Journal of Computing and Electronic Information Management
ISSN: 2413-1660 | Vol. 12, No. 1, 2024

Mask wearing detection algorithm based on improved Yolov7

Xu Zhou, Guojun Lin*

School of Automation and information engineering, Sichuan University of Technology, Zigong 643000, China
* **Corresponding author**: Guojun Lin (Email: 386988463@qq.com)

Abstract: Manual inspection of the mask is too time-consuming and laborious. In order to detect whether a mask is worn in a crowded public place, a mask-wearing detection method based on improved YOLOV7 is proposed, which uses Depth wise separable convolution instead of conventional convolution, in order to integrate the local feature information and the whole image information deeply, Dilated Convolution was used to improve the Pyramid Pooling Module (DC-PPM) , at last, the loss function of target location is optimized, which makes it not only have the ability of feature extraction to fuse the whole and local information, but also have the ability of not losing the detail information. The experimental results show that the detection accuracy and speed of the algorithm are 95.07% and 79 frames/s respectively, which are 3.4% and 14 frames/s higher than the original YOLOV7 algorithm, very good to meet the actual application needs.

Keywords: YOLOV7 algorithm; Depth separable convolution; Dilated convolution; Loss function; Target detection.

1. Introduction

It has become a daily routine to wear masks in public places such as subways, train stations and airports. It is an important research topic to develop an efficient and easy-to-deploy mask detection algorithm in recent years, because it is too time-consuming and laborious to carry out artificial detection in public places.

At present, there are two methods of target detection based on depth learning: one is one stage target detection algorithm based on regression problem, which transforms the problem of target boundary location into a regression problem without generating candidate boxes, direct regression to the predicted target. The classical algorithms are SSD and Yolo [3-7]. The other is a two-stage target detection algorithm based on candidate regions, which generates a series of candidate boxes from the samples, classifies the samples according to the convolutional neural network, and finally determines the location of the boundary boxes, representative works include R-cnn and other detection algorithms. On this basis, there are one-stage and two-stage improvements, such as Faster R-CNN, Yolo, SSD and other updates of classic papers, including Faster R-CNN, DSSD algorithm and so on. Because of the difference between one stage and two stages, the former is superior to the latter in detection accuracy and positioning accuracy, while the latter is superior to the former in detection speed.

In the face mask detection algorithm, Bo Jingwen et al. based on Yolov3, ShuffleNetv was used as the backbone network, SKNet attention mechanism was introduced, and CIOU was used as the bounding box regression loss function. The experimental results show that the accuracy of mask detection is 93.38% and the detection speed is 59.09 frames/s. Zhang Lieping based on Yolov2, MobileNetV2, a feature extraction network using parameter migration learning, also has better detection effect under the condition of insufficient light and dense detection, the experimental results show that the average prediction accuracy of the proposed algorithm is 91.3% and the detection speed is 22 frames/s. Item Fusion, etc. based on Yolov5s, introduces self-attention mechanism, and adopts feature fusion target detection technology based on two scales. Experimental results show that the average accuracy of the proposed algorithm reaches 94.7% . The network structure of the above algorithm is complex, the detection speed is slow, and the ability of target detection and recognition in complex background is poor.

In practical applications, it is necessary to use some mobile devices and embedded devices with low memory consumption to detect the real-time mask, but the real-time parameters of the model and the amount of calculation are large, and the real detection task is more complex. In order to solve this problem, this research improves the YOLOV7 model. Yolov7 uses faster convolutional operations and smaller models in the YOLO series to achieve higher detection accuracy with the same computational resources. In addition, YOLOV7 provides higher accuracy, ability to detect more fine-grained objects. Compared with Yolov5, although Yolov5 is a lighter model with less parameters and computation, Yolov7 is suitable for more complex target detection tasks and outperforms YOLOV5 in both detection accuracy and speed.

In this study, the YOLOV7 model is improved by using depth-separable convolution to replace the conventional convolution layer in the YOLOV7 model, thus reducing the computation and parameters of the model, eiou loss is used as the target location loss function to further improve the location accuracy and robustness, and to improve the accuracy and robustness of target location, and have more accurate loss value.

