network has only three detection layers, the sizes of which are 80*80, 40*40, and 20*20. In order to enhance the detection capability of small targets, we added an additional 160*160 detection layer for detecting small targets. Compared with the traditional structure of three detection layers, four detection layers can provide more receptive fields and capture richer context information, so that the model can better understand the complex semantic information of the target object. It can greatly reduce the difficulty of small target detection, and the detection effect of small objects has been greatly improved.

D. Transformer Structure

Transformer[8] architecture is a deep learning model architecture for sequence modeling, originally proposed by Vaswani et al, and widely used in natural language processing (NLP) tasks. Compared to traditional recurrent neural networks (RNN) [9] and long short-term memory networks (LSTM), Transformer uses a new architecture that is primarily based on attention mechanisms to capture long-distance dependencies in the sequence.

Transformer structure is mainly composed of encoders and decoders. The main task of encoder is to capture the semantic information of input sequence. In the encoder, each input term is converted into a fixed dimensional vector representation through an embedding layer. These vectors are then processed through multiple layers of self-attention and feedforward neural networks to capture dependencies and semantic information between words. The decoder is the core part of the Transformer model, its main task is to generate new output sequences based on the input sequences that have already been processed. The decoder receives the output sequence from the encoder and then makes multiple rounds of prediction through the self-attention layer and the feedforward neural network layer to generate a new output sequence. Each step of prediction relies on the results of all previous predictions, which allows the decoder to capture more complex linguistic phenomena.

Overall, Transformer enables models to capture global context information in images. Compared with traditional convolutional neural networks (CNNS), it is not limited by the fixed window size, and can better handle input images of different sizes, and it is easier to capture long-distance dependencies while focusing on different parts of the input sequence, so as to realize multi-scale information processing. This is very useful for processing features at different scales in an image and helps to better understand the hierarchy of the image.

IV. Experiment and Result

A. Experimental environment and implementation details

In this paper, the experimental training and testing are carried out under the PyTorch framework. PyTorch framework can run on the GPU, operated by the basic library, the built-in neural network library model training function supports shared memory multi-process concurrent library, which is very helpful for data loading and training. The OS is Ubuntu18.04.5. The CPU is Intel XeonE5-2698 v4. Train the GPU to use four TeslaV100 graphics cards.

After setting up the experimental environment, it is necessary to adjust the hyperparameters before the model can be trained. The batch_size is related to the performance and number of GPU blocks. Set the batch_size to 32. epoch is used to control the iteration and data processing of the training process. The epoch is set to 600 based on the loss function and mAP convergence. init_lr represents the variation amplitude of neural network parameters in the gradient direction. A reasonable setting of the learning rate can be used to converge to the minimum point of the model instead of the local optimal point, which is set here as 0.01. optimizer_type represents the type of optimizer, which is used to update and adjust network parameters during the training of the model and find a set of appropriate parameters so that the model trained by these parameters can perform as well as possible on the test set. In this case, it is set to sgd optimizer. weight_decay adjusts the effect of model complexity on the loss function, set to 0.0005 here. lr_decay_type indicates the learning rate reduction mode. Set this parameter to cos.

B. *Experimental result*

In order to verify the influence of different modules on the algorithm performance of the model, this paper will test the performance of the improved model through ablation experiments. For the data set used in this paper, ablation experiments will be performed on three improved modules in YOLOX-tiny to test the final performance and prove the contribution of each improved module to the overall network model. The improved modules are located in backbone layer, neck layer and YOLO-Head layer of the model respectively. Evaluate and detect performance indicators using mAP.

In this paper, we first evaluate in the test set VisDrone2019-DET-test-dev. We try to analyze the impact of each step by gradually incorporating the impact of each step into the method. Firstly, increasing the number of YOLO-Heads from 3 to 4 can significantly improve the accuracy of the model, improving the mAP by 3.38%, and enhancing the ability to identify small targets; By adding iAFF module to the neck layer and integrating the feature layer output of the backbone network, the mAP is improved by 0.4%, which helps the model better understand the context information of the target and is more friendly to images with different target scales. Finally, add backbone to transformer module, mAP increased by 1.39%. The final network model improved the mAP by 5.17% compared to the baseline model (YOLOX-tiny). Table 1 shows the ablation experiment of the improved network in VisDrone2019.

Table1. Ablation experiment

Baseline	Head	iAFF	Transformer	mAP(%)
YOLOX-tiny	3			31.03
YOLOX-tiny	4			34.41
YOLOX-tiny	4	√		34.81
YOLOX-tiny	4	√	√	36.20

In addition to the ablation experiment, this paper also conducted experiments with other existing network models and compared them, as shown in Table 2. Generally speaking, although the number of parameters of the model proposed in this paper is higher than that of some algorithms, its accuracy has been greatly improved.

Table2. Compare with the existing network model

Method	mAP(%)	Parameters(M)
YOLOv3	27.33	61.57
YOLOv4	29.36	63.98
YOLOv5-s	21.40	7.08
YOLOv7-tiny	27.64	6.01
YOLOX-tiny	31.03	5.03
YOLOX-s	33.16	8.93
YOLOX-m	36.49	25.28
ours	36.20	10.58

V. CONCLUSION

With the popularity of UAV technology, UAV target detection algorithm has a long-term development prospect in many fields such as urban planning, emergency rescue, environmental monitoring, etc. However, there are some difficulties in the images captured by UAV, such as a large number of small targets and occlusion, and UAV target detection has certain requirements on the number of parameters and inference time of the model.

Based on the above situation, this paper proposes an IT-YOLOX target detection algorithm for UAV images, which is improved by taking YOLOX-tiny as the baseline model to apply to UAV aerial images. The first point adds a detection layer to the original three detection layers of YOLOX-tiny, improving the ability to enhance the acquisition of small targets. The second point is to use iAFF module at the neck layer to fuse features with inconsistent semantics and scale to better capture the correlation between input data. The third point is to add the YOLOX-tiny backbone network into Transformer structure at last, so that context-relevant feature representations can be extracted through multi-head attention mechanism. The results show that when VisDrone2019 is used as the dataset and the size of the input image is 640×640, the mAP of the improved model is 36.2%, which is 5.17% higher than that of the original YOLOX-tiny.

REFERENCES

[1] Zou Z ,Shi Z ,Guo Y , et al. Object Detection in 20 Years: A Survey.[J]. CoRR,2019,abs/1905.05055.

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

[3] Redmon J ,Divvala K S ,Girshick B R , et al. You Only Look Once: Unified, Real-Time Object Detection.[J]. CoRR,2015,abs/1506.02640.

[4] Dai Y, Gieseke F, Oehmcke S, et al. Attentional feature fusion[C]//Proceedings of the IEEE/CVF winter conference on applications of computer vision. 2021: 3560-3569.

[5] Ge Z , Liu S , Wang F ,et al.YOLOX: Exceeding YOLO Series in 2021[J]. 2021.DOI:10.48550/arXiv.2107.08430.

[6] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 2117-2125.

[7] Du D, Zhu P, Wen L, et al. VisDrone-DET2019: The vision meets drone object detection in image challenge results[C]//Proceedings of the IEEE/CVF international conference on computer vision workshops. 2019: 0-0.

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[9] Sherstinsky A. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network[J]. Physica D: Nonlinear Phenomena, 2020, 404: 132306.