Transformer-based computer vision technology empowers drones

Mingzheng Lai [1st]
[1st]Nation's Research Laboratory: Alibaba Cloud College of Big Data Application,Zhuhai College of Science and Technology,Zhuhai,China,
Corresponding author: grimm@stu.zcst.edu.cn

PengJie Wang [2nd]
[2nd]Nation's Research Laboratory: Alibaba Cloud College of Big Data Application,Zhuhai College of Science and Technology,Zhuhai,China,
Corresponding author: 1297276320@stu.zcst.cn

YiFan Zeng [3rd]
[3rd]Nation's Research Laboratory: Alibaba Cloud College of Big Data Application,Zhuhai College of Science and Technology,Zhuhai,China,
Corresponding author: yokey@stu.zcst.edu.cn

Wei Lv *[4th]
*[4th]Nation's Research Laboratory: Alibaba Cloud College of Big Data Application,Zhuhai College of Science and Technology,Zhuhai,China,
Corresponding author: luwei@jluzh.edu.cn

Abstract——With the rapid development of drones. The traditional "manual feature extraction + classifier-based" object detection algorithm can no longer meet the accuracy requirements. Aiming at the problem that the reasoning speed is slowed down due to the complex background of UAV aerial images, we propose a target detection model based on a multi-head attention mechanism: Swim-Transformer. This model introduces a multi-head attention mechanism based on the YOLOv5, and introduces a decoupling head method in the head to improve detection accuracy and speed up network convergence. Experiments show that the new target detection framework is superior to the traditional target detection algorithm on the UAV aerial photography data set VisDrone, and increasing mAP by about 0.0067%.

Keywords——Transformer、drone、YOLOv5

I. INTRODUCTION

Real-time recognition and tracking of obstacles in dynamic scenes has always been a difficult problem for unmanned systems. The current mainstream visual target detection methods face problems such as deformation, background noise, and scale transformation. At the same time, the target tracking task also has problems such as geometric deformation, feature occlusion, and distance estimation. Compared with the traditional target detection algorithm, the target detection algorithm based on Transformer[1] has better detection accuracy for abstract targets, and can be applied to the dense target, target overlap, target size, and perspective changes in pictures taken by drones and other equipment. So we will learn from the design ideas of traditional target detection algorithms and give full play to the potential of Transformer on images.

A. Traditional UAV detection methods

The main process of the traditional target detection algorithm is shown in Figure 1, which is mainly divided into the training phase and the testing phase.

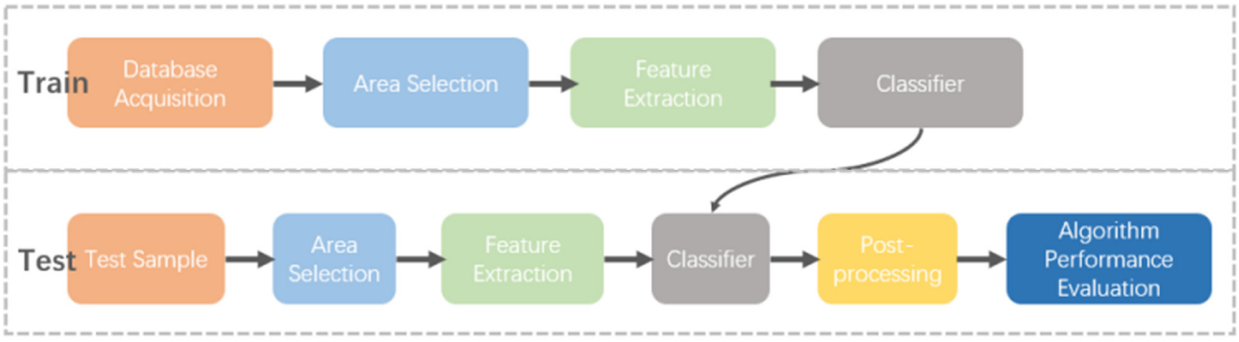

Fig.1 Main process of traditional target detection algorithm

Most of the features used in the early target detection algorithms are calculated based on the pixel values in the image. The well-known V-J detector [2] is a typical algorithm based on manual features. The haar feature descriptor used by VJ is composed of a series of rectangular templates. This method is a grayscale change. . The haar feature description is better at capturing the boundary area, as well as local areas such as straight lines. The corresponding formula (1) is as follows:

2023 4th International Conference on Computer Vision, Image and Deep Learning (CVIDL) | 979-8-3503-2644-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/CVIDL58838.2023.10166286

$$F = \sum_{(x,y)\in R} m_{(x,y)} * I_{(x,y)} \quad (1)$$

Among them, (x, y) represents the index coordinate of the pixel point, R is the area where the template is located on the original image, and is the value of the rectangular template at (x, y), indicating the value at (x, y) of the original image . In order to speed up the feature extraction process, the integral map is introduced, and the calculation method (2) is as follows:

$$C(x,y) = \sum_{i\le x}\sum_{j\le y} I_{(i,j)} \quad (2)$$

Local haar features can be quickly computed by simple addition of integral maps.

