IOP Conference Series: Earth and Environmental Science

PAPER • OPEN ACCESS

# An improved method of Tiny YOLOV3

To cite this article: Xiaotian Gong *et al* 2020 *IOP Conf. Ser.: Earth Environ. Sci.* **440** 052025

View the article online for updates and enhancements.

You may also like

- Research on the use of YOLOv3 image processing algorithm in power plants
Junjie Zhao

- A two-stage CNN for automated tire defect inspection in radiographic image
Zhouzhou Zheng, Sen Zhang, Jinyue Shen et al.

- YOLO-RD: A lightweight object detection network for range doppler radar images
Long Zhou, Suyuan Wei, Zhongma Cui et al.

Recent citations

- Deep Learning-Based Detection of Articulatory Features in Arabic and English Speech
Mohammed Algabri *et al*

This content was downloaded from IP address 108.173.253.238 on 21/12/2021 at 09:23

# An improved method of Tiny YOLOV3

**Xiaotian Gong[a], Li Ma[b] and Hangkong Ouyang[*]**

School of Mechatronic Engineering and Automation, Shanghai University, Shanghai, China

*Corresponding author e-mail: 15202151595@163.com, [a]gongxt@shu.edu.cn, [b]1379076820@qq.com

**Abstract**. Target detection is the basic technology of self-driving system. In this paper, the problem of high detection rate of pedestrians and other small targets is studied in real-time detection of Tiny YOLOV3 target detection algorithm, and the network structure of Tiny YOLOV3 algorithm is improved. 2-step convolutional layers are added to the network, and deep separable convolution constructs are used to replace the traditional convolutions. On the basis of the original two-scales prediction target of the network, a scale is added to form a three-scales prediction, which can makes the detection of small targets such as pedestrians more accurate. The experimental results show that the average accuracy of the improved target detection algorithm is 8.6% higher than that of Tiny YOLOV3, and it meets the real-time requirements and has certain robustness.

## 1. Introduction

With the development of society and the advancement of science and technology, the self-driving system has become a major trend in modern transportation. The target detection method is an indispensable part of the self-driving system and can be divided into shallow learning and deep learning methods. Shallow learning generally obtains the corresponding relationship between the template and the object in the scene according to the feature points of the template to detect the target, and then uses AdaBoost [1] algorithm to extract features and support vector machine [2] for classification and so on to achieve target detection. Deep learning has developed into two categories: one is the target detection method combined with prediction box and convolutional neural network classification, which is a two-stage algorithm; the other is to convert target detection into regression problem for processing, which is a single-stage algorithm. The two-stage algorithm starts from the R-CNN [3] network and uses the selective search method instead of the sliding window to greatly improve the target detection accuracy. The single-stage algorithm is started by Redmon J et al., which proposes the YOLO [4] algorithm. Liu W et al. proposed the SSD [5] algorithm, which combines the regression ideas of Faster R-CNN and YOLO. Redmon J et al. continued to propose the YOLOv2 [6] and YOLOv3 [7] algorithms to further improve target detection speed and accuracy.

Tiny YOLOV3 is a simplified version of YOLOV3. It is a real-time detection algorithm developed for embedded devices with poor data processing capabilities. The model structure is simple and is currently the most Fast target detection algorithm, but the detection accuracy is low, especially in small target detection such as person, the miss detection rate is higher. In this paper, Tiny YOLOV3 is improved for this problem. In the feature extraction network stage, a 2-step long convolution layer is

Content from this work may be used under the terms of the Creative Commons Attribution 3.0 licence. Any further distribution of this work must maintain attribution to the author(s) and the title of the work, journal citation and DOI.
Published under licence by IOP Publishing Ltd

added to replace the maxpooling layer in the original network for downsampling. More convolutional layers can better extract target features, but will also add more parameters and calculations. Therefore, we use the deep separable convolution construct anti-residual block to replace the traditional convolution to ensure fast calculation to meet real-time detection requirements. On the basis of the two-scale prediction of Tiny YOLOV3's prediction network, a scale is added to form a three-scale prediction, which further improves the detection accuracy of small targets such as person.

## 2. Improved tiny-yolov3 network

### *2.1. Introduction to Tiny YOLOV3*

Tiny YOLOV3 is a lightweight target detection algorithm applied to embedded platforms based on YOLOv3. Although the detection accuracy is lower than YOLOv3, the model size compression is implemented. Tiny YOLOV3 reduced the YOLOv3 feature detection network darknet-53 to a 7-layer traditional convolution and a 6-layer Max Pooling layer, using a 13*13, 26*26 two-scale prediction network to predict the target. The network structure is shown in Figure 1.

