2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)

Target Detection Algorithm for Drone Aerial Images based on Deep Learning

1st Tao Liu
College of Computer Science and Technology
Heilongjiang Institute of Technology
Harbin, Heilongjiang, China
290785353@qq.com

2nd Bohan Zhang
[1]*School of Information Engineering*
China University of Geosciences
Wuhan, Hubei, China
[2]*State Key Laboratory of Information Engineering in Surveying Mapping and Remote Sensing*
Wuhan University
Wuhan, Hubei, China
332918692@qq.com

2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC) | 979-8-3503-7544-2/24/$31.00 ©2024 IEEE | DOI: 10.1109/ICDSCNC62492.2024.10939462

***Abstract*—In recent years, the rapid development of drones has brought tremendous changes to many fields. From refined management in agriculture and forestry to aerial surveying in urban planning, the popularity of UAVs provides efficient and accurate data support for various industries, promoting social intelligence and technological progress. Meanwhile, deep learning methods represented by computer vision provide powerful tools for research such as image processing and object recognition. The research results of HE-YOLOX-ASFF (Highly Efficient You Only Look Once eXtend with Adaptive Spatial Feature Fusion) algorithm for target detection in UAV aerial images in this article can help improve the recognition accuracy and efficiency of drone remote sensing images, and apply it to fields such as smart agriculture, urban management, and environmental monitoring. At present, detection technologies for large-scale targets have achieved good detection results in clear backgrounds. However, drone aerial images often have complex backgrounds, small target volumes, limited available features, and high target density, making them prone to occlusion. Due to the above reasons, existing target detection algorithms have low detection accuracy for weak and small targets in aerial images under complex backgrounds, and are prone to missed and false detections. Therefore, the detection method of weak aerial targets in complex backgrounds remains a challenging issue. The experimental results show that in identifying pedestrians and bicycles, the parameter values are relatively high, at 33.3 and 65.1, respectively, while the FPS (frames per second) is relatively low, at 63 and 38, respectively. This indicates that the algorithm requires more parameters to process these targets and may affect processing speed.**

Keywords—drone aerial images, object detection, yolo algorithm, computer vision technology.

I. Introduction

Object detection in drone aerial images is an important research direction in computer vision. In the past decade, target detection methods in drone aerial images have been increasingly applied in fields such as traffic control, fire warning, high-altitude operations, and military operations. However, the large number and dense distribution of small objects, large scale changes, and complex and variable environments in drone aerial images make it difficult for conventional perspective-based target detection methods to be applicable to target detection problems in bird's-eye views.

This article introduces an efficient small object detection algorithm HE-YOLOX, which is improved based on YOLOX-S (You Only Look Once eXtend - Small). The backbone of this algorithm adopts CSP (Cross Stage Partial) Darknet, which consists of multiple CSP Layer structures and utilizes residual convolution to achieve feature extraction. CSP Layer is divided into main branch and residual branch, and feature fusion is achieved through residual connections. In terms of feature fusion, the ASFF module is applied to replace the traditional PANet (Path Aggregation Network) model, achieving adaptive fusion of multi-scale features. The experimental results show that the algorithm performs outstandingly in the detection of large vehicles, reaching the best level in the detection of vehicles and buses, but there is still room for improvement in the recognition of small targets. On the whole, the method proposed in this article has good application prospects in drone aerial images.

II. Related Works

For drone aerial image target detection, there have been experts specializing in this for a long time. Khan S modeled the target using an efficient Convolutional Neural Network (CNN) and tracked the target in real-time using a PID (Proportional-Integral-Differential) control algorithm to fly around the target for a few seconds. Tests have proved that the system can adapt to the dynamics of various complex situations [1]. Huang F adopted K-Means and TensorRT inference methods for YOLOv4, which improved the NVIDIA JetsonTX2 in terms of computational accuracy and computational speed. The overall confidence level was 89.6% and 3.8% in the test including static test [2]. Chen J combined the Double Route Attention (BRA) mechanism with YOLOv7 (You Only Look Once Version 7) for weak target detection in drone aerial images [3]. Guo L studied a super-resolution reconstruction algorithm based on recurrent convolutional neural networks, which achieved multi-level super-resolution reconstruction by adjusting the number of cycles. He conducted simulation experiments on the system. Experimental results have shown that this method is reliable [4]. Tian G calculated the feature vectors obtained from the denoised sparse autoencoder and then filtered them [5].

