Improved Yolov4-tiny Algorithm Based on Cascade Residual Dilated Fusion

Qinggang Gong
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan 430063, China
e-mail: 517422877@qq.com

Yefu Wu*
School of Computer Science and Artificial Intelligence
Wuhan University of Technology
Wuhan 430063, China
e-mail: wuyefu1988@qq.com

***Abstract*—Performance of object feature extraction and fusion is not high in the detection process of yolov4-tiny algorithm, which leads to the low accuracy. An improved algorithm, dcca-yolov4-tiny based on cascaded residual dilation fusion mechanism embedded in location information, is proposed. The dilated convolution is used to increase the receptive field, and based on the idea of residual, the cascade residual fusion mechanism is used for feature fusion to improve the effect. The spatial features containing position information are embedded into the dilation convolution, so that the network can adaptively learn the spatial feature information. Two fusion methods are used combined with the position information to further improve the fusion effect of different object receptive field features and improve the accuracy of object detection. Experimental results on the VOC(Visual Object Classes) test set show that the proposed cascade residual dilated fusion module can promote the algorithm to effectively improve the accuracy of object detection. The mAP can reach 80.86%, which is 4.73% higher than the original model.**

Keywords—cascaded residual dilated convolution, position information, feature fusion, receptive field

I. INTRODUCTION

With the rapid improvement of computer hardware, object detection based on deep learning has made great progress in speed and accuracy, and is widely used in Unmanned Aerial Vehicle(UAV), vehicle detection and Intelligent Transportation System(ITS).

At present, a variety of improvement schemes are proposed for the accuracy and speed of target detection. Lin et al. [1] fused the shallow features with rich spatial information and the deep features with rich semantic information, on the premise of adding a small amount of calculation, effectively fused the shallow and deep multi-scale features to improve the accuracy of object detection. This idea of multi-scale feature fusion is reflected in documents [2,3]. Liu et al. [4] uses the human visual mechanism and uses parallel dilated convolution to increase the receptive field to form a branch structure, which effectively improves the object detection accuracy. Fu et al. [5] used deconvolution in SSD(Single Shot multibox Detector) to fuse shallow and deep features in a top-down manner, thereby improving the accuracy of object detection. Xu et al. [6] improved SSD algorithm by using lightweight network and applied it to helmet wearing detection, which improved the object detection speed with less loss of accuracy. Gao et al. [7] effectively improved the object detection accuracy by introducing the attention mechanism into CenterNet and combined with the auxiliary detection module. Redmon et al. [8] proposed yolo(you only look once) object detection algorithm to realize real-time detection.

II. CASCADE RESIDUAL DILATED FUSION MECHANISM

Through research on the above improved scheme, a cascaded residual dilated fusion module embedded with location information is proposed to improve the detection performance of yolov4-tiny which is designed based on yolov4 [9].

A. Dilated Convolution

Dilated convolution [10] uses different dilated factors, which increases the receptive field when the convolution kernel size is the same. In formula 1, K_{m+1} is the size of the current convolution kernel, RF_m and RF_{m+1} are the receptive fields of the previous layer and the current layer respectively, and S is the product of the stride of the previous m layer.

$$\mathrm{RF}_{m+1} = \mathrm{RF}_m + (K_{m+1} - 1) \times S \qquad (1)$$

Because the convolution kernel is expanded by filling the convolution gap with zero value, the discontinuity of convolution characteristic information may be caused. As shown in Fig 1, this paper uses dilated convolution with dilated factors of 1, 2, 3 for feature superposition to reduce feature discontinuity.

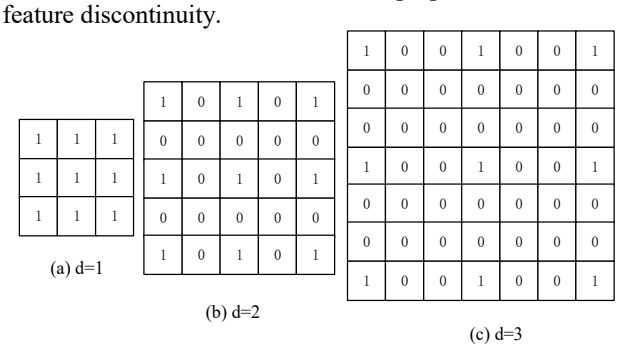


Figure 1. Dilated convolution

B. Cascade Residual Dilated Convolution

ResNet [11] uses a deeper network to extract more effective features through the idea of residuals. According

2473-3636/21/$31.00 ©2021 IEEE
DOI 10.1109/DCABES52998.2021.00041

to the idea of residuals, in order to make effective use of the original features, this paper uses residual to fuse the input features and new features, which deepens the network structure and extracts more features. The fusion module as shown in Figure 2.