The subsequent SIFT features [3] and HOG features [4] are popular due to their better properties.In general, the effectiveness of the target detection algorithm based on manual features has been verified in many experiments, but the feature description adopted by this traditional target detection algorithm did not consider its semantic information at the beginning of the design, which is based on manual features. The natural defects of feature-based target detection algorithms are also the key to the performance constraints of traditional target detection algorithms when faced with complex scenes.

B. *Transformer-based target detection method*

Transformers were originally used in natural language processing because their self-attention components can model information over long distances. More recently, Transformers have been used as CNN-like feature extractors in computer vision. The feasibility of the Transformer architecture in computer vision tasks has been demonstrated by ViT[5] research. Transformer blocks are introduced into CNNs as separate architectures for image classification and object recognition, exploiting long-distance dependencies. The DETR method [6] has been used to model global connections between functions on a serial basis by extracting local features from CNN Transformer encoders. Different from previous work, our proposed method defines an efficient deep learning network structure that directly fuses features. This structure not only organically inherits the structural advantages of CNN and Transformer, but also keeps the model size and computation at a low level.

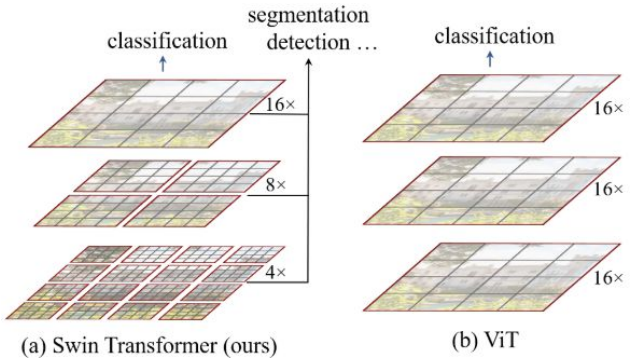

Fig.2 Comparison between Swin-Transformer and ViT

B. *Neck*

In order to make better use of the features extracted by Backbone, Neck was designed. The feature maps extracted by Backbone are reprocessed and used reasonably in different stages. A neck usually consists of several bottom-up paths and several top-down paths. Neck is a key link in the target detection framework. The earliest Neck used up and down sampling blocks, which was characterized by no feature layer aggregation

II. MODEL

Object detectors developed in recent years often insert layers between the backbone and head, and people often call this part the neck of the detector.. These three structures will be described in detail next.

A. *Backbone*

Commonly used backbones include VGG[7], RestNet[8], DenseNet[9], MobileNet[10], EfficientNet[11], Swin-Transformer[12], etc. Among them, Swin-Transformer not only introduces the hierarchical construction method commonly used in CNN to build a hierarchical Transformer, but also introduces the idea of locality to perform self-attention calculations in non-overlapping window areas. It can greatly reduce the computational complexity, as shown in Figure 2. With the deepening of the depth of Swin-Transformer, the hierarchical Transformer can be built by gradually merging image blocks, which can be used as a general visual backbone network for tasks such as image classification, object detection, and language segmentation. So we choose Swin-Transformer as our backbone.

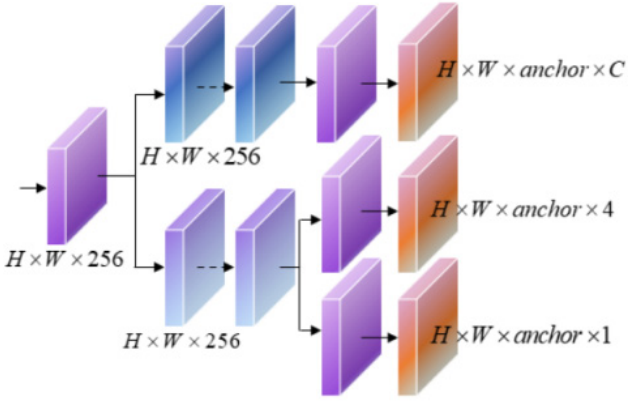

Fig.3 Anchor-based decoupling head method

operation, such as SSD, and the multi-level feature map directly follows the head. As shown in Figure 3.