Barisic A proposed an approach that utilizes texture randomness to emphasize shape-based target representations and built a set of 3D model libraries containing multiple types of features based on Blender. The average accuracy of his proposed algorithm on real images is improved by 17, 3.7 percentage points on average [6]. Dianqing Y enhanced the image using a combination of grayscale transformation and Gaussian filtering to improve the quality of the image and then applied the feature pyramid network structure [7]. Ren K researched a novel detection network based on Region Super Resolution Generative Adversarial Network (RSRGAN). The experimental results indicated that the algorithm can effectively detect small-sized infrared targets, and its

performance is significantly improved compared to existing algorithms [8].

Wang C proposed a method based on U-shaped production adversarial networks, which integrated visible light and infrared images. To enhance the recognition ability of the adversarial network for pedestrians, he proposed a convolutional block attention model. The experimental results showed that the transfer learning algorithm based on fused images had good recognition performance [9]. Momin M A utilized YOLOv4 (You Only Look Once Version 4) Tiny to design a lightweight convolutional neural network model for detecting vehicles in the dataset. He found through the analysis of experimental data that this method has a higher average accuracy than previous methods [10]. Qi J combined drone remote sensing images with landslide characteristics to achieve classification of ground object detection, and provided evaluation criteria for ground object detection and several typical ground object detectors. He proved the feasibility of the designed system through statistical analysis of experimental data [11]. In order to prevent feature information loss caused by small target features during the sampling stage, Li Y added Feature Alignment Module (FAM) and Feature Selection Module (FSM) to the Feature Pyramid Network (FPN). He used a transformer encoder block to form a Transformer Prediction Head (TPH) to replace the prediction head of the original model. The experimental results showed that compared with the original YOLOv6 (You Only Look Once Version 6) network, when the LOU (Intersection over Union) was 0.5, the average accuracy of the improved network on the validation set was 59.73%, which was 6.02% higher than before [12].

Wang Q designed and implemented the ODTDS (online drone-based target detection system), which can perform online data processing and autonomous navigation under limited resources. The effectiveness and reliability of the system were demonstrated through outdoor high-altitude experiments [13]. Delplanque A compared three multi-class CNN algorithms for detecting African mammalian species based on high-resolution aerial images: Faster-RCNN, Libra-RCNN and RetinaNet [14]. Li R proposed an object detection technique for Unmanned Surface Vehicle (USV) based on the EfficientDet algorithm. Compared to Faster-RCNN and YOLO V3, this method improved the accuracy of ship target detection to 87.5% [15]. The main bottleneck faced by current drone image target recognition technology is insufficient response to small target detection and occlusion problems. In complex environments, identifying small targets and targets obscured by other objects remains a challenge, which can easily lead to missed or false detections, limiting the accuracy and reliability of the system.

III. Methods

A. Algorithm Structure

The efficient small object detection algorithm HE-YOLOX introduced in this article is improved based on YOLOX-S. The backbone network still uses CSPDarknet. The entire backbone is composed of residual convolution, which is mainly composed of multiple CSPLayer structures. CSPLayer is divided into two parts. The main branch stacks N residual blocks based on the depth set by the network. The residual branch is directly concatenated with the main branch after only a small amount of processing. After passing through convolutional layers and BN (Batch Normalization) layers, the input is activated into the next layer of CSPLayer structure through SiLU activation function. Among them, the residual block is also divided into two parts, the backbone part is 1×1 convolution and 3×3 convolution, and the residual edge uses skip connections to directly combine the input and output of the backbone. The advantage of residual networks is that they can increase network depth while avoiding network degradation and improving detection accuracy. SiLU is a combination of sigmoid activation function and ReLU activation function [16]. The calculation formula is:

$$SiLU(x) = x.\frac{1}{1+e^{-x}} \tag{1}$$

B. Multi-Scale Feature Extraction

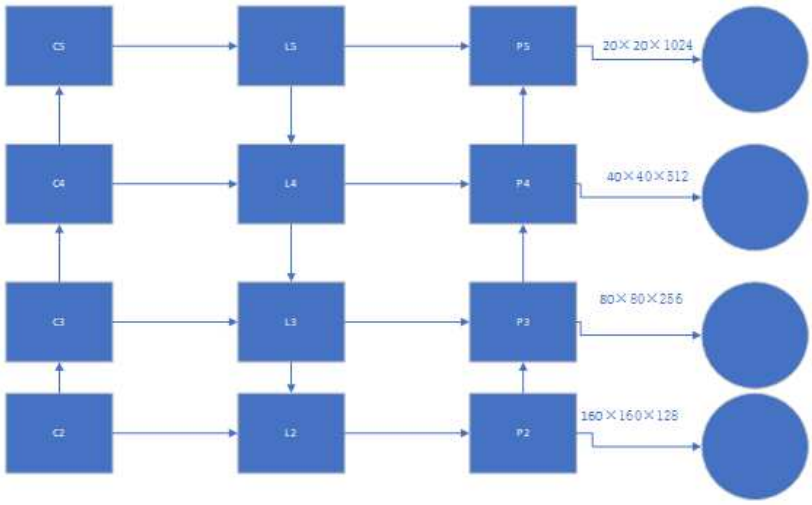

Fig. 1. Adding small target layer fusion process

The detection of existing single-stage small object detection algorithms is mostly achieved by downsampling the input image (640 pixels×640 pixels) at 8 times, 16 times, and 32 times, as well as feature pyramids and path aggregation networks to obtain the detection head. However, this is accompanied by the acquisition of shallow to deep feature maps. Among them, multiple convolutional pooling operations lose some positional information. Due to the large

number of small targets and small inter class variance in the aerial target dataset, the existing network structure cannot play its maximum role. After multiple downsampling, the original feature maps of three scales are difficult to detect such extremely small targets. In order to improve the feature information extraction of small targets, a low level and scale feature information is added to input into PANet. The feature map C2 obtained from the early convolution of the backbone network is input into the neck network and fused with {C3, C4, C5} to obtain a new feature map. For detecting small targets, the layer fusion process is added as shown in Figure 1. Among them, the feature layers input to the detection head by YOLOv7 are {P2, P3, P4, P5}, and their sizes are shown in Figure 1 [17-18].

C. ASFF Module

The problem of small target object recognition involves different semantic information contained in feature layers at different scales. However, due to the shallow nature of drone aerial images, significant losses often occur during feature fusion. In response to this issue, this article introduces a new ASFF-based method, which uses the ASFF model to replace the original PANet model and performs feature fusion on it. This method plays a dominant role in the fusion process by adaptively learning the weights of different feature layers. This method involves two aspects: feature adjustment, which maps features from other scales to the corresponding scale to ensure that the scale remains unchanged during fusion. In adaptive fusion, the network is trained to obtain the important weight parameters α, β and γ for three different feature layers, and then multiplied with the feature maps of each scale respectively. Taking ASFF-1 as an example, X_1, X_2 and X_3 are the feature maps from the three scales of YOLOX path aggregation network output respectively, and $X^{2\to1}$ and $X^{3\to1}$ are the feature layers of the same size as X_1 that are generated from X_2 and X_3 after feature scaling respectively. Then, $X^{1\to1}$, $X^{2\to1}$, $X^{3\to1}$ are multiplied with the corresponding weight parameters α, β, and γ, and then summed to output Y^1, which is fused by Formula (2).

$$Y_{ij}^1 = \alpha_{ij}^1 \cdot X_{ij}^{1\to1} + \beta_{ij}^1 \cdot X_{ij}^{2\to I} + \gamma_{ij}^1 \cdot X_{ij}^{3\to I} \tag{2}$$

In the formula, Y_{ij}^1 represents the new feature map $\alpha_{ij}^1 + \beta_{ij}^1 + \gamma_{ij}^1 = 1$ obtained through ASFF-1, and $X_{ij}^{n\to I}$ represents the feature vector at (x, y) on the nth layer feature map before fusion of the lth layer feature map obtained after the above correlation operation, as shown in Figure 2.

