IMG-CenterNet: An optimized algorithm based on CenterNet for pedestrian detection

Muxian Li[1], Huayong Ge[1], Huixuan Wang[1]
1. School of information science and technology, Donghua University, Shanghai, China
974073492@qq.com, gehuayong@dhu.edu.cn, w18375357956@163.com
Corresponding Author: Huayong Ge Email: gehuayong@dhu.edu.cn

2022 IEEE 6th Information Technology and Mechatronics Engineering Conference (ITOEC) | 978-1-6654-3185-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ITOEC53115.2022.9734594

Abstract—With the development of deep learning in the field of target detection, there are more and more types of algorithms with higher accuracy. Different from the traditional anchor_based algorithm, CenterNet proposed in recent years is an anchor_based algorithm which has the advantages of high accuracy, simple network and fast detection speed. Although CenterNet is already light and its accuracy can be maintained at a relatively high degree, there is still a space for further improvement. In order to get a better detection effect, IBN_Net normalization method and group convolution are introduced into the residual module of CenterNet to improve the accuracy of network detection and reduce parameters. In addition, Mosaic data enhancement method is also used to optimize the training mode of the algorithm to enrich the detection background and optimize the detection performance. Compared with the original network, the MAP has been improved by nearly 3% and the number of parameters is also 30% lower.

Keywords—CenterNet; anchor free; mosaic; IBN_Net; pedestrian detection

I. INTRODUCTION

With the improvement of hardware level and the continuous development of computer technology, target detection based on deep learning has been applied more and more widely in our life. Among them, Pedestrian Detection is a popular application in this field, which is generally combined with Pedestrian tracking, Pedestrian re-recognition and other technologies, and is widely used in security, retail and other fields. At present, the mainstream target detection algorithms are anchor-based, such as RCNN[1], Fast RCNN[2], Faster RCNN[3], SSD[4], YOLOv2[5], YOLOV3[6], YOLOV4[7], etc. Such anchory-based algorithms can obtain higher accuracy and extract richer features.So,this kind of algorithm can usually reach the level of SOTA.However, its network structure is complex. The number of parameters is large and the detection process is cumbersome. For example, the appropriate size of the prior box and indispensable post-processing such as non-maximum suppression are required. In contrast, the algorithm based on anchors-free has a lightweight network and fast detection speed which may be more friendly for industrial applications.But its detection accuracy is often not satisfactory. In order to solve such a problem, we choose to improve CenterNet[8] which is a classic anchor-free target detection algorithm through some methods. The improved CenterNet not only provides greater accuracy, but also further reduces the number of parameters.

Although CenterNet is already light and its accuracy can be maintained at a relatively high degree, there is still a space for further improvement. In order to get a better detection effect, we have done some work on the network structure and training mode and used the improved CenterNet for pedestrian detection.

This paper presents an improved CenterNet that can improve the accuracy of pedestrian detection and greatly reduce the number of parameters. The improvement of this paper in CenterNet is mainly manifested in the following aspects: (1) The original BN layer is replaced by IBN[9] construction. (2) Mosaic data enhancement method is introduced into algorithm. (3) Group Convolution is added into CenterNet to further reduce parameters and lighten the network.

In this paper, we firstly introduce the related work in section II; Then we describe three improvements of the original CenterNet algorithm in section III; The process and result of the experiment is shown in section IV; section V is the conclusion and some shortcomings found at the present stage as well as the prospect of future work.

II. RELATED WORK

A. The Network Structure Of Centernet

CenterNet is relatively simple and it can be divided into three parts: "backbone", "up-sampling" and "head". The functions of these three parts are to carry out preliminary feature extraction, obtain high resolution feature map and obtain prediction results respectively. The network structure of CenterNet is shown in Figure 1.

In actual 2D target detection, Resnet50 is usually used as backbone. Resnet50 consists of two residual network structures named Conv Block and Identity Block. The former requires a convolution of 1×1 to change the number of channels in order to concatenate. The function of Conv Block is to change the dimension of the network. Identity Block can be directly concatenated and used to deepen the network. We can get a 16×16×2048 effective feature layer for further feature processing when the input image is 512×512×3.