C. *Head*

Backbone acts as a network classifier, but it cannot complete the localization task, and Head is responsible for detecting the location and category of the target through the feature map extracted by Backbone. Head is generally divided into two types: One-Stage detector and Two-Stage detector.

Two-stage detectors have been the primary method in the field of object detection. The most representative of these is the RCNN series .Compared with Two-Stage detectors, One-Stage detectors must predict both box and object categories. The speed advantage of One-Stage detector is obvious, but the accuracy is lower. However, in view of the real-time nature of drones, we selected the most representative yolo series in One-Stage as our basic experimental algorithm.

In yolo x, the method of decoupling head [13] is used to speed up network convergence and improve accuracy. Therefore, we also want to introduce the method of decoupling the head in the detection head of YOLO v5, so as to improve the detection accuracy and speed up the network convergence, but here is slightly different from the detection method used by the YOLO x decoupling head. The decoupling head introduced in YOLO v5 is still based on the anchor detection method.

III. Experiment

A. Data

The data set used in the experiment is VisDrone. In the official data, the training set is 6471, and the verification set is 548. There are 11 classes in total, which are: 'pedestrian', 'people', 'bicycle' ', 'car', 'van','truck', 'tricycle', 'awning-tricycle', 'bus', 'motor', 'others', where others are non-valid target areas, as shown in Figure 4.

In order to increase the diversity of samples and enhance the robustness of the network, we use Mosaic data enhancement, as shown in Figure 5. This data enhancement method is to splice the four photos by random arrangement, random cropping, and random arrangement. This can enrich the background and small targets of detected objects, and calculate the data of four pictures at a time when calculating batch normalization, thereby reducing the amount of calculation.

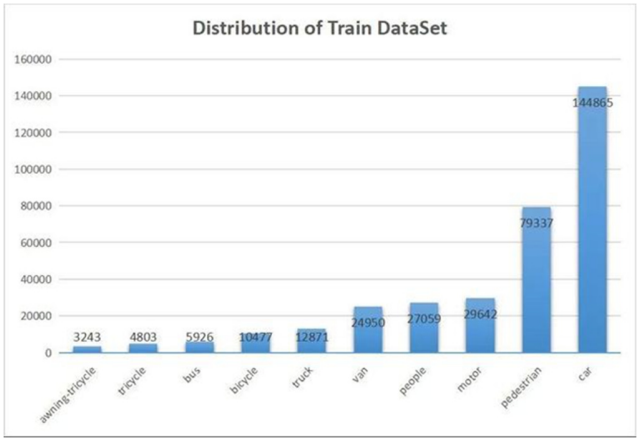

Fig.4 Data Distribution Statistics

Fig.5 Mosaic data enhancement results of a batch after target detection (yolov5m6_transformer)

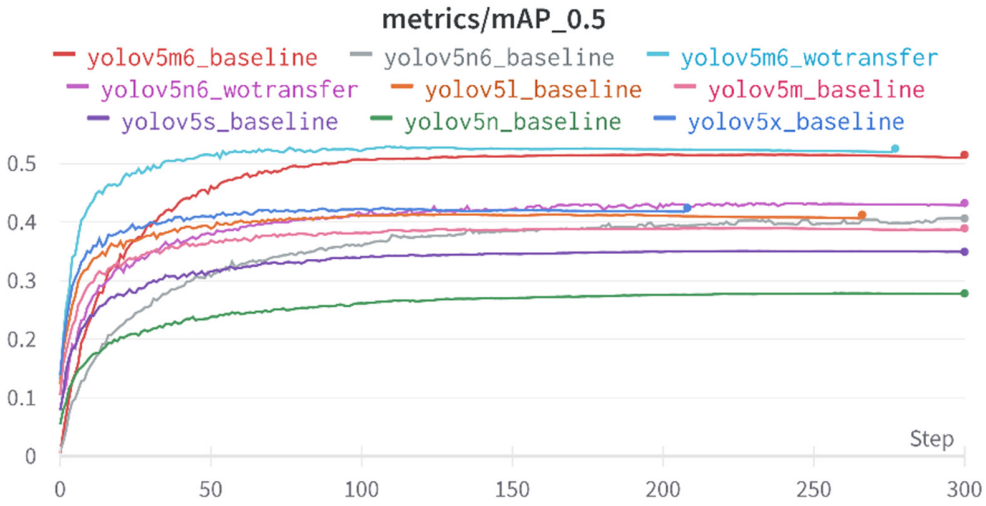

Fig.6 Comparison of mAP of each model

B. *Experimental results*

We compared 9 models, namely yolov5x, yolov5n, yolov5m, yolo5s, yolov5l, yolov5n6, yolov5n6+swin-transformer, yolov5m6, yolov5m6+swin-transformer.