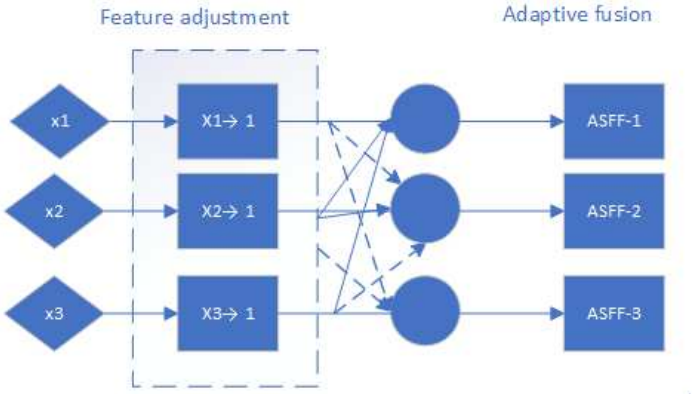

Fig. 2. ASFF module

IV. Results and Discussion

A. Datasets

The dataset used in this article is the VisDrone2019 dataset released by a certain laboratory, consisting of 10209 high-altitude drone images, including various climate change shooting angles and brightness changes. Among them, there are 6471 training sets, 548 validation sets, and 3190 testing sets. This dataset consists of 13 categories, namely pedestrians, people, bicycles, cars, vans, trucks, tricycles, awning-tricycles, buses, trees, road signs, buildings, and motorcycles.

B. Comparison of Experimental Results

From Figure 3, it can be seen that this set of data shows the Average Precision (AP) values of object detection based on the YOLO-ASFF algorithm in different categories. The YOLO-ASFF algorithm has the most outstanding performance in automotive object detection, with an AP value of around 81.2, followed by buses and trucks, with AP values of around 66.4 and 55.6, respectively. In the categories of pedestrians, trucks, and motorcycles, the AP values are around 42.6, 47.0, and 45.5, respectively, indicating good performance. Relatively low AP values appear in the categories of people, bicycles, tricycles, and covered tricycles, at around 30.6, 19.4, 29.8, and 27.1, respectively. Overall, the YOLO-ASFF algorithm has high accuracy in detecting categories of cars and large vehicles, while its accuracy is relatively low in detecting categories of bicycles and some tricycles.

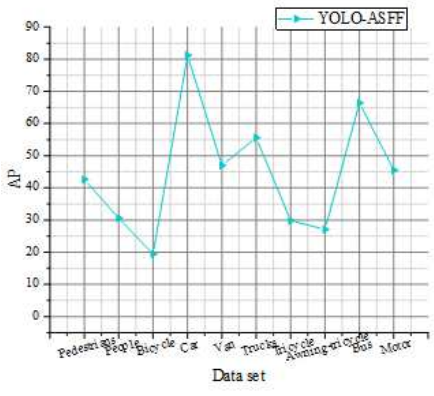

Fig. 3. Comparative experiment of object detection algorithms

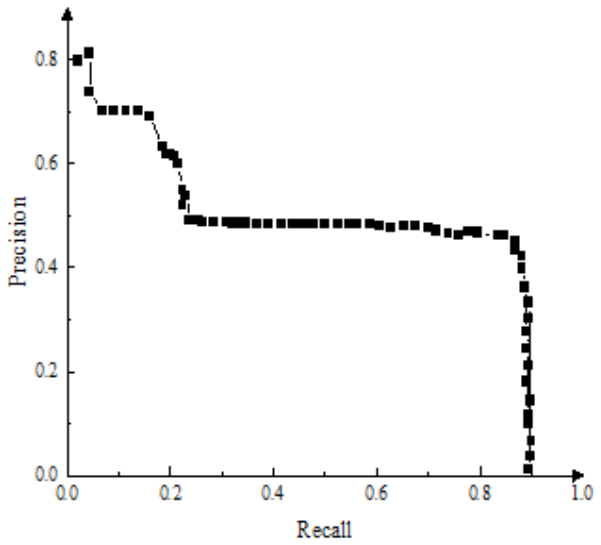

Fig. 4. P-R curve of object detection using YOLO-ASFF algorithm on drone image dataset

Figure 4 shows the P-R curve of the target detection algorithm for drone aerial images based on the YOLO-ASFF algorithm. As the recall rate increases, the precision overall shows a downward trend, but there are fluctuations. For example, when the recall rate increases from around 0.02 to around 0.07, the accuracy decreases less and remains around 0.70; as the recall rate further increases to around 0.22, the accuracy gradually decreases to around 0.52. It is worth noting that some improvements in recall rates don't significantly reduce accuracy, and even when the recall rate reaches around 0.68, accuracy remains at around 0.48. These data reflect the performance changes of YOLO-ASFF algorithm in target detection of drone aerial images, indicating that the algorithm can maintain relatively stable accuracy at high recall rates, but as the recall rate continues to increase, the accuracy significantly decreases.