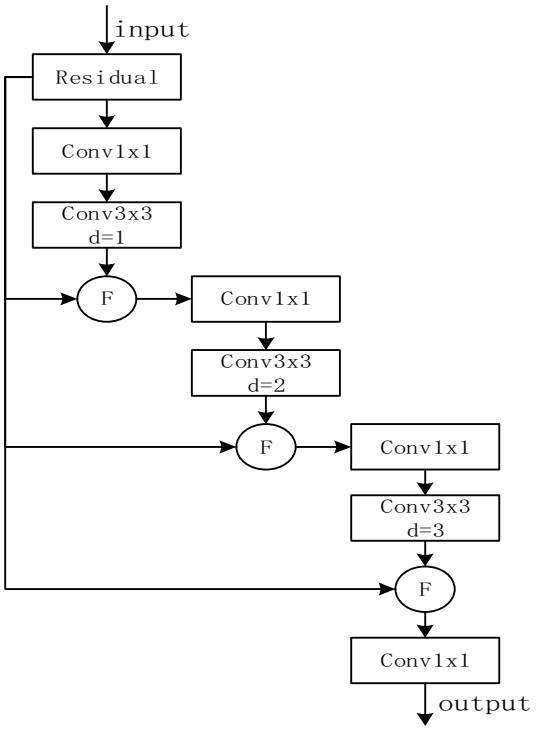


Figure 2. Cascade residual dilated fusion module

The dilated convolution with *d=1* is used to extract the local features of the small receptive field in the center of the convolution kernel. The dilated convolution with *d=2* is used to extract the surrounding features of the medium receptive field in the center of the convolution kernel. The dilated convolution with *d=3* is used to extract the long-distance features of large receptive fields in the center of the convolution kernel. In this paper, F represents add or concat for feature fusion.

C. Fusion Module with Coordinate Attention

In order to make the network pay attention to the features with accurate location information, Coordinate Attention(CA) [12] mechanism encodes the input features from the horizontal and vertical directions respectively according to formula 2 and formula 3, and decomposes the global pooling into the average pooling in two directions, so that the network can learn the location information of channel features and spatial features at the same time.

$$z_c^{h}(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h,i) \qquad (2)$$

$$z_c^{w}(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j,w) \qquad (3)$$

Based on the idea, when using cascaded dilated convolution to extract multi receptive field features, the target location information features are embedded into the hierarchical fusion module, and the network is combined with cascaded dilated convolution and location attention module to enhance the extraction of target features. A cascaded residual dilated fusion module embedded with location information is proposed as shown in Figure 3 below.

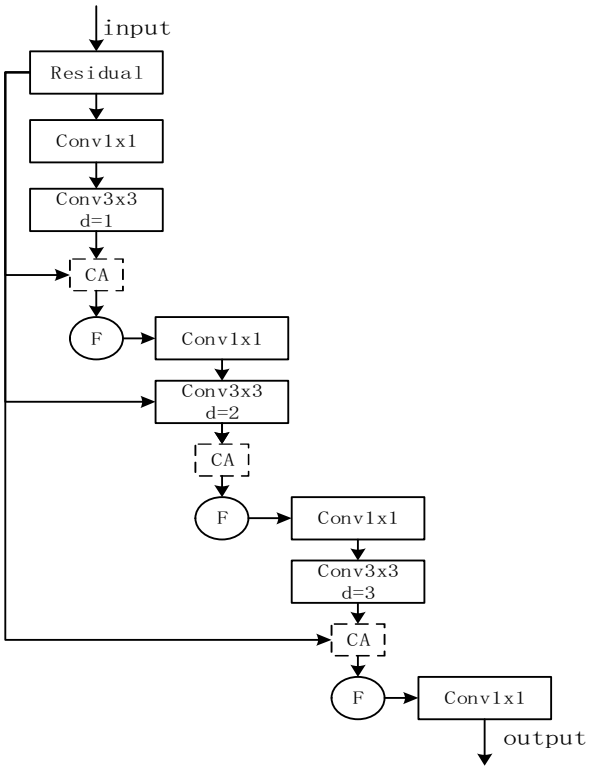


Figure 3. Cascaded residual dilated fusion module embedded with location information

