ISSN: 2414-1895 DOI: 10.6919/ICJE.202009_6(9).0006

BiNet: Vehicle and Pedestrian Detection Network Based on Attention Mechanism Combined with Two-way Weighted Feature Fusion

Jinhui Diao

College of Information Engineering, Shanghai Maritime University, Shanghai, 201306, China.

1421142577@qq.com

Abstract

In order to solve the environment detection problem of autonomous driving, this paper proposes a vehicle and pedestrian detection network BiNet based on deep learning, which is an improved YOLO v3 target detection algorithm based on convolutional neural network. The traditional top-down feature fusion method is essentially limited by the one-way information flow, making the deep semantic information unable to experience the fine-grained features of the shallow layer. Aiming at the above problems, this paper proposes a two-way weighted feature fusion method. At the same time, an attention mechanism algorithm is introduced to learn the correlation of features between different channels by weighting the features of each channel. And use Focal loss, and GIOU loss to design the loss function of the detector. Experiments show that compared with the original Yolov3, BiNet has increased the average detection accuracy of pedestrians and vehicles in the data sets PascalVOC2007, PascalVOC2012, and MS COCO by 1.9%, 2.7%, and 2.5%, respectively.

Keywords

Convolutional Neural Network; Pedestrian and Vehicle Detection; Feature Fusion; Attention Mechanism; BiNet.

1. Introduction

As the basis of unmanned driving, environmental perception has become a hot field in academic circles in recent years. Among them, pedestrian and vehicle detection, as an important content of environmental perception, has also become a current research focus. In recent years, detection has become a research hotspot. With the continuous development of technology, many neural network-based algorithms have been proposed. Compared with traditional algorithms, neural network-based algorithms have more powerful generalization capabilities and automatic learning goals. Semantic information.

At present, target detection algorithms based on convolutional neural networks can be divided into two categories: the first category is two-stage target detection algorithms, such as R-CNN[1], Fast R-CNN[2], Faster R- CNN[3],etc.these algorithms first use Selective Search or Region Proposal Network (Region Proposal Network, RPN) to generate candidate regions, and then use the detection network to further regress the location and type of the target; the second type is One-stage (one-stage) target detection algorithms, such as YOLO[4], SSD[5], YOLO v3[6], etc.these algorithms directly use the detection network to return the location and category information of the target. The two-stage target detection algorithm has higher detection accuracy, and the single-stage target detection algorithm has faster detection speed.

In order to balance accuracy and speed, the traditional multi-scale feature pyramid is widely used in the target detection network, and the robust semantic information of the high-level features is effectively integrated through the top-down cross-layer path to generate the feature pyramid Network, YOLO v3 is based on the neural network Darknet53 to extract multi-scale target detectors with different resolutions. FPN uses a top-down, side-to-side connection method to fuse the features of two adjacent scales. The high-resolution feature map contains more fine-grained features of the target, and the low-resolution feature map contains more semantic information. Feature fusion can effectively improve the accuracy of target detection.

Although the traditional multi-scale target detector of FPN [7] has achieved good detection results, there is still room for improvement. First, the fusion of features between adjacent scales does not make full use of the resolution from other scales. The low-level information has a higher resolution, which can achieve good results for the detection of small targets.

the main contributions of this article are as follows:

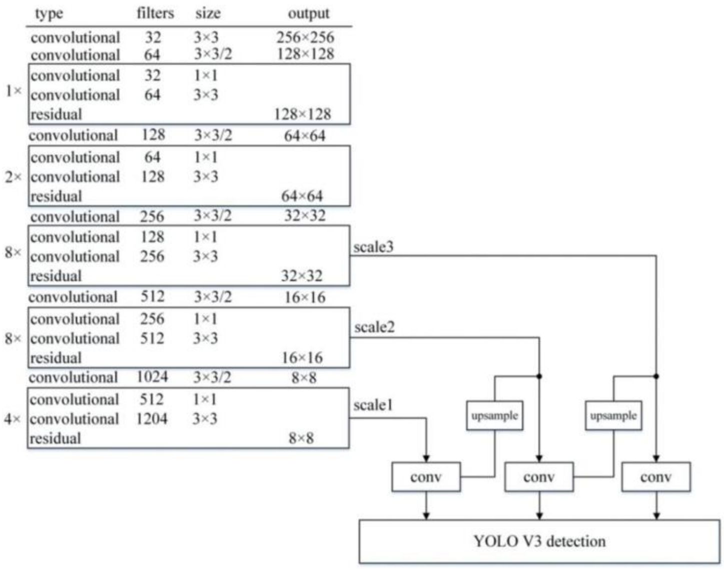
in order to make full use of the complementarity of different scale feature layers in visual semantic information, according to the difference of different resolution feature layers in visual semantic information and precise positioning information, this paper proposes a two-way weighted The feature fusion network realizes the weighting of different levels of feature information without increasing any time consumption, so that deep semantic features and shallow fine-grained features can be more fully integrated to enhance the semantic robustness of the feature pyramid And positioning accuracy.