TABLE I. PERFORMANCE COMPARISON OF ALGORITHMS ON DIFFERENT DATASETS

Data sets	Params	FPS
Pedestrian	33.3	63
People	9.5	371
Bicycle	65.1	38
Car	7.3	96
Van	9.7	60
Trucks	9.3	50
Tricycle	6.1	113
Tricycle with canopy	8.8	64
Bus	4.9	78

According to the data in Table 1, it can be observed that the algorithm performs differently on different datasets. For example, in identifying pedestrians and bicycles, the parameter values are higher at 33.3 and 65.1, respectively, while FPS (frames per second) is lower at 63 and 38, respectively. This indicates that the algorithm requires more parameters to process these targets and may affect processing speed. On the contrary, in terms of identifying people and buses, the parameter values are lower at 9.5 and 4.9, respectively, while the FPS is higher at 371 and 78, respectively, indicating that the algorithm may be more efficient in identifying these targets.

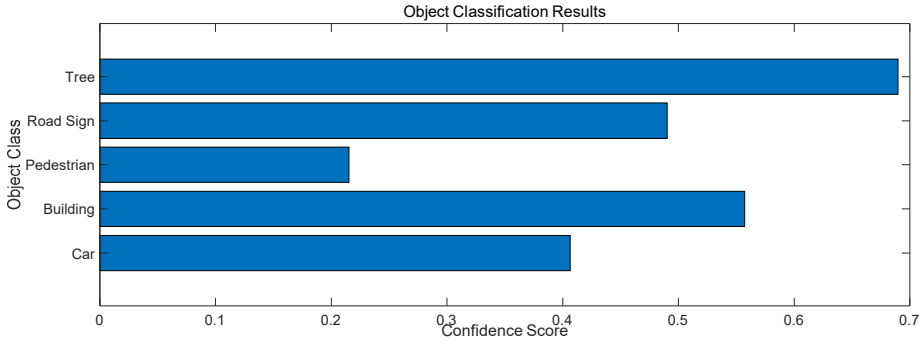

Fig. 5. Classification scores for different targets

Figure 5 shows the results obtained from target detection and classification of drone aerial images based on the YOLO-ASFF algorithm, showing the classification scores for different target categories, including cars, buildings, pedestrians, road signs, and trees.

C. Discussions

The HE-YOLOX-ASFF algorithm developed in this study shows significant improvement in the accuracy of large vehicle detection in UAV aerial images, especially when dealing with complex backgrounds and high-density targets. This is attributed to its enhanced feature extraction capability and adaptive spatial feature fusion. However, there is still room for improvement in recognising smaller and irregular targets such as pedestrians and bicycles. Although the HE-YOLOX-ASFF algorithm performs well in large object detection, it has lower AP values for small object detection with higher parameter values and lower frame rates, suggesting that there is a trade-off between accuracy and processing speed. These results are important for applications such as smart agriculture, urban management and environmental monitoring, which can provide more reliable data support in these fields. Future research should focus on optimising the algorithms to improve the detection accuracy of small objects, possibly involving the development of more advanced feature extraction and fusion methods combined with other advanced deep learning techniques. Comparisons with existing methods show that HE-YOLOX-ASFF has an advantage in large object detection, but combining successful strategies such as attentional mechanisms or transformer-based models may further enhance its capabilities. In conclusion, the HE-YOLOX-ASFF algorithm has great potential for UAV aerial image analysis, but further improvements are needed to achieve comprehensive target detection performance enhancement.

V. Conclusions

With the rapid development of computer vision, object detection methods based on deep learning have been increasingly applied due to their superior performance. This study proposes an efficient small target detection method based on the YOLO-ASFF algorithm. The experimental results show that the algorithm performs outstandingly in the detection of large vehicles, reaching the optimal level in the detection of cars and buses. However, there is still room for improvement in the recognition of small targets, especially unconventional targets such as pedestrians and bicycles. Future research directions include but are not limited to further optimizing algorithms to improve the accuracy and robustness of small object detection, exploring more effective feature extraction and fusion methods, and combining more advanced technologies such as deep learning and computer vision to promote the application of drone aerial images in fields such as smart agriculture, urban management, and environmental monitoring. At the same time, strengthening the adaptability research of algorithms under different scenarios and environmental conditions, improving the stability and reliability of algorithms in practical applications, can be an important direction for future research.

