

YOLO Series Target Detection Technology and Application

Yuan Zhang *

United International College, Beijing Normal University - Hong Kong Baptist University, Zhuhai, China

* Corresponding author email: q030026211@mail.uic.edu.cn

Abstract. Recently, YOLO is the most popular algorithm in machine learning. The algorithm has developed rapidly, and there are several versions at present. Each version of the framework is different, and they also have their own application areas. And maybe in one area, not only one version can be used. This paper summarizes the process of target detection, the structures of YOLO network. In addition, this work also analyzed the development, advantages and disadvantages of YOLO target detection. Finally, the application of YOLO in automatic driving and UAV detection are discussed. YOLO may develop faster in the future. YOLO model is a variable model, which has unique functions when detecting different things under different circumstances. At the end of the paper, the thesis is summarized, and the related research has certain reference value.

Keywords: YOLO; Target Detection; Automatic Driving; UAV Target Detection.

1. Introduction

Object detection is a research hotspot in computer vision, and it meets the requirements in many fields, such as monitoring security, driving, traffic monitoring and machine vision. The target detection is generally to detect some predefined target instances. YOLO target detection method is used in many target detection fields. YOLO is the most popular and representative target detection method. YOLO (You Only Look Once) summed up marker detection as regression problems to achieve end-to-end training and detection [1]. Because of its good speed precision balance, YOLO has been in the leading position in the field of marker detection in the last few years, and has been successfully studied, improved and applied to many different fields. YOLO directly looks for the position of possible objects from the picture, so it is also called one stage model.

Although many improved extension algorithms based on YOLO have been proposed, the analysis and comparative review of relevant methods of killing you is not comprehensive enough. For this reason, this paper mainly describes YOLO and its extension algorithms, analyzes the differences, advantages and disadvantages between YOLO and its extension algorithms, and outlines the application of YOLO and its extension algorithms in the field of automatic driving and UAV target recognition.

2. YOLO Target Detection and its Series

YOLO is an advanced single stage marker detection framework. In the detection process of YOLOv1, first adjust the output diagram to 448×448 , send to CNN to extract features, and then process the prediction results of network to achieve end-to-end detection of landmarks [2]. Subsequent YOLO series inherit this basic idea. In addition, YOLOv5, developed based on PyTorch, has just matured. Its feature is that the code is easier to understand and expand, and it has many applications in many fields. In the development process of YOLO algorithm, five versions have their own fields, YOLOv1, YOLOv2, YOLOv3 which is a milestone, YOLOv4 which have better performance on the COCO (Common Objects in Context) dataset and YOLOv5. In addition to this five, YOLOR (You Only Learn One Representation) which combines traditional compression sensing, and YOLOX which does not rely on the anchor frame.

2.1 The Analysis of YOLOv1

YOLOv1 abandons the traditional sliding window technique, and its CNN separates the input photograph into $S \times S$ grid, and each cell is responsible for detecting those targets whose center points fall within the grid. Each cell predicts B bounding box and the confidence of the bounding box. The confidence includes the small probability that the bounding box contains a marker and the accuracy of the bounding box. Each bounding box predicts five elements: (x, y, w, h, c) , representing the position, magnitude and confidence of the bounding box. Each cell predicts $(B \times 5 + C)$ values, where C is the categories number. The grid prediction is then performed using Non-Maximum Suppression (NMS) algorithm [3].

YOLOv1 makes main network like Google Net, with 24 convolution layers and 2 full connection layers. It is pre-trained on ImageNet, then migrated to the detection task and validated on the VOC (Visual Object Classes) dataset.

YOLOv1 divides the input graph into 7×7 . Each mesh predicts two bounding boxes, so there are $7 \times 7 \times 2$ bounding boxes. Each mesh predicts two bounding boxes. Because of this, a maximum of 49 targets can be identified. YOLOv1 takes a long time to predict, has few targets to identify and is not conducive to identifying intensive beacons.