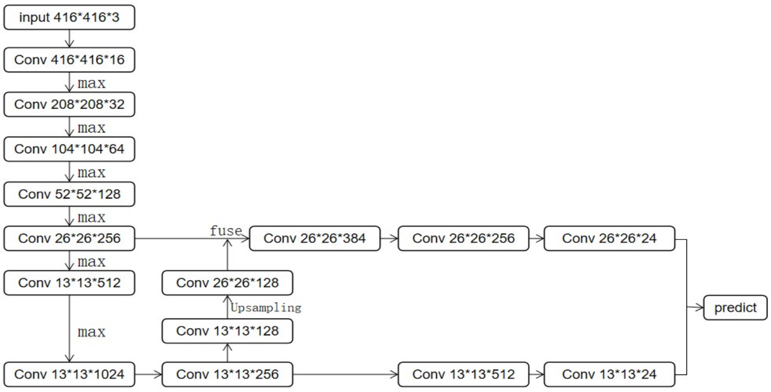


**Figure 1.** Tiny YOLOV3 network structure.

The loss function of Tiny YOLOv3 is mainly defined from three aspects: the bounding box position error term, the bounding box error term and the classification prediction error term:

$$\begin{aligned} Loss= & \lambda_{coord}\sum_{i=o}^{s^2}\sum_{j=o}^{B} l_{ij}^{obj}[(x_i-\overset{\wedge}{x}_i)^2+(y_i-\overset{\wedge}{y}_i)^2]+\lambda_{coord}\sum_{i=o}^{s^2}\sum_{j=o}^{B} l_{ij}^{obj}[(\sqrt{w_i}-\sqrt{\overset{\wedge}{w}_i})^2+(\sqrt{h_i}-\sqrt{\overset{\wedge}{h}_i})^2] \\ & +\lambda_{coord}\sum_{i=o}^{s^2}\sum_{j=o}^{B} l_{ij}^{noobj}(C_i-\overset{\wedge}{C}_i)^2+\lambda_{coord}\sum_{i=o}^{s^2}\sum_{j=o}^{B} l_{ij}^{obj}(C_i-\overset{\wedge}{C}_i)^2 \\ & +\sum_{i=o}^{s^2} l_i^{obj}\sum_{(c\in classes)}(P_i(c)-\overset{\wedge}{P}_i(c))^2 \end{aligned} \tag{1}$$

Where, the first term is the position coordinates error term, the second term is the confidence error term, and the third term is the classification prediction error term.

### *2.2. Constructing an anti-residual block*

The calculation of the parameters of the traditional convolution increases exponentially with the increase of the number of network layers, which leads to the increase of the model size and the difficulty of meeting the requirements of real-time rapid detection. This paper replaces traditional convolution with deep separable convolution, transforming traditional convolution into deep convolution and point-by-point convolution: deep convolution is spatially convolved independently on each input channel; point-by-point convolution maps the output of the deep convolution to the new channel space. This decomposition can effectively reduce the amount of calculation and the size of the model, and improve the real-time of the convolution. However, this method causes the number of network layers to deepen, and the gradient disappears. Although the residual structure can solve this problem, the residual structure has the problem that the compression feature map leads to the damage of the feature expression. Therefore, this paper uses the anti-residual module in the feature extraction process. The channel is first expanded by a 1x1 convolutional layer, then the 3x3 deep convolutional layer is applied to extract the high-dimensional features, and finally, the depth convolution result is mapped to the new channel space by the 1x1 point convolutional layer. The residual structure and anti-residual structure are shown in Figure 2.

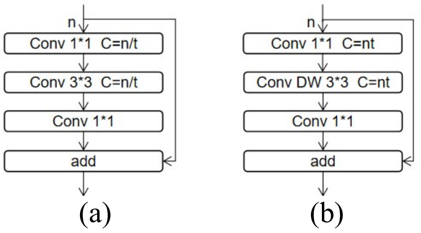


**Figure 2.** (a) residual structure (b) anti-residual structure.

Where: n is the number of input channels, t is a multiple of the extended or compressed channel, and C is the number of channels.

The calculation of the parameters of the anti-residual block in this paper is shown in Table 1.

**Table 1.** Anti-residual block.

| Input | Operation | Output |
|---|---|---|
| h*w*n | point conv 1*1, Relu | h*w*2n |
| h*w*2n | depthwise conv 3*3, Relu | h/s*w/s*2n |
| h/s*w/s*2n | point conv 1*1 | h/s*w/s*2n |

Where: h, w is the height and width of the feature map; n is the number of channels of the feature map; s is the step size.

### *2.3. Network model improvement*

In order to solve the problem that yolov3-tiny has low precision and high false detection rate for small targets such as person, this paper improves the network model of Tiny YOLOV3. The network structure is shown in Figure 3. Among them, The ARB is an anti-residual block, and the dotted line box is a network feature extraction part.