In the process of feature fusion, the attention mechanism is introduced to redistribute the weight of feature maps of different channels.

In order to further improve the detection accuracy of the network, Focal loss[8] and Generalized Intersection over Union (GIOU) loss [9] are used to design the loss function of the detector.

2. YOLO v3 Network

The Darknet53 network structure adopted by the YOLO v3 network contains a total of 53 convolutional layers. See Fig. 1, Darknet53 is composed of 5 residual blocks, which draws on the idea of Resnet[10] residual neural network. Each residual block is composed of multiple residual units, and the residual unit is constructed by inputting the residual operation with two digital cumulative modeling (DBL) units, see Fig. 2(a). Among them, the DBL unit includes convolution, batch normalization and activation functions, Fig. 2 (b). The residual unit is introduced to increase the depth of the network and avoid the disappearance of the gradient.



	type	filters	size	output
	convolutional	32	3×3	256×256
	convolutional	64	3×3/2	128×128
1×	convolutional	32	1×1	
	convolutional	64	3×3	
	residual			128×128
	convolutional	128	3×3/2	64×64
2×	convolutional	64	1×1	
	convolutional	128	3×3	
	residual			64×64
	convolutional	256	3×3/2	32×32
8×	convolutional	128	1×1	
	convolutional	256	3×3	
	residual			32×32
	convolutional	512	3×3/2	16×16
8×	convolutional	256	1×1	
	convolutional	512	3×3	
	residual			16×16
	convolutional	1024	3×3/2	8×8
4×	convolutional	512	1×1	
	convolutional	1204	3×3	
	residual			8×8

Fig. 1 Yolo v3 network structure

YoloV3 performs 5 downs on the input image using YOLOV3 and predicts the target in the last 3 downsampling. The last 3 downsampling contains 3 scale target detection feature maps, and the small feature maps provide deep semantic information, The large feature map provides the location information of the target, and the small feature map is up-sampled and merged with the large feature map, so the model can detect both large targets and small targets.

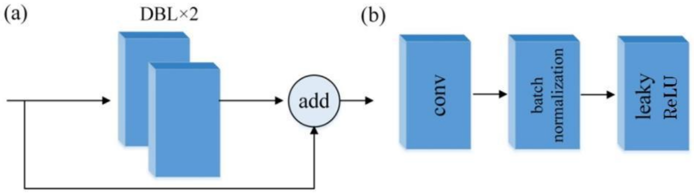


Fig. 2 Darknet 53 structural unit

3. Improved YOLO V3 network model introduction

The network proposed in this paper is improved based on yolov3, darknet53 is used as the feature extraction network, and the traditional FPN feature fusion method is improved to two-way weighted feature fusion, and the attention mechanism algorithm is combined with the fused channel. And use Focal loss and GIOU loss to design the network loss function, the improved network structure is see Fig. 3.