III. OVERVIEW OF NETWORK STRUCTURE

Considering the detection speed, yolov4-tiny model is selected in this paper. Because it uses a shallow feature extraction network, it will lose some accuracy. In order to balance the detection accuracy and speed, the following improvements are made for the original yolov4-tiny:

(1) After the output feature of the second CSP(Cross Stage Partial) block, the proposed cascade residual dilated fusion module embedded with location information is added to improve the ability of network feature extraction by adding receptive fields;

(2) This module is also added to the second output feature of the backbone network, which aims to make the deep network pay more attention to spatial features through the module embedded with location information, combine semantic features with spatial features, and strengthen the learning of features.

So, the improved model dcca-yolov4-tiny detection structure of cascaded residual dilated fusion mechanism based on embedded location information is shown in Figure 4 below.

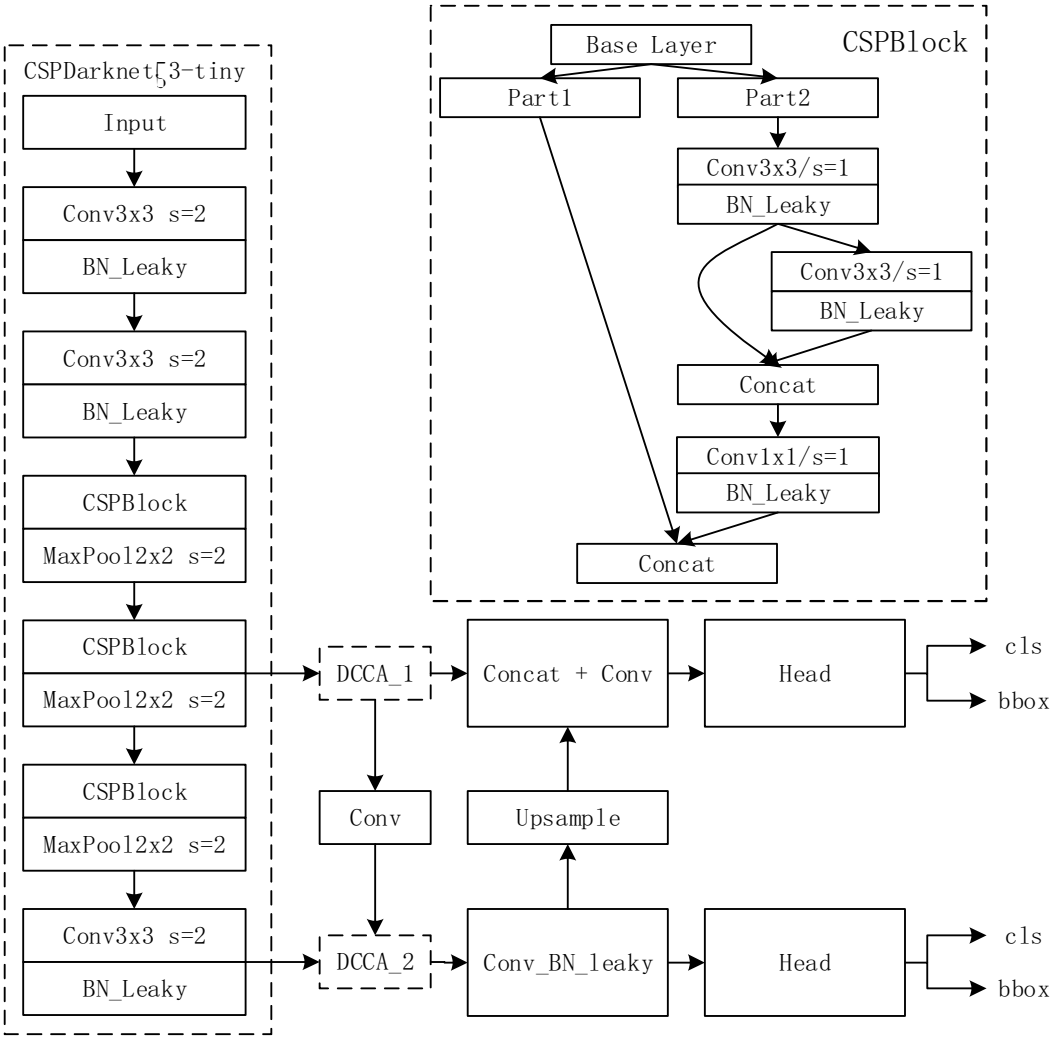


Figure 4. Structure of proposed dcca-yolov4-tiny

IV. EXPERIMENTAL RESULTS AND ANALYSIS

In this paper, experiments and evaluation are carried out on VOC data sets, including VOC2007(Visual Object Classes Challenge 2007) and VOC2012(Visual Object Classes Challenge 2012). This paper selects deep learning framework called PyTorch and uses 1080Ti GPU for training. The input picture size is 416x416, batch_size 32 is selected. This paper uses mAP(mean Average Precision) to verify the effectiveness of the model, and FPS(Frame Per Second) to evaluate the detection speed of the model.