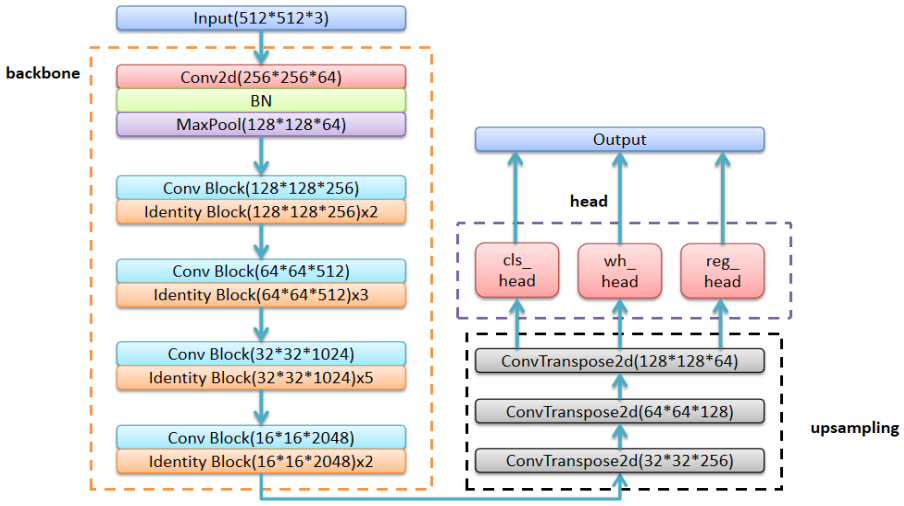

Fig. 1. The network of CenterNet

The up-sampling module of CenterNet actually uses Transposed Convolution to expand the feature layer of 16×16×2048 in resolution and reduce the number of channels. The height and width of the feature layer will be doubled after a Transposed Convolution. So, we can obtain an effective feature layer (high resolution feature map) of 128×128×64 after up-sampling. Then, this effective feature layer will be delivered to the head part for final prediction result processing.

The "head" part can be split into three branches including "cls_head", "wh_head" and "reg_head". The former is responsible for predicting the type and confidence of the object detected corresponding to each point in headmap; The "wh_head" regress to the width and height odetection box; The "reg_head" is used to predict the deviation of point in headmap along the X and Y axes.

B. IBN_Net

Batch Normalization(BN) and Instance Normalization(IN) are both methods of Normalizing the data. The difference is that the mean and standard deviation of BN are calculated in a channel. The width, height and batch are averaged by BN. But IN fixed channel and batch at the same time.It averaged width and height. The research about them shows that IN is invariable to the changes of object appearance, such as illumination, color, style, virtual and real. While BN can save the information related to content. The adaptability of the model to image appearance changes can be improved while the learning and generalization ability can also be maintained if BN is combined with IN according to certain principles.

C. Group Convolution

The idea of Group Convolution first appeared in AlexNet. At that time, it was to solve the problem that AlexNet could not be trained in a GPU due to too many training parameters. At that time, the feature map was divided into multiple GPUs for processing and the processing results of multiple GPUs were concatenated to avoid loading too many parameters at one time. This idea is actually Group Convolution.

Now the application of Group Convolution[10] is mainly to divide the input feature layer into several groups along the depth direction, and each group is composed of C1/ G channels (C1 is the original number of channels and G is the number of groups). Similarly, the output channel is similarly split to get g groups. The grouping of inputs corresponds to the grouping of outputs and is computed separately using convolution. Instead of generating only one output feature layer, we can now get glayers with the same number of parameters. In other words, the number of parameters required is reduced by g times! Specific application process is analyzed in Section III in conjunction with the improved CenterNet network.

D. Mosaic Data Augmentation

Mosaic is a image data enhancement technology proposed in YOLOV4 in 2020. Different from the general way of data enhancement is to perform some operations on one image such as: flip, twist, gamut change.Mosaic data enhancement takes advantage of four images which are pieced together to obtain a new image. And then the neural network can be trained by using this new image. This is equivalent to calculating the data of four images during training and normalization which enriching the background of the detected object greatly! Some argue that the biggest reason for YOLOV4's performance improvement is the use of Mosaic.