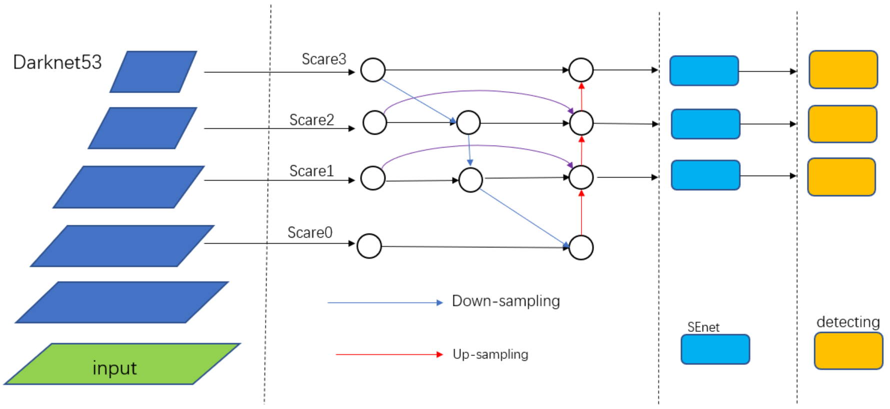


Fig. 3 Improved network structure of YOLO V3

3.1 Fusion algorithm based on attention mechanism.

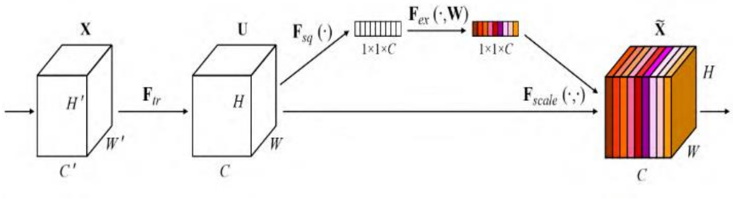


Fig. 4 Senet attention unit

For convolution operations, a large part of the work is to improve the receptive field, that is, to fuse more feature fusion spatially, or to extract multi-scale spatial information. In this paper, the attention mechanism is added to the process of feature fusion, and the attention mechanism is to focus on the relationship between channels, hoping that the model can automatically learn the importance of different channel features, and adjust the weights between different channels[11], enhanced volume The expressive ability of product neural network. The goal is to adjust the semantic information of the original network by relearning the interdependence between the convolution feature channels and changing the weight. The attention unit of Senet [12] is see Fig. 4.

The SE module first performs Squeeze (global pooling) operation on the feature map obtained by convolution to obtain the channel-level global receptive field, and then uses the fully connected layer to perform dimensionality reduction operations, and uses the Relu activation layer to learn the nonlinearity between the feature channels At the end of the relationship, the feature map is upgraded using the fully connected layer. The Excitation operation uses the Sigmoid activation function to output the weight of the dimension C*1*1. The feature's reweight operation uses the input feature map and the weight value obtained from the Excitation operation to perform a product operation to redistribute the weight of each channel feature. With the Squeeze-and-Excitation Block module, the network can automatically learn the correlation and importance of feature channels.

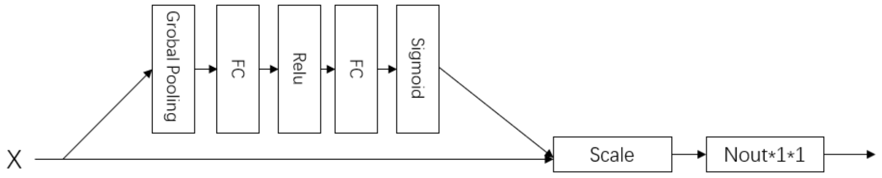
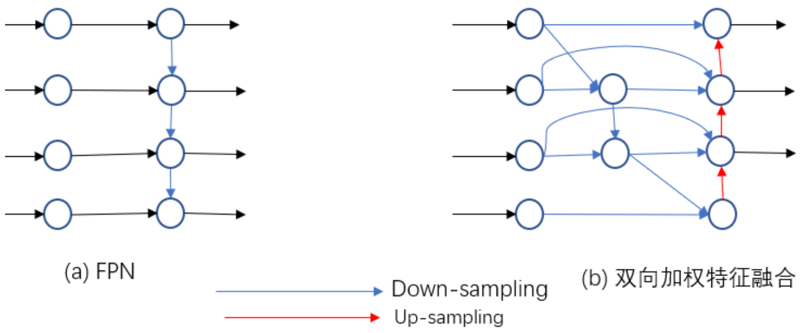


Fig. 5 The improved Senet attention unit in this paper

The attention mechanism fusion process proposed in this paper is based on the Senet of Squeeze-and-Excitation and Nout*1*1 convolution structure, as see Fig. 5, so that the network can automatically learn feature channels between The relevance and importance of. Among them, Nout*1*1 convolution is used to predict the position and type of the target, Nout=Nboxs*(Nclasses+4+1), 4 and 1 represent the 4 offsets of the predicted target position and the confidence of the target.

3.2 Tow-way weighted feature fusion detection network.

In the neural network, the high-level feature maps of the network will have the characteristics of low resolution and high semantic information, while the low-level feature maps have the characteristics of high resolution and low semantic information, but the information of these layers has the characteristics of small target objects. The classification is very helpful. The role of low-level semantic information can often not be ignored. The primary goal of establishing an information fusion mechanism is to make full use of the low-level information of the network .