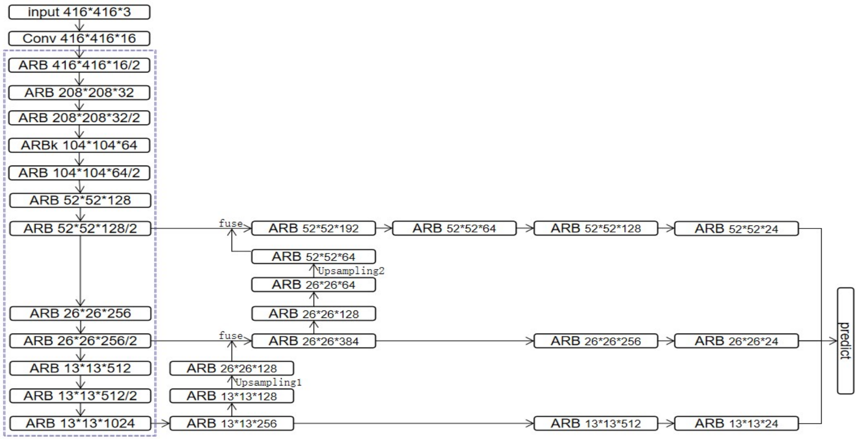


**Figure 3.** Improved Tiny YOLOV3 network model.

In the feature extraction network, the feature extraction is increased by adding convolution layers, the maxpooling layer in the original network is replaced by 2-step convolutions, and the anti-residual block constructed by the depth separable network is used to replace the traditional convolution. The improved feature extraction network consists of 12 anti-residual blocks. While increasing feature extraction, the computational complexity of the model is effectively reduced. At the same time, based on the 26x26, 13x13 two-scales detection target of the original network, an upsampled layer upsample2 is added to form 52x52, 26x26, 13x13 three-scales detection, which can better detect small targets such as person.

## 3. Experiments

### *3.1. Test environment and dataset*

The whole experiment in this paper is implemented in keras with TensorFlow as the back end. The experimental environment is: Cuda 9.0 acceleration; hardware configuration is AMD Ryzen 5 2600 Six-Core Processor@ 3.4GHz, GeForce GTX 1080Ti graphics card; operating system: Ubuntu 18.04.1 LTS.

For the problem under study, this article uses a mixed dataset. The autopilot dataset KITTI was first used and the original dataset 8 categories were re-divided: cars, trucks and trucks were merged into motor vehicles; pedestrians, pedestrians and cyclists were merged into humans; the last two were deleted. Then mix the INRIA pedestrian dataset and the VOC2007 pedestrian dataset, and finally get a mixed dataset of a total of 50,000 images, with the ratio of motor vehicles to pedestrians approaching 1:1. This prevents over-fitting problems caused by excessive sample gaps in the dataset. 70% of the new data set is used for training, 20% for verification, and 10% for testing. The surroundings of the mixed dataset are complex, and the person's posture is diverse. The size of the target and the degree of occlusion on the image in the dataset are also different. These can enhance the generalization of training models and meet the roads of complex traffic scenarios.