III. THE IMPROVMENT OF CENTERNET ALGORITHM

A. The change of the normalization processing method

As a method of Normalization,Batch Normalization (BN) which preserves semantic information about features is used in original CenterNet network. However, relevant studies show that in the shallow layer of the network, the feature divergence of data sets with the same content and different styles is much larger than that of

semantic information. It is only with the deepening of the network that the characteristic divergence declines continuously. This means that low-level features reflect cosmetic differences, while high-level features reflect semantic information. In order to improve the model's adaptability on appearance differences, Instance Normalization(IN) is introduced into CenterNet. The introduction of IN needs to follow two principles .Firstly, IN should be put at the lower level.Because this can preserve the high-level semantic information while the low-level filtering reflects the appearance information. Secondly, half of BN layer of the original low-level layer is retained while the other half features are replaced by IN. This preserves the content of the image at low-level layer.Figure 2 shows the structure of residual block which introduce IN into network.

As for residual Block, in addition to two principles, there are three reasons for the optimization of BN layer: (1) Original ResNet had proved that it is much better to optimize ResNet without adding layers to Identity Block. (2) In order to ensure consistency with Identity Block, IN is placed in the first Normalization layer. (3)It can optimize the appearance features and retain the semantic information features without increasing any computation with half IN and half BN.

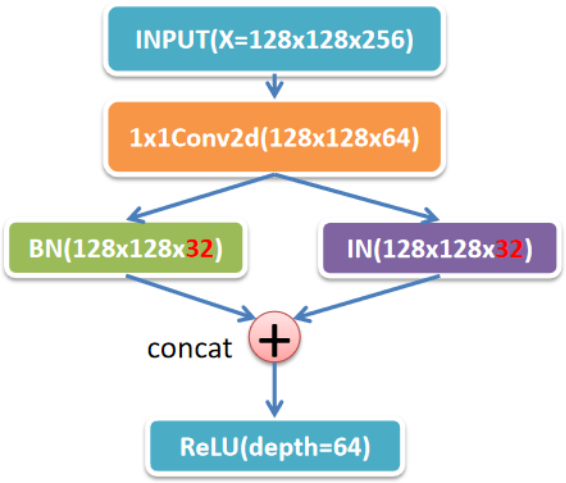

Fig. 2. The residual block introduced IN

B. *Mosaic data enhancement was introduced*

Mosaic data enhancement method was introduced into the training algorithm in order to increase CenterNet's ability to detect small targets and enrich the background information of detected objects. As for the number of image splicing, four images are appropriate. Because the Mosaic effect of two images is not so obvious and six or even eight images will make the originally small target become smaller which impacts the generalization ability of the model . In addition, it is worth mentioning that the data processing method of directly calculating 4 pictures does not require a large batch size. So a GPU can achieve a better effect which reduce the training pressure of the network. The effect is shown in figure 3.

Specific implementation process is roughly divided into four steps: (1) Reading four images randomly from the data set every time.(2) Those four images are flipped (the original image is flipped left and right), zoomed (scaling original images), gamut changed (brightness, saturation and tone change) and other operations. Placing four pictures in the position of top left, bottom left, bottom right and top right in order when those operation are completed. (3) Using matrix approach to intercept down fixed area of 4 images and forming a new image by putting them together. The new image contains the coordinates of the original real box that have been recalculated (mainly zooming and moving). (4) Sometimes a part of image that be placed first will be covered by the image that be placed later during the process of stitching images.This part may contain an object marked by a real box. So we need to marginalize the real box in this case. Similarly, part of the intercepted original image may be lost and we also need this part for edge filling.Finally, this new picture can be put into mini-batch for training.