2. YOLOV7 algorithm framework

The structure of Yolov7 deep learning target detection model is shown in Figure 1. Yolov7 is divided into three parts: Backbone Feature extraction network (Backbone) , FPN (Feature Pyramid Networks) and Yolo Head (classifier and regression) . The work of the whole Yolov7 network is feature extraction-feature enhancement-prediction of the objects corresponding to a priori box.

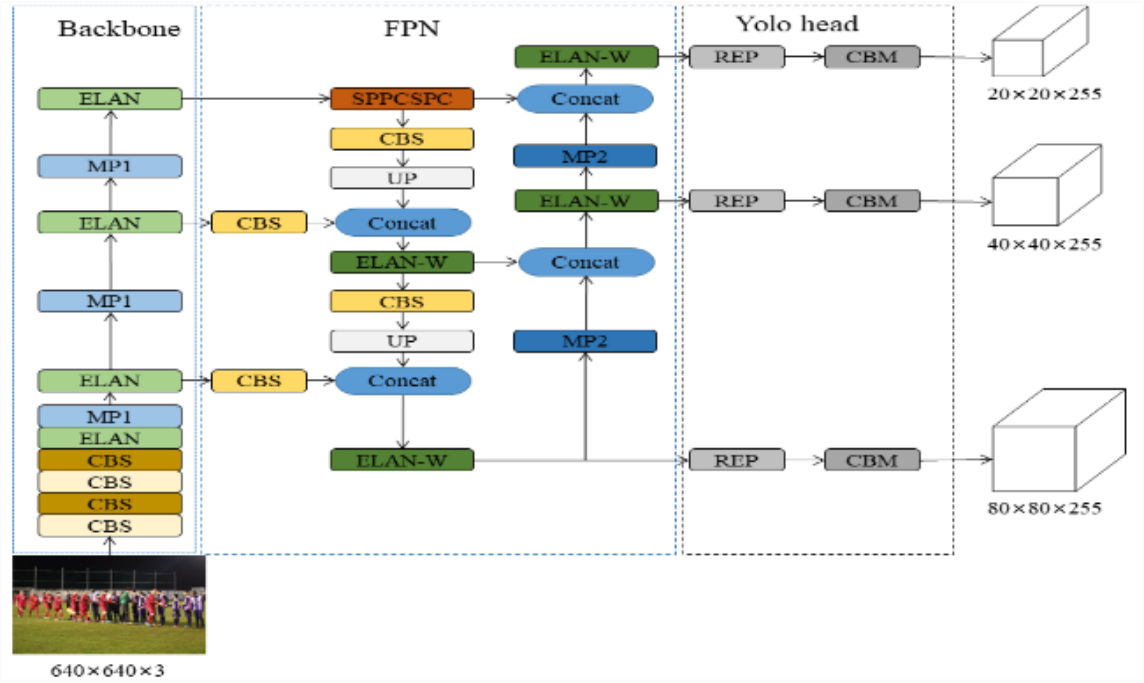


Fig. 1 YoLov7 structure of object detection model for deep learning

3. Improved YOLOV7 algorithm

3.1. Depth separable convolution

In order to detect whether a pedestrian is wearing a mask, a deep learning target detection algorithm is usually deployed to an edge terminal for real-time detection of surveillance video or images, then the detected information file back to the management room to carry out whether to wear a mask monitoring. Because of the high cost of edge terminal, and can only deploy the model with few parameters, fast reasoning speed, light-weight model, so the memory and real-time of the model are required. Even the most lightweight I model in the Yolov7 Target Detection Model series uses many traditional convolution layers, and many CBS traditional convolution layers are used in Elan and ELAN-w modules, the model parameters and computational load are large, which is unfavorable for the model to be deployed to the edge terminal. In this paper, the depth-separable convolution is used to replace the traditional convolution layers in CBS, Elan, MP1, MP2, ELAN-w and SPPCSPC modules to reduce the model parameters and computation. The improved modules are named DS-CBS, DS-ELAN, DS-MP1, DS-MP2, DS-ELAN-W and DS-SPPCSPC modules.