(a) FPN

(b) 双向加权特征融合

Fig. 6 Traditional FPN and two-way weighted feature fusion structure

The traditional FPN, as see Fig. 6(a), only includes top-down feature fusion. It can be seen that the fusion process only includes the fusion from low resolution to high resolution, that is, from the deep features of the network to the network The shallow feature fusion is the top-down mode. The traditional top-down FPN is essentially limited by one-way information flow, which makes the deep semantic information unable to experience the fine-grained features of the shallow layer,such as EfficientNet[13], and does not consider the complementarity of different information see Fig. 6(b).

3.3 Loss function

In order to further improve the detection accuracy of the model, GIOU loss and Focal loss are used to design the loss function of the multi-scale target detector. Compared with Intersection over Union (IOU), GIOU can reflect the distance between the prediction box and the target box. Moreover, GIOU loss can effectively avoid the problem that the gradient of the loss function is 0 because the prediction frame does not intersect with the ground truth. Therefore, this article uses GIOU loss to regress the position of the prediction box. GIOU can be expressed as:

$$GIOU_{B_{GT},B_P} = \frac{|B_{GT} \cap B_P|}{|B_{GT} \cup B_p|} = \frac{|\frac{B_{GT} \cup B_P}{B}|}{|B|} \tag{1}$$

Where: B_{GT} represents the ground truth of the target frame, B_P represents the prediction frame, and B represents the smallest rectangular frame enclosing B_{GT} and B_P. Then GIOU loss can be expressed as.

$$GIOUloss = 1 - GIOU_{B_{GT},B_P} \tag{2}$$

Since the training sample contains a large number of easily distinguishable negative samples, these simple negative samples will play a major role in the loss function. In order to enhance the network's ability to predict difficult samples, this paper uses Focal loss to regress the confidence of the target.

$$Loss_{conf} = -\alpha_1 * (y_{GT} - y_P)^2 * y_{GT}\, log\, y_P - \alpha_2 * (y_{GT} - y_P)^2 (1 - y_{GT})\, log(1 - y_P) \tag{3}$$

Where: y_{GT} represents the ground truth of the target confidence, and y_P represents the confidence of the predicted target.

For classification loss, binary cross entropy is used, as shown in the figure:

$$Loss_{cls} = C_{GT}\, log\, C_p - (1 - C_{GT})\, log(1 - C_p) \tag{4}$$

Where: C_{GT} represents the ground truth of the target category, and C_p represents the predicted target category.

The total loss can be expressed as:

$$Loss_{total} = \sum_{i=1}^{3} GIOUloss^i + Loss_{conf}{}^i + Loss_{cls}{}^i \tag{5}$$

4. Exprimental results and analysis

4.1 Experimental software and hardware configuration and experimental parameter settings

In terms of hardware, the training machine is equipped with GTX1080Ti GPU and 16G memory. In terms of software, 64bit-Ubuntu16.04 operating system is used; Tenorflow deep learning framework, Python programming is used, and dependent libraries such as Cuda 8.0 with cudnn and OpenCV are used.

The initial learning rate is set to 0.005, the number of iterations is 30, the maximum number of iterations is set to 30K, and the momentum is 0.9. The Adam optimization function is used to optimize the objective function of the network layer. The β1 coefficient is the exponential decay rate, and the weight distribution is controlled. The value is 0.9, the β 2 coefficient is the exponential decay rate, which controls the influence of the previous gradient square, generally the value is 0.999, the batch_size of the input image is set to 10, and the weight decay rate used to prevent overfitting is set to 0.005 . On each data set, training can be completed in about 21h.

4.2 Evaluation Index

When evaluating the effect of target detection, the Mean average precision (MAP) index is used. The specific formula is:

$$MAP = \int_0^1 P(R)dR \tag{6}$$

When calculating the accuracy and recall rates, set the coincidence rate of the prediction frame and the label frame to α, and consider the prediction frame with α ≥ 0.5 as a positive example:

$$\alpha = \frac{Box_{pre} \cap Box_{gt}}{Box_{pre} \cup Box_{gt}} \tag{7}$$

Where: Box_{pre} is the prediction box; Box_{gt} is the label box; ∩ is the intersection area of theprediction box and the label box; ∪ is the combined area of the prediction box and the label box.