Fig. 3. Using Mosaic to process images

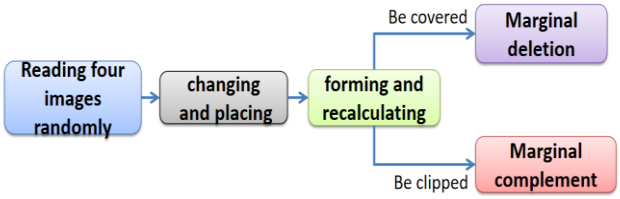

Fig. 4. The flowchart of Mosaic

C. *Parameters are reduced by Group Convolution*

Backbone which is composed of residual network blocks has the main number of parameters in the CenterNet. Therefore, in order to further reduce the number of parameters in entire network, we choose to replace a part of normal 2D image convolution in the residual module that forms backbone with grouping convolution in this improvement of CenterNet.As shown in the figure 4,the second Conv2d is replaced.The reason why we choose the second Conv2d is that the size of this convolution kernel is 3 × 3 which will make the effect more obvious.

In the original 2D convolution, the input feature layer and the Filters are two independent wholes. When the size of the input feature map is $H\times W\times C_1$ and the size of the convolution kernel is $H_1\times w_1\times C_1$, a feature layer with the size of $H\times W\times C_2$ is obtained. At this time, the number of parameters to complete this process is $H\times w_1\times c_1\times c_2$. Replacing Group Convolution is actually to divide the input feature layer and the Filters into G groups (g=8 in this improvement). After replacement, the size of each input feature layer is $H\times W\times (C_1/g)$ and the convolution kernel is $H_1\times w_1\times (C_1)$.The corresponding size of each output feature map will change to $H\times W\times (C_2/g)$.Then all of those output feature map will be concatenated to obtain the same feature layer of $H\times W\times C_2$ as before. The number of parameters to complete this process is $h_1\times w_1\times (C_1/g)\times(C_2/g)\times g=h_1\times w_1\times C_1\times C_2\times(1/g)$. It is obvious that parameters were reduced by g times compared with the previous Conv2d.

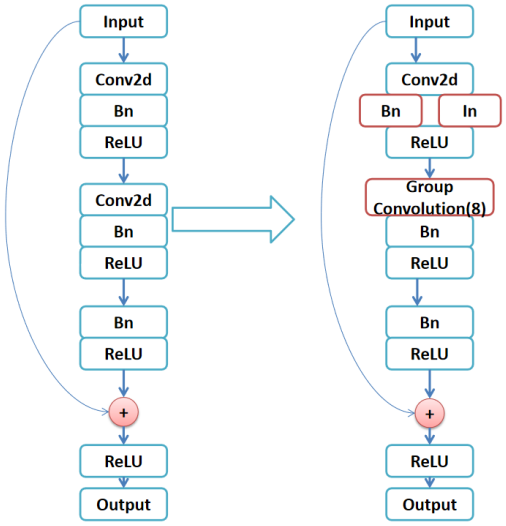

Fig. 5. Position of Group Convolution on the network

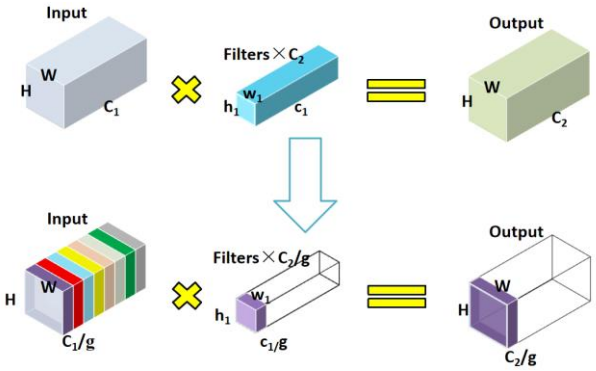

Fig. 6. Reducing parameters by grouping

In addition, it is worth mentioning that adopting Group Convolution can not only optimize the number of parameters, but also increase the diagonal correlation between filters which is similar to the effect of regularization. So that the network is not easy to over-fit.