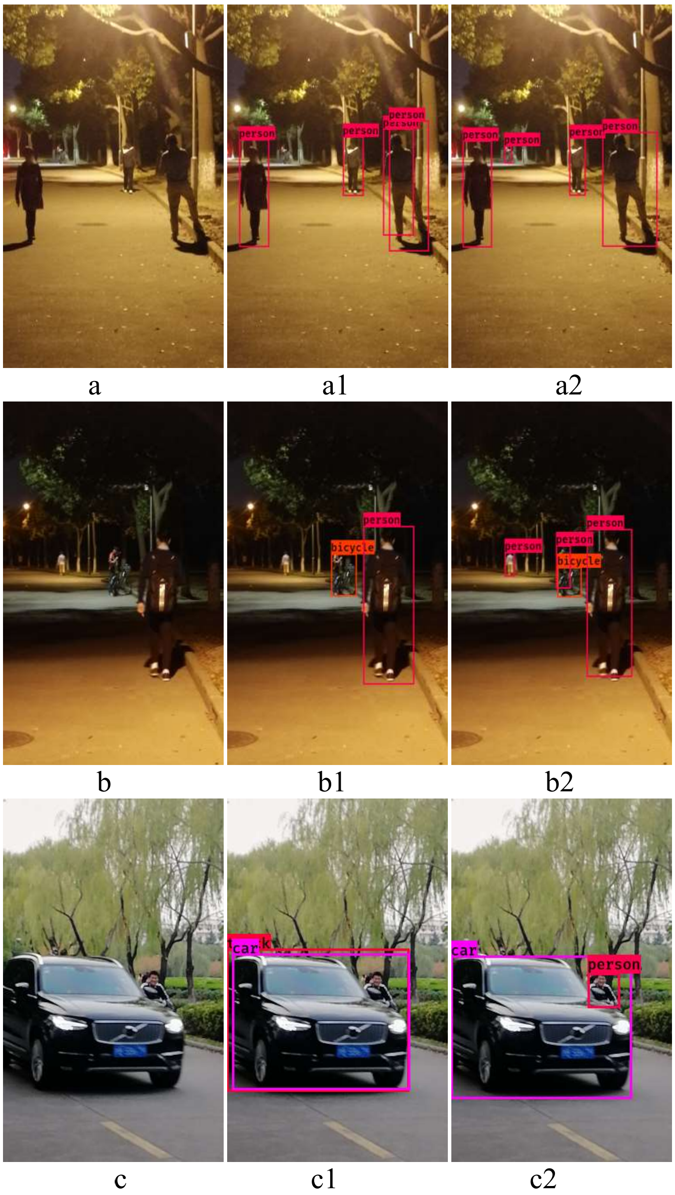
### *3.2. Experimental result*

In order to evaluate the effectiveness of the improved algorithm, the data set is trained in various advanced target detection algorithms and improved algorithms, and the mean precision mean (mAP) and speed fps are calculated. The test results are shown in Table 2.

**Table 2.** Test results of each algorithm.

| algorithms | mAP/% | fps/(frame·s$^{-1}$) |
|---|---|---|
| Faster R-CNN | 81.4 | 7.5 |
| SSD500 | 81.6 | 18.4 |
| YOLOV3 | 82.1 | 22.6 |
| Tiny YOLOV3 | 51.8 | 37.4 |
| Our algorithm | 60.4 | 34.6 |

As can be seen from the above table, the detection accuracy of large networks such as fast R-CNN, SSD500, and YOLOV3 is much higher than that of Tiny YOLOV3 and our improved algorithm, but these large networks have many parameters and large calculations, resulting in low fps. Comparing Tiny YOLOV3 with our improved algorithm, although the fps is reduced by 2.8 and the detection speed is slightly slower, it still meets the real-time detection requirements. Moreover, the average accuracy of our improved algorithm is 60.4%, which is 8.6% higher than the accuracy of 51.8% of Tiny YOLOV3 algorithm. The improved algorithm improves the detection accuracy while ensuring real-time monitoring.



a a1 a2

b b1 b2

c c1 c2

**Figure 4.** The improved algorithm and the detection effect of the Tiny YOLOV3 algorithm in real-time scenarios.

Figure 4 shows the improved algorithm and the detection effect of the Tiny YOLOV3 algorithm in real-time scenarios. A,b,c show three original images, a1,b1,c1 show Tiny YOLOV3 detection results, and a2,b2,c2 show the improved algorithm detection results. It can be seen that the improved algorithm has better detection effect on small targets such as person, and still has good detection effect under weak light environment, which indicates that the improved algorithm also has good adaptability to complex environments. In summary, the improved algorithm has better accuracy than the original algorithm, and can meet the requirements of real-time detection, and has certain robustness.

## 4. Conclusion

In this paper, an improved tiny yolov3 algorithm is proposed. The convolution layers are added in the feature extraction stage, and the original network maxpooling layers are replaced by 2-step convolutions for downsampling. The deep convolutional convolutions are used to replace the traditional convolutions to solve the computational problem of a large number of parameters caused by the increase of the convolutional layers, and the anti-residual blocks are constructed according to the depth scalable convolution to extract high-dimensional spatial features and reduce information loss. On the basis of the two-scales prediction target of tiny yolov3, a scale is added to form a three-scales prediction, which can better detects small targets such as person and improves detection accuracy. However, the detection accuracy of the algorithm is still low, and there is still a big gap compared with the large-scale network. How to improve the accuracy of target detection on the basis of satisfying the real-time detection is the direction for the next step.

## References

[1] Lowe D G. Distinctive image features from scale-invariant key points [J]. International Journal of Computer Vision, 2004, 60(2):91-110.

[2] Freund Y, Schapire R E. Experiments with a new boosting algorithm [C] // Proceedings of Thirteenth International Conference on International Conference on Machine Learning, 1996:148-156.

[3] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // Proceedings of the IEEE conference on computer vision and pattern recognition, 2014:580-587.

[4] Redmon J, Divvala S, Girshick R, et al. You only look once:Unified, real-time object detectiono [C] // Proceedings of CVPR 2015, 2015:779-788.

[5] Liu W, Anguelov D, Erhan D, et al. SSD: Single shot multibox detector [C] // Proceedings of European Conference on Computer Vision, 2016: 21-37.

[6] Redmon J, Farhadi A. YOLO9000: Better, faster, strong-er [C] // Proceedings of CVPR 2016, 2016.

[7] Redmon J, Farhadi A. Yolov3: An incremental improvement [J]. arXiv preprint arXiv: 1804.02767, 2018.