From Figure 6, it can be seen that the mAP of the yolov5m6 model and yolov5n6 model after adding the swin-transformer are improved.

Figure 7 is the result of target detection, it can be seen that the accuracy rate is very impressive. Its accuracy rate can reach up to 91.1%.

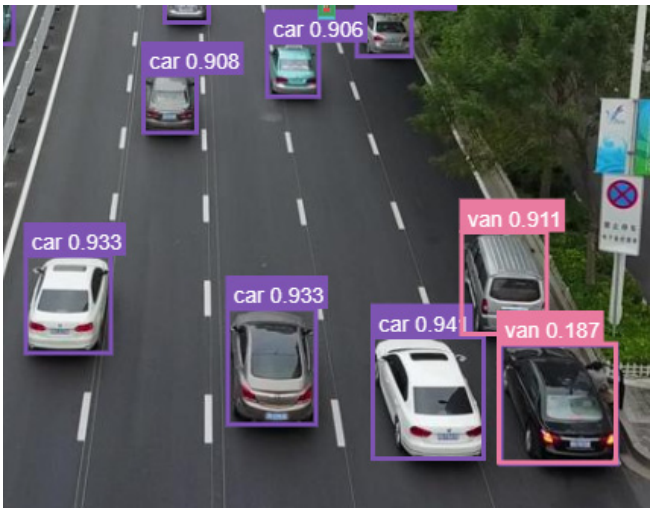

Fig.7 Object detection result

IV. Conclusion

The method proposed in this paper is based on the one-stage classic algorithm yolov5 combined with the swin-transformer method, which improves the accuracy of the model on the visdrone data set. And reduce the amount of calculation. It plays an important role in the detection of UAV targets with small image size, single background, large target size and low density.

But the accuracy can continue to be improved through subsequent experiments, such as combining Transformer and CNN. Or introduce the CBAM [14] attention mechanism template. CBAM is a simple and effective attention module for feed-forward convolutional neural networks. CBAM is a lightweight general-purpose module, so it can be seamlessly integrated into any CNN architecture by ignoring the overhead of this module, and can be trained end-to-end with the underlying CNN. This enables better feature extraction.

Acknowledgement

This research was partially supported by Zhuhai College of Science and Technology grant: S202213684041

References

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

[2] Viola P, Jones M. Rapid object detection using a boosted cascade of simple features[C]//Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001. Ieee, 2001, 1: I-I.

[3] Beck K D, Luine V N. Food deprivation modulates chronic stress effects on object recognition in male rats: role of monoamines and amino acids[J]. Brain research, 1999, 830(1): 56-71.

[4] Dalal N, Triggs B. Histograms of oriented gradients for human detection[C]//2005 IEEE computer society conference on computer vision and pattern recognition (CVPR'05). Ieee, 2005, 1: 886-893.

[5] Y. Wang, Y. Yang, and X. Zhao, "Object detection using clustering algorithm adaptive searching regions in aerial images," in Proc. Eur. Conf. Comput. Vis., Aug. 2020, pp. 651-664.

[6] Z. Wu, K. Suresh, P. Narayanan, H. Xu, H. Kwon, and Z. Wang, "Delving into robust object detection from unmanned aerial vehicles: A deep nui- sance disentanglement approach," in Proc. IEEE/CVF Int. Conf. Comput. Vis., 2019, pp. 1201-1210.

[7] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition[J]. arXiv preprint arXiv:1409.1556, 2014.

[8] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[9] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2017: 4700-4708.

[10] Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.

[11] Tan M, Le Q. Efficientnet: Rethinking model scaling for convolutional neural networks[C]//International conference on machine learning. PMLR, 2019: 6105-6114.

[12] Liu Z, Lin Y, Cao Y, et al. Swin transformer: Hierarchical vision transformer using shifted windows[C]//Proceedings of the IEEE/CVF international conference on computer vision. 2021: 10012-10022.

[13] Ge Z, Liu S, Wang F, et al. Yolox: Exceeding yolo series in 2021[J]. arXiv preprint arXiv:2107.08430, 2021.

[14] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 3-19.