2.2 The Analysis of YOLOv2

YOLOv2 builds a new main network, Darknet-19, based on YOLOv1 by borrowing the VGG grid. Compared with the backbone network of Google Net used in YOLOv1, VGG network has only a several small filter convolution layers, and its structure is very simple. YOLOv1 uses fully connected layers to predict the bounding box, which has lost more spatial information and leads to inaccurate positioning [4]. Therefore, YOLOv2 introduces the anchor frame instead of the fully connected layer of v1 to predict boundary frame. At the same time, YOLOv2 trims the input size to 416×416 , resulting in a 13×13 feature map with an odd number of dimensions, and a feature map with exactly one center. The resulting feature map has exactly one center. By using this central location to predict the target where the central point falls, the detection of this type of target will be easier.

The dashed rectangular box is a priori, and the box is a prediction bounding box obtained by the offset of the prediction. (cx, cy) denotes the upper-left corner of the cell. (pw, ph) is the width and height of the prior box. (tx, ty) and (tw, th) are the center offset and the width-to-height ratio of the bounding box of the grid; the truth coordinates in the feature map are (Gx, Gy, Gw, Gh) . (bx, by, bw, bh) is the final predicted bounding box. The conversion process from the predefined bounding box to the predicted bounding box is shown in the right-hand side, where σ function is a Sigmoid function, which aims to scale the prediction offset to 0-1 and accelerate the convergence of the grid.

YOLOv2 pioneered the use of combined classification and detection training method to extend the target detection samples. Obviously improve the accuracy of prediction while maintaining the advantage of fast detection.

2.3 The Analysis of YOLOv3

The basic network of YOLOv3 is Darknet-53, which extracts the features of 8, 16 and 32 times down sampling. 1×1 convolution and 3×3 convolution are widely used in the network, of which 1×1 is mainly used for channel expansion and reduction. It uses the residual structure of ResNet for reference. Its essence is the idea of matrix decomposition, which is used to reduce the number of parameters. While deepening the network structure, the problem of network convergence caused by network gradient explosion is prevented. In the process of forward propagation, the pooling layer and the full junction layer are removed and the tensor scale is changed by changing the step size of the convolution kernel. Similar to v2, Darknet-53 scales down the output features to 1/32 of the input. Therefore, the resolution of the output image is usually required to be a multiple of 32. At the same time, YOLOv3 uses tensor splicing to expand the dimensions of tensors to extract more information. The specific operation is to conduct upper sampling and splicing of the intermediate layer of Darknet-53 and a later layer.

In a word, the prediction box of YOLOv3 is more than 10 times that of YOLOv2, and they are advanced at different scales. The overall detection accuracy and the detection accuracy of objects have been greatly improved, so it becomes one of the key algorithms in single-stage detection.

2.4 The Analysis of YOLOv4

YOLOv4 summarized the various improvement methods after v3, divided into free packs and special packs. The former refers to the modules that have no influence on reasoning speed by improving training and the latter refers to the modules that have more influence on reasoning time and have more performance returns [5]. The former can maintain a higher reasoning speed and have a higher accuracy because of the local cross stage partial (CSP) structure in the main network. At the same time, YOLOv4 is more suitable for training on a single video card. The latter uses CSPResNeXt-50 for detection and its accuracy is lower than that of the former. At the same time, free bags and Mish improve the accuracy of detection. Therefore, YOLOv4 chooses CSPDarknet-53 as the main network.

For the basic network, the overall architecture of YOLOv4 is the same as that of YOLOv3, but each structure has been improved. CSPDarknet-53 network is used to replace Darknet-53. YOLOv4 deletes the last pooling layer, full connection layer and Softmax layer. Its main network has five CSP modules.

2.5 The Analysis of YOLOv5 and YOLOX's

The basic network structure of YOLOv5 is similar to that of YOLOv4. The biggest difference is that it scales according to different sizes. The model of YOLOv5-N/S/M/L/X 5 is built from small to large according to the model. Based on YOLOv3 and YOLOv5, YOLOX used CSPNet, SiLU activation function and PANet, and designed four models of YOLOX-S/M/L/X following the scaling rules; then the models were further scaled down to build YOLOXTiny and YOLOX- Nano for mobile edge devices.

2.6 The Compare of Different YOLO

In order to further compare the differences between different versions of YOLO, we summarized the structure, performance and subsequent improvements of YOLO V1-V4. The specific results are shown in the table 1.