IV. Experimental Results And Analysis

A. DataSet

INRIA Person Data Set is the most widely used static Person detection data set and its original pictures are from the GRAZ 01 data set and some pictures on the network. Human posture and all kinds of light conditions are relatively comprehensive which make this data set is suitable for pedestrian detection. Each image also demarcates the pedestrian area and records the length and width of the upper left fixed point of the rectangular box.The training set contains 614 images and test set has 288 images.So we use 902 images in this training set. In this experiment, we expand the number to 1,495 using data enhancement.

B. Experiment Settings

The experiment settings of CenterNet and IMG-CenterNet are both on same condition. For this experiment we used an NVIDIA GeForce RTX 2080TI graphics card with 11GB of memory. Framework that we choose here is Pytorch.

Specifically, we first trained 20 times on the INRIA dataset using CenterNet with VOC2007 pre-training weights. The results of the 20th session were then used as the new pre-training weights In the end, the original CenterNet and the IMG-CenterNet were trained 100 times (It was found that the fit was usually completed in about 50 times). We took the best performing data of the epoch 100 times and compared them. Since CenterNet is the Anchor-free target detection algorithm, it does't need Anchor or final non-maximum suppression processing.

C. Results

The test results of the original CenterNet algorithm and the improved IMG-CenterNet algorithm under the INRIA data set are shown in Table I:

TABLE I. Detection Results

Algorithm	recall	Precision	Model size/MB
CenterNet	69.01%	93.21%	124.8
IMG-CenterNet	**77.72%**	**97.27%**	**87.1**

It can be seen that the improved algorithm has a relatively high degree of improvement in terms of accuracy and recall rate. Recall increased from 69.01% to 77.72% which greatly reduced the missed detection rate. Precision also improved by four percentage points from 93.21% to 97.27%. This is mainly due to the introduction of IBN normalization which improves the model's adaptability to image appearance changes and Mosaic Data Augmentation which greatly enriched the background information of detected objects. On the other hand, the number of model parameters decreased by 30%, from 124.8MB to 87.1MB. The detection comparison

results of the original CenterNet algorithm and the improved algorithm are shown in Figure 7.

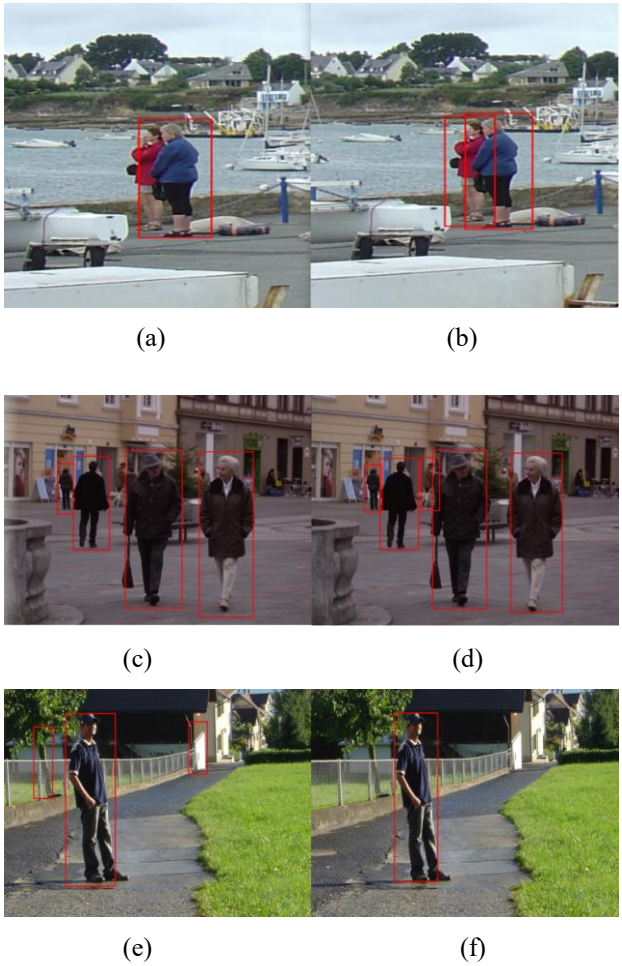

Fig. 7. Comparison of test results

The comparison images above are taken from the test data set. Images (a), (b) and (c) in the left column show the original CenterNet test results.Images (d), (e) and (f) are the results of IMG-CenterNet. It can be found that in the comparison of (a) and (b), the improved algorithm can detect the effective objects that could not be detected originally. Moreover, in the group of images (c) and (d), the detection of small targets is more accurate, which is more attributed to the use of Mosaic data enhancement in the algorithm. On the other hand, it can be seen from the comparison of (e) and (f) that the improved algorithm can also reduce the error detection rate and improve the accuracy rate.The original algorithm mistakenly targets backgrounds that do not belong to humans.