Author Contributions

The corresponding author is Bohan Zhang.

References

[1] Khan S, Tufail M, Khan M T, et al. A novel framework for multiple ground target detection, recognition and inspection in precision agriculture applications using a UAV[J]. Unmanned Systems, 2022, 10(01): 45-56.

[2] Huang F, Chen S, Wang Q, et al. Using deep learning in an embedded system for real-time target detection based on images from an unmanned aerial vehicle: Vehicle detection as a case study[J]. International Journal of Digital Earth, 2023, 16(1): 910-936.

[3] Chen J, Wen R, Ma L. Small object detection model for UAV aerial image based on YOLOv7[J]. Signal, Image and Video Processing, 2024, 18(3): 2695-2707.

[4] Guo L, Yang R, Zhong Z, et al. Target recognition method of small UAV remote sensing image based on fuzzy clustering[J]. Neural Computing and Applications, 2022, 34(15): 12299-12315.

[5] Tian G, Liu J, Zhao H, et al. Small object detection via dual inspection mechanism for UAV visual images[J]. Applied Intelligence, 2022, 52(4): 4244-4257.

[6] Sunitha, D., Balmuri, K. R., de Prado, R. P., Divakarachari, P. B., Vijayarangan, R., & Hemalatha, K. L. (2023). Congestion centric multi-objective reptile search algorithm-based clustering and routing in cognitive radio sensor network. Transactions on Emerging Telecommunications Technologies, 34(11), e4629.

[7] Dianqing Y, Yanping M. Remote sensing landslide target detection method based on improved Faster R-CNN[J]. Journal of Applied Remote Sensing, 2022, 16(4): 044521-044521.

[8] Ren K, Gao Y, Wan M, et al. Infrared small target detection via region super resolution generative adversarial network[J]. Applied Intelligence, 2022, 52(10): 11725-11737.

[9] Wang C, Luo D, Liu Y, et al. Near-surface pedestrian detection method based on deep learning for UAVs in low illumination environments[J]. Optical Engineering, 2022, 61(2): 023103-023103.

[10] Momin M A, Junos M H, Mohd Khairuddin A S, et al. Lightweight CNN model: automated vehicle detection in aerial images[J]. Signal, Image and Video Processing, 2023, 17(4): 1209-1217.

[11] Qi J, Chen H, Chen F. Extraction of landslide features in UAV remote sensing images based on machine vision and image enhancement technology[J]. Neural Computing and Applications, 2022, 34(15): 12283-12297.

[12] Jangam, N. R., Guthikinda, L., & Ramesh, G. P. (2022). Design and analysis of new ultra low power CMOS Based flip-flop approaches. In Distributed Computing and Optimization Techniques: Select Proceedings of ICDCOT 2021 (pp. 295-302). Singapore: Springer Nature Singapore.

[13] Wang Q, Gu J, Huang H, et al. A resource-efficient online target detection system with autonomous drone-assisted IoT[J]. IEEE Internet of Things Journal, 2022, 9(15): 13755-13766.

[14] Delplanque A, Foucher S, Lejeune P, et al. Multispecies detection and identification of African mammals in aerial imagery using convolutional neural networks[J]. Remote Sensing in Ecology and Conservation, 2022, 8(2): 166-179.

[15] Li R, Wu J, Cao L. Ship target detection of unmanned surface vehicle base on efficientdet[J]. Systems Science & Control Engineering, 2022, 10(1): 264-271.

[16] Diwan T, Anirudh G, Tembhurne J V. Object detection using YOLO: Challenges, architectural successors, datasets and applications[J]. multimedia Tools and Applications, 2023, 82(6): 9243-9275.

[17] Nazil Perveen, Debaditya Roy and C Krishna Mohan, "Facial Expression Recognition in Videos using Dynamic Kernels," IEEE Transactions on Image Processing, vol. 29, pp. 8316-8325, 10.1109/TIP.2020.3011846, 2020.

[18] Hurtik P, Molek V, Hula J, et al. Poly-YOLO: higher speed, more precise detection and instance segmentation for YOLOv3[J]. Neural Computing and Applications, 2022, 34(10): 8275-8290.