Table 1. Comparison between YOLO V1-V4

Method	Features	Advantages	Disadvantages	Improved afterwards
YoloV1	Anchor-free model	The center of each object in the original image is in one of the grids. Therefore, simply make each grid in charge of predicting the objects whose centroids fall in the current grid.	● Inaccurate positioning ● Low recall rate	● Network improvements ● Add bath Normalization to the network ● Add High Resolution Classifier to the network
YoloV2	Anchor-based model	● Better, Faster, Stronger	● The anchor mechanism is difficult to design with the super-references that need to be set ● Lots of redundant boxes ● Anchor box for target category classification with IOU threshold over-reference setting	● FPN structure added ● Object classification with logistic instead of softmax

YoloV3	Darknet-53 model	● Multi-scale detection ● Support for multiple label objects	● 1.The strategy construction loss function for binary classification leads to biased outcome values ● The same point, the same output layer, the same anchor first filled data will be overwritten by the latter ● Not high precision	● Data input using mosaic ● Added CSP darknet53 ● Added spp to increase the feeling of wildness ● PANet was used after Ssp for feature fusion ● Adopting CIOU loss
YoloV4	CSP Darknet-53 model	● The same accuracy, YOLOv4 is faster, the same speed, YOLOv4 is more accurate ● Can be trained and used on a normal GPU ● Have the right tricks to improve detector performance	● Multi-scale training is too inefficient, and SNIPER is complex ● Classification should not be too accurate, too accurate easily over-fitted	

3. Application of YOLO in Various Fields

After deep learning is widely used in image classification and speech recognition, deep neural networks are being applied to self-driving cars, including with regards to planning and determination, sensing, mapping and localization, etc. The YOLO series algorithms satisfy the real-time performance of target detection, but with some sacrifices in detection accuracy [6]. In particular, YOLO v1 is easy to miss detection for small objects with dense distribution. YOLO v3, YOLO v4 and the subsequent YOLO v5 and YOLO X updates make YOLO series algorithms more suitable for engineering practice, with their accuracy and real-time have changed significantly (Table 2). The following will focus on the application of YOLO in automatic driving and unmanned aerial vehicle.

Table 2. Performance comparison between different YOLO's

	Network	Advantage and disadvantage	FPS	Using scene
YOLO v1	VGG-16	Simple network, fast detection, but poor positioning, small target detection effect is poor	45.0	Detection of objects
YOLO v2	DarkNet-19	Reduce positioning errors, high classification accuracy, but not high accuracy	40.0	Detection of objects
YOLO v3	Darknet-53	Accuracy is improved, more than three times faster, but the boundary box prediction is not accurate enough in strict cases	19.6	Multi-scale target detection
YOLO v4	CSPDarknet-53	The small target detection effect is improved and the model complexity is high	65.0	High precision real-time target detection
YOLO X	DarkNet-53	Improved over-optimization of pipelines based on anchor points, multiple network frameworks can be selected	68.9	High precision real-time target detection
SSD	VG-16	The detection speed is fast, the accuracy is not high, and the detection effect of small target is not good	19.0	Multi-scale target detection

3.1 Application of YOLO in Automatic Driving

3.1.1 Detection of Traffic Signs

Zhang et al. reduce the computational complexity by reducing the convolutional layers at the top layer. yang et al. trains the training set using YOLO v3 and YOLO v4, respectively. Dewi et al. use generative adversarial networks using YOLO, YOLO v3 and YOLO v4 iterations. The recognition performance was improved after image blending. Mohd-Isa et al. further recognized far smaller traffic signs in real environments by including spatial pyramid pooling (SSP) in the YOLO v3 framework.

Mangshor et al. used the YOLO training model in order to perform recognition of five warning traffic signs, including intersection, right intersection, left intersection, Malaysian traffic sign of school children crossing and collision [7]. Wo-Feng Lu et al. adapted the loss function of the YOLO v5 algorithm to enhance the accuracies of the loss algorithm, achieve faster target recognition, and improve the accuracy of generating check frames.

In addition, traffic signal images are one of the most important pieces of information on the road. Traffic signal detection is an important component of intelligent vehicles and driver assistance systems. Jensen et al. applied the implementation of the target detection system YOLO to detect datasets such as public LISA traffic. Possatti et al. used the YOLO v3 model, in combination with an a priori map used by smart cars, to identify traffic lights associated with predefined routes. Gao used the YOLO v3 and YOLO v4 algorithms for the detection and recognition of traffic signal images on the road experimentally.