Based on yolov4-tiny, this paper designs the following four modules: (1) the cascade residual dilated module using add fusion(dc_add); (2) the cascade residual dilated module using concat fusion(dc_concat); (3) the cascade residual dilated module embedded with location information using add fusion(dcca_add); (4) the cascade residual dilated module embedded with location information using concat fusion(dcca_concat). The results are shown in tables 1 below:

TABLE I. ORIGINAL YOLOV4-TINY AND OUR PROPOSED MODEL

	mAP(%)	FPS(frame/s)
yolov4-tiny	76.13	85
dc_add	78.00	74
dc_concat	79.03	69
dcca_add	79.12	72
dcca_concat	80.86	68

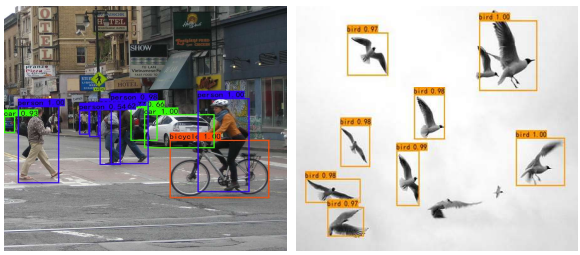
As shown in table 1, the mAP of dc_concat is 1.03% higher than that of dc_add, and the detection speed is reduced by 5FPS; the mAP of dcca_concat is 1.74% higher than that of dcca_add, and the detection speed is reduced by 4FPS. The experimental results show that the detection accuracy of the proposed module using different fusion methods is 1.87%, 2.90%, 2.99% and 4.73% higher than the original yolov4-tiny model, and the detection speed is lower than the original yolov4-tiny model, but it can still achieve the effect of real-time detection. Table 2 shows the detection results of each

category of the four different modules proposed in this paper.

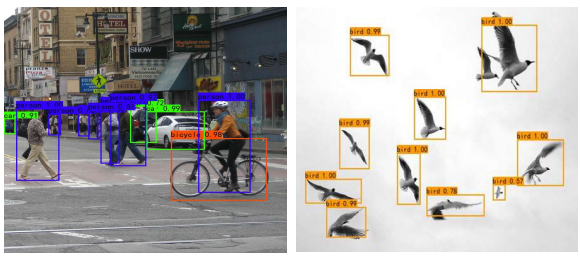
TABLE II. COMPARISON OF MAP OF THE PROPOSED MODULES