4.3 Data set

The experiment uses 3 data sets for training, namely PascalVOC2007, PascalVOC2012, MS COCO. Target detection general data set. In addition to people and vehicles, the labeled objects include cats, dogs, airplanes, machines, chairs, TVs, etc. The picture scenes include a variety of rich indoor and outdoor scenes. PascalVOC2007 contains 9,963 pictures and 24,640 annotated targets, PascalVOC2012 contains 11,530 pictures and 27,450 annotated targets, MS COCO contains about 328,000 pictures and 2.5 million annotated targets. The experiment merges the labels of the vehicles in the three data sets, focusing only on vehicles and pedestrians, and other labels are not involved in training and testing, and are only used as background images

4.4 Analysis of experimental results

The AP and MAP of the feature fusion network based on the attention mechanism combined with two-way persuasion to detect vehicles and pedestrians on three public data sets are shown in the Table 1 and Table 2. The average accuracy rate of most pedestrian detection is slightly lower than the average accuracy rate of vehicle detection. This is because in the training set and test set, the image area occupied by pedestrians is often smaller than the image area occupied by vehicles, so it is more difficult to detect. In particular, on the PascalVOC 2012 data set,The average accuracy rate of pedestrian detection is higher than the average accuracy rate of vehicle detection. Analysis of the data set found that in PascalVOC 2012, pedestriansThe number of labels is much larger than vehicle.

Table 1 Vehicle and pedestrian AP on the three data sets

Network/dataset	PascalVOC2007		PascalVOC2012		MS COCO	
	VehicleAP	Pedestrian AP	VehicleAP	Pedestrian AP	VehicleAP	Pedestrian AP
YOLO V3	82.6	75.3	75.6	78.2	65.3	56.5
BiNET	84.4	77.2	77.3	81.9	68.7	58.2

Table 2 Vehicle and pedestrian MAP on the three data sets

Network/dataset	PascalVOC2007 Vehicle Pedestrian MAP	PascalVOC2012 Vehicle Pedestrian MAP	MS COCO Vehicle Pedestrian MAP
YOLO V3	78.9	76.9	60.9
BiNET	80.8	79.6	63.4

It can be seen from the table that the detection effect of the network proposed in this paper is higher than that of YOLO V3. It can be analyzed that the two-way weighted feature fusion breaks the traditional top-down fusion effect of FPN, and high resolution is achieved on each channel. Rate and low resolution are only merged together. By introducing an attention mechanism, the network can automatically learn the correlation and importance of feature channels, and improve the accuracy of the network's detection of targets. At the same time, the improved YOLO V3 loss function using GIOU loss and Focal loss can make the network's target location and classification more accurate. The specific detection effect see Fig. 7.