D. Ablation Experiment

In order to observe the specific effects of IBN, Mosaic, and Group Convolution on IMG-CenterNet algorithm, further elaborate experiments are carried out. In this engraved experiments, not only are the map of the algorithm considered, but also the number of parameters are included in the comparison. The results are shown in Table II:

TABLE II. ABLATION EXPERIMENTAL RESULTS

Algorithm	**CenterNet**	-	-	**IMG-CenterNet**
IBN		√	√	√
Mosaic			√	√
Group Convolution				√
mAP(%)	91.42	92.38	**94.65**	94.24
Model size/MB	124.8	124.8	124.8	**87.1**

It can be seen that the original CenterNet Model size is only 124.88MB. CenterNet is already a fairly lightweight algorithm compared to other anchor-base target detection algorithms. This also means that the network is relatively refined, with fewer redundant parameters. On the other hand, CenterNet's MAP performance is 91.42% which is not very good. The results of ablation experiments show that the accuracy of the algorithm is well optimized by adding IBN and Mosaic and MAP is improved from 91.42% to 94.65% without increasing the number of parameters in the model. In addition, Group Convolution introduced to further reduce the number of parameters also achieves good results. Substituting Group Convolution for some ordinary 2D convolution operations in the residual module reduces the number of parameters by 30.47% while only 0.41% map is sacrificed.

V. CONCLUSION

In this paper, the proposed IMG-CenterNet improves the detection accuracy and reduces the model parameters by combining the IBN normalization method and Group Convolution. The former can improve the model's comfort on appearance differences. The latter greatly reduces the parameters of the network by grouping input feature layer and convolution kernel along the depth direction.In addition, Mosaic data enhancement algorithm is introduced to enrich the background information of the detected objects which can further improve the accuracy of network training. With the combination of the above improvements, The detection accuracy and number of model parameters of IMG-CenterNet are both better than that of CenterNet. However, there are still two problems in the image detected through the experimental results. Firstly, CenterNet is not good at distinguishing between dummies and real person.secondly, Those people whose center are overlapped cannot be detected. We'll study how to address these issues and further improve performance in future work.

REFERENCES

[1] R. Girshick, J. Donahue, T. Darrell, and J. J. I. C. S. Malik, "Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation," 2013.

[2] R. J. a. e.-p. Girshick, "Fast R-CNN," 2015.

[3] S. Ren, K. He, R. Girshick, J. J. I. T. o. P. A. Sun, and M. Intelligence, "Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks," vol. 39, no. 6, pp. 1137-1149, 2017.

[4] W. Liu et al., "SSD: Single Shot MultiBox Detector," 2016.

[5] J. Redmon, A. J. I. C. o. C. V. Farhadi, and P. Recognition, "YOLO9000: Better, Faster, Stronger," pp. 6517-6525, 2017.

[6] J. Redmon and A. J. a. e.-p. Farhadi, "YOLOv3: An Incremental Improvement," 2018.

[7] A. Bochkovskiy, C. Y. Wang, and H. Liao, "YOLOv4: Optimal Speed and Accuracy of Object Detection," 2020.

[8] X. Zhou, D. Wang, and P. Krhenbühl, "Objects as Points," 2019.

[9] X. Pan, P. Luo, J. Shi, and X. Tang, "Two at once: Enhancing learning and generalization capacities via ibn-net," in Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 464-479.

[10] Y. Ioannou, D. Robertson, R. Cipolla, and A. Criminisi, "Deep Roots: Improving CNN Efficiency with Hierarchical Filter Groups," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.