3.1.2 Detection of Traffic Vehicles

With the increase in traffic requirements, the pressure on traffic on the road is also increasing. The types of vehicles that require testing vary and include motorized and non-motorized vehicles.

Ye Jialin et al. designed feature fusion structure and used GIOU loss function to improve localization accuracy in YOLO v3 framework. Zhou et al. implemented deep fusion for target detection and recognition of vehicles by YOLO v2 algorithm. Chengbiao Zhang et al. changed the original YOLO v2 network framework and added and improved it to improve the detection accuracy for targets of different sizes in images [8].

Chen et al. proposed YOLO v3-live, a cut of the Tiny-YOLO v3 network layer structure, which is much faster. Wu et al. presented YOLO v5- Ghost based on YOLO v5's neural network structure. adjusted the network layer structure of YOLO v5s to reduce the computational complexity, which is better suited for embedded devices.

3.1.3 Pedestrian Detection

In transportation, pedestrians are an equally important part of the research. Intelligent or driverless vehicles need to detect pedestrians, recognize their limb movements, and understand the significance of their behavior before making appropriate response decisions.

Lan et al. changed the number of penetration layer connections from 16 to 12 layers in the network structure of YOLO v2. The detection speed was improved, while the missed detection rate was reduced by the experimental results. Gao Zong et al. proposed to change the YOLO network structure to combine the characteristics of the small aspect ratio of pedestrians in images in the original YOLO network structure to form a network structure suitable to pedestrian detection [9]. Hao, Xuzheng et al. proposed a pedestrian detection and recognition method based on deep residual network and YOLO model Zhang et al. based their Caps-YOLO detection model on YOLO v3, which used dense connection instead of shortcut connection in the original network to reduce the probability of missing and false detection of pedestrians with complex postures. Liu et al. used K-means clustering to directly calculate the frame size of anchor using the dataset to improve the detection accuracy of small-scale pedestrian targets in intelligent driving systems.

Zhang et al. proposed to add three layers of convolution to the original network based on the Tiny-YOLOv3 network and introduced a 1×1 convolution kernel to downscale the features to ensure the real-time detection.

3.1.4 Lane Line Detection

Achieving efficient lane detection is one of the important parts of the road environment perception module for unmanned vehicles. The accuracy of lane detection is related to the safety of driverless cars.

Zhang et al. constructed a two-stage learning network based on YOLO v3. The structural parameters are modified, which can improve the accuracy of off-road detection in complex scenes. Cui Wenliang et al. combined the original YOLO v3 with the K-means ++ clustering algorithm to

optimize the network anchor parameters, and improved the convolutional layer structure of the YOLO v3 algorithm, that is, the accuracy of the final result was improved. Ji et al. divided the lane line map into $S \times 2S$ grid based on the characteristics of inconsistent density of longitudinal and lateral distribution of lane line map, and simplified the grid to adjust the convolution layers in the original YOLO v3 from 53 layers to 49 layers, this change is more suitable for the detection of small targets such as 3 lanes. Xiang Zhang et al. also used the YOLO v3 algorithm with $S \times 2S$ grid density and proposed a lane line detection model based on the lane line distribution relationship, which improved the accuracy of the model to detect small size and large aspect ratio objects.

3.2 Application of YOLO in Unmanned Aerial Vehicle

The unmanned aerial vehicle (UAV) market is showing rapid development, bringing great convenience to people's colleagues, and unmanned aerial vehicle incidents are gradually rising. In unmanned aerial vehicle detection, technologies based on acoustic, radio and radar monitoring are more common, but these technologies often require expensive equipment and strict configuration. With the development of deep learning technology, it has been gradually applied to the identification and detection of various types of targets.