	dc_add	dcca_add	dc_concat	dcca_concat
car	91.43	**92.13**	92.21	91.97
horse	86.03	87.11	87.74	**89.25**
train	87.09	87.56	88.00	**89.02**
motorbike	87.65	85.30	86.04	**87.98**
person	86.05	85.72	85.82	**86.72**
bus	84.16	87.47	**88.70**	87.07
bicycle	85.86	85.49	85.94	**86.49**
cat	83.22	84.60	**86.00**	85.81
aeroplane	80.81	83.12	82.95	85.90
cow	82.47	**85.94**	85.86	85.89
sheep	81.28	83.15	**83.63**	83.31
tvmonitor	79.37	79.98	79.80	**81.57**
dog	76.62	79.54	79.48	**80.51**
bird	76.43	74.16	74.43	**79.21**
sofa	70.68	75.99	74.87	**78.30**
diningtable	71.94	75.38	72.65	**76.54**
boat	67.83	68.37	68.02	**71.32**
bottle	65.82	64.21	63.93	**68.43**
chair	62.54	62.98	61.50	**63.73**
pottedplant	52.90	54.22	53.02	**58.14**
mAP	**78.00**	**79.12**	**79.03**	**80.86**

In table 2, dcca_concat fusion method has the highest detection accuracy in small target boat, pottedplant and bird, which are 71.32%, 58.14% and 79.21% respectively, which is compared with dc_concat, the accuracy of small object detection can be improved by 3.30%, 5.12% and 4.78% respectively. It is proved that embedding location information into cascade residual fusion module can further improve the accuracy of small object detection.



(a) yolov4-tiny



(b) dcca-yolov4-tiny

Figure 5. Visualization of test results

Figure 5 above shows the visual results of the improved algorithm and the original yolov4-tiny model detection. It can be seen from the result on the left of the first picture that the original yolov4-tiny missed a person, but this paper uses the concat dense connection method in feature fusion, so the missed target is detected. Due to the use of dilated convolution to increase the receptive field and the way of embedding position information, it can be seen from the second picture that the improved dcca-yolov4-tniy model proposed in this paper detects the birds at the smallest lower right corner in the image.

V. CONCLUSIONS

Through the theoretical research on the attention mechanism of dilated convolution to increase receptive field and location information, a cascaded residual dilated fusion module embedded with location information is proposed in this paper, which is added to yolov4-tiny and used for object detection. The comparative experiments on VOC data sets show that the different modules proposed in this paper can improve the object detection accuracy of the original model. At the same time, the comparative experiments of feature fusion methods show that in practical application, add method can be selected from the perspective of parameters and model; In terms of feature reuse and accuracy, concat can be selected for fusion.

ACKNOWLEDGMENT

This work is supported by Hubei Provincial Key Laboratory of Transportation Internet of Things Technology under Grant 2017III028-002 and the Special Fund for Basic Scientific Research Business Expenses of Central Universities under Grant 2019III137CG.

REFERENCES

[1] Lin T, Dollar P, Girshick R, et al, Feature Pyramid Networks for Object Detection, Computer vision and pattern recognition,2017.

[2] Tan M, Pang R, Le Q V J a P A, Efficientdet:Scalable and efficient object detection, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[3] Liu S, Qi L, Qin H, et al, Path Aggregation Network for Instance Segmentation, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[4] LIU S, HUANG D, Receptive field block net for accurate and fast object detection, Proceedings of the European Conference on Computer Vision (ECCV), 2018.

[5] FU C Y, LIU W, RANGA A, et al, DSSD:deconvolutional single shot detector, https://arxiv.org/abs/1701.06659.

[6] XU X F, ZHAO W F, ZOU H Q, et al, Detection algorithm of safety helmet wear based on MobileNet-SSD, Computer Engineering, vol. 47, pp. 298-305, 2021.

[7] GAO Q Q, HUANG B C, LIU W Z, et al, Detection method of bamboo strip surface defects based on improved CenterNet, Journal of Computer Applications, vol. 41, pp. 1933-1938, 2021.

[8] REDMON J, DIVVALA S, GIRSHICK R, et al, You only look once:unified, realtime object detection, IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[9] Bochkovskiy A, Wang C Y, Liao H Y M, YOLOv4: Optimal Speed and Accuracy of Object Detection, https://arxiv.org/abs/2004.10934.

[10] YU F, KOLTUN V, FUNKHOUSER T, Dilated residual networks, Proceedings of 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] HE K M, Deep residual learning for image recognition, Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR), 2016.

[12] HOU Q B, ZHOU D Q, FENG J S, Coordinate Attention for Efficient Mobile Network Design, https://arxiv.org/abs/2103.02907.