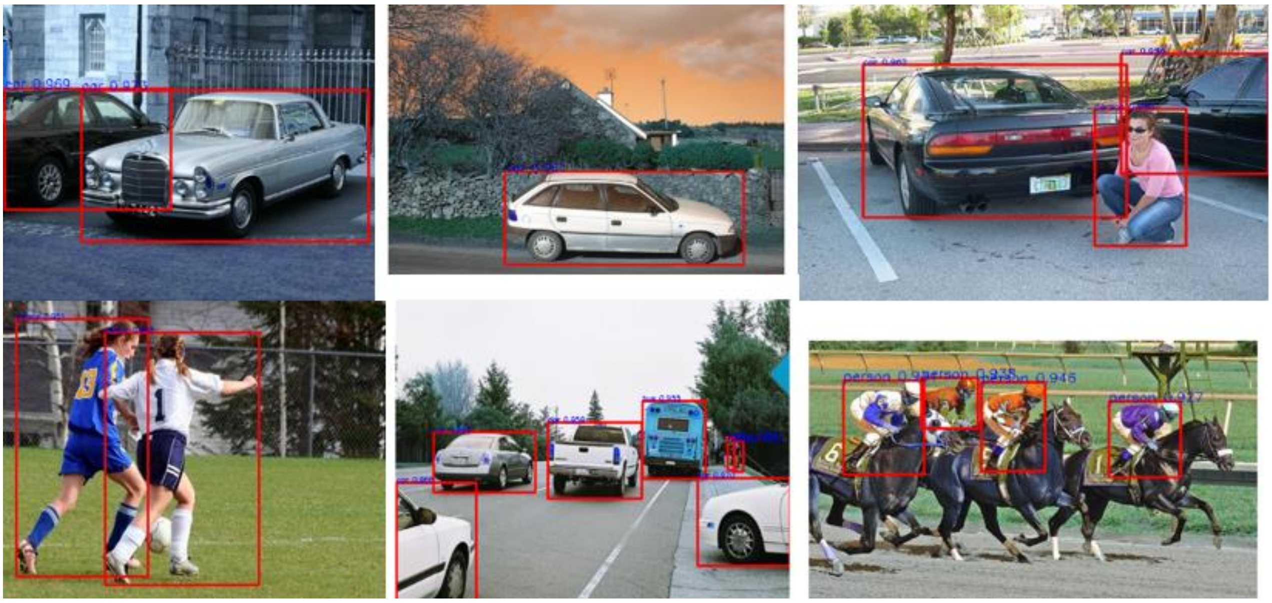


Fig. 7 Part of the detection effect on the test set

5. Conclusion

This paper proposes a pedestrian and vehicle detection network based on deep learning. In the network structure, the traditional top-down feature fusion is improved, and the two-way weighted feature fusion is proposed, so that the network can integrate high-level features in the fusion stage. Semantic information and low-level semantic information are closely integrated on each channel, and the attention mechanism algorithm is introduced to redistribute the weights of features at various scales, so that the network can automatically learn the correlation and importance between feature channels In order to further improve the detection performance of the detector, use GIOU loss and Focal loss to design the loss function of the detector. Experiments show that compared with the original YOLO V3, the algorithm pedestrian and vehicle detection accuracy has been significantly improved. The model can effectively improve the accuracy and robustness of pedestrian and vehicle detection, and the speed can reach real-time. It can be used as an important basis for unmanned driving path planning and behavior decision-making, and is of great significance to the realization of unmanned driving. There is still room for improvement in the detection speed of the model. The next step will be to do more in-depth research in compressing the amount of network parameters and further improving the detection speed. While ensuring the accuracy of the network, it also improves the response speed of the network, making it easy to use in embedded devices.

References

[1] GIRSHICK R, DONAHUE J, DARRELL T, et al. Region Based Convolutional Networks for Accurate Object Detection andSegmentation[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2016, 38(1): 142-158.

[2] GIRSHICK R. Fast R-CNN[C]// International Conference on Computer Vision. Santiago, 2015: 1440-1448.

[3] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: towards Real-time Object Detection with Region Proposal Networks[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2017, 39(6): 1137-1149.

[4] REDMON J, DIVVALA S, GIRSHICK R, et al. You Only Look Once: Unified, Real-time Object Detection[C] // IEEE ComputerVision and Pattern Recognition. Las Vegas: IEEE, 2016: 779-788.

[5] LIU W, ANGUELOV D, ERHAN D, et al. SSD: Single Shot MultiBox Detector[C]//European Conference on Computer Vision.Amsterdam, The Netherlands, 2016:21-37.

[6] Redmon J, Farhadi A. YOLOv3: an incremental improvement[EB/OL]. (2018-04-08)[2018-12-25]. https: //arxiv.org/abs/1804.02767.

[7] LIN T, DOLLAR P, GIRSHICK R, et al.Feature pyramid networks for object detection[C]// IEEE Conference on Computer Vision and Pattern Recognition. Piscataway, NJ:IEEE Press, 2017:2117-2125

[8] Lin T Y, Goyal P, Girshick R, He K, Dollár Piotr. Focal loss for dense object detection. IEEETransactions on Pattern Analysis and Machine Intelligence[J]. 2020, 42(2): 318-327.

[9] Rezatofighi H, Tsoi N, Gwak J Y, Sadeghian A, Reid, I, Savarese S. Generalized intersectionover union: a metric and a loss for bounding box regression. 2019 IEEE Conference onComputer Vision and Pattern Recognition (CVPR), June 16-20, 2019, Long BeachConvention Center, CA, USA, 2019.

[10] He K, Zhang S, Ren S, et al. Deep residual learningfor image recognition[C]// IEEE Conference on Computer Vision and Pattern Recognition. Boston,USA,2015:770-778.

[11] Bell S, Lawrence Zitnick C, Bala K, et al. Inside-outside net:Detectingobjects in context with skip pooling and recurrentneural networks[C]//Proceedings of the IEEE conference oncomputer vision and pattern recognition. 2016: 2874-2883.

[12] Hu J, Shen L, Albanie S, Sun G, Wu E. Squeeze-and-excitation networks. 2018 IEEEConference on Computer Vision and Pattern Recognition (CVPR), Salt Lake City, June 18 -22, 2018, Utah, USA. 2018.

[13] Mingxing Tan, Ruoming Pang, Quoc V. Le EfficientDet: Scalable and Efficient Object Detection[J]. arxiv, 2019(20): 1-9