3.2.1 Features of UAV Target Detection

- Complex detection background. Low flight altitude, urban background as the detection scene, wake complex, more obscurants, weather, light, noise, flying birds and other uncertainties, adding to the difficulty of detection.
- Detection targets are not significant. The UAV is small and has a long distance to the target. Since the UAV is in motion, the captured images are prone to motion blur, and there are multiple angles and attitudes of the in-flight UAV images captured from different viewpoints, making them more difficult to detect [10].
- Short detection time frame. UAVs move fast, have short acceleration time and strong mobility, and are not easy to capture images, which require timely detection of valid information in order to buy time for early warning and disposal.

3.2.2 UAV Target Detection Requirements

- Target classification. Because there is more than one type of UAV, the task of target classification is first to determine whether a UAV target appears in a digital image or video, and then to output a label for the category to which the UAV belongs.
- Target location. Outputs the location and range of the UAV target, marking out the center of the UAV or drawing the closed boundary of the UAV with a square enclosing box.

The YOLO target algorithm uses the idea of regression to speed up the detection and improve the background accuracy, but the positioning accuracy and recall rate are still not ideal, and it is difficult to detect small targets. Liu et al. proposed SSD, YOLO, and Faster RCNN to solve the localization accuracy problem. Because the RCNN series algorithm has a more accurate detection accuracy for the target, the YOLO series algorithm has an effective fusion between accuracy and speed.

4. Conclusion

Due to the rapid update of YOLO target detection technology and a large number of applications in various target detection fields, this paper summarizes the characteristics and applications of YOLO technology. This paper mainly analyzes the main characteristics of YOLO technology and the technology and characteristics changed after its update. The mechanism, structural characteristics and advantages and disadvantages of yolo series are analyzed. The application of YOLO series detection technology in the field of unmanned aerial vehicle and UAV target recognition is summarized. Especially in the development of the field of unmanned driving, many researchers use YOLO to solve problems in this field. YOLO target detection technology is a flexible technology. Researchers can make specific changes to the YOLO for different scenarios, making it applicable to more scenarios.

References

- [1] Bevilacqua V, Triggiani M, Gallo V, et al. Intelligent Computing Theories and Applications. Springer Berlin Heidelberg, 2010.
- [2] Zhihong Zhang, Huajun Gong, Xinhua Wang, Yadong Feng, Minjie Xu. An Improved Yolov3 Object Detection Algorithm for UAV Aerial Images, 2021 International Conference on Intelligent Computing, Automation and Applications (ICAA), 2021, 191(13): 153-158.
- [3] Malik Haris, Jin Hou, Xiaomin Wang. Lane Lines Detection under Complex Environment by Fusion of Detection and Prediction Models, Transportation Research Record: Journal of the Transportation Research Board, 2014, 110(10): 503-511.
- [4] Ghada Hamed Aly, Mohammed Abd El- Rahman Marey, Safaa El-Sayed Amin, Mohamed Fahmy Tolba. Chapter 15 YOLO V3 and YOLO V4 for Masses Detection in Mammograms with ResNet and Inception for Masses Classification, Springer Science and Business Media LLC, 2021: 1342-1351.
- [5] Huang D S , Mcginnity M , Heutte L , et al. Advanced Intelligent Computing Theories and Applications. Communications in Computer and Information Science, 2010, 93.
- [6] JIN Zhao-zhao, Zheng Yu-fu. Research on Application of Improved YOLO V3 Algorithm in Road Target Detection, Journal of Physics: Conference Series, 2020, 23(12): 435-452.
- [7] Shujian Liu, Haibo Zhou, Chenming Li, Shuo Wang. Analysis of Anchor-Based and Anchor- Free Object Detection Methods Based on Deep Learning, 2020 IEEE International Conference on Mechatronics and Automation (ICMA), 2020, 1132-1143.
- [8] Zongyi Xing, Zhenyu Zhang, Xiaowen Yao, Yong Qin, Limin Jia. Rail wheel tread defect detection using improved YOLOv3, Measurement, 2022,674-686.
- [9] Zhang B , Qian X , Yang R , et al. Water Surface Target Detection Based on Improved YOLOv3 in UAV Images, 2021 9th International Conference on Communications and Broadband Networking. 2021 24 (3): 348-359.
- [10] Mengying Cao, Junsheng Wang, Yantong Chen, Yuezhu Wang. Detection of Microalgae Object Based on the Improved YOLOv3 Model, Environmental Science: Processes & Impacts, 2021, 19(102): 43-48.