In order to reduce the parameter and computation, the depth separable Convolution is used to extend the feature information by using 1 × 1 Pointwise Convolution (PW, Pointwise Convolution) to increase the dimension of the input feature matrix, then, 3 × 3 DW (Depth wise Convolution) is used to extract the depth feature, and 1 × 1 PW Convolution is used to reduce the depth feature. The depth-separable convolution operation flow is shown in Figure 2, Figure 3, and Figure 4. The parameters are shown in formulas (1), (2), and the computations are shown in formulas (3), (4)

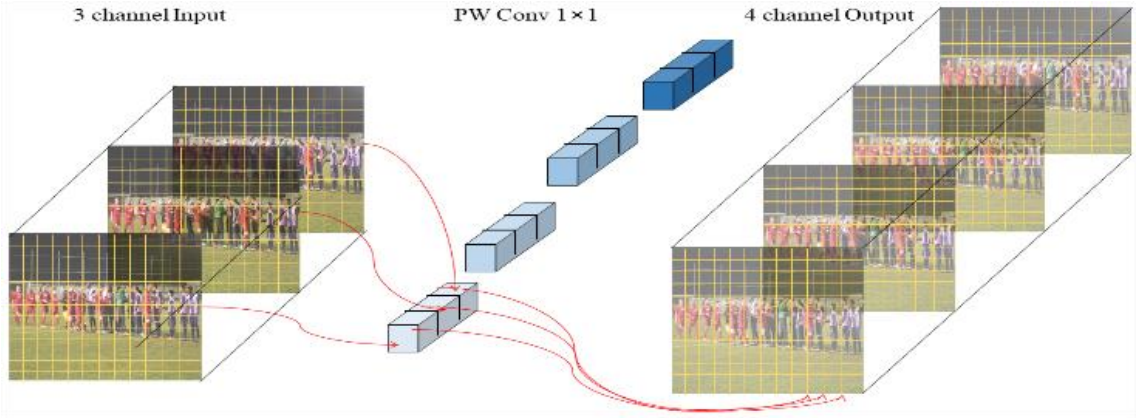


Fig. 2 PW convolution

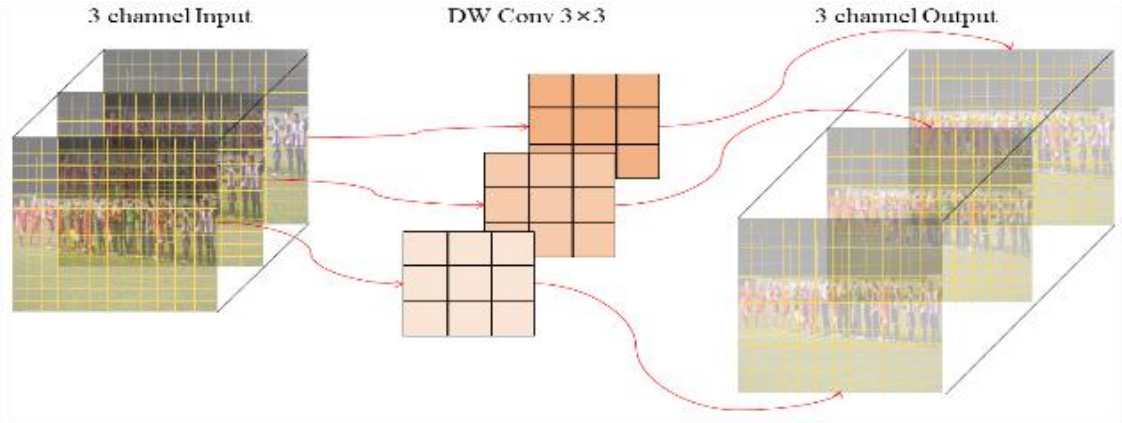


Fig. 3 DW convolution

$$n_p = e_c \times o_c \quad (1)$$

$$n_d = c_w \times c_h \times e_c \quad (2)$$

$$f_p = e_c \times o_c \times f_w \times f_h \quad (3)$$

$$f_d = c_w \times c_h \times e_c \times f_w \times f_h \quad (4)$$

In the formula, the number of parameters in PW and DW Convolution, the number of channels of input and output characteristic matrix, the width and height of convolution core, the number of calculations in PW and DW Convolution, the width and height of the input feature matrix. Taking an input feature matrix of 320 × 320 × 512, the size of the convolution kernel is 3 × 3, and the number of channels of the output feature matrix is 1024 as an example, the required parameters of the convolution kernel are: = 512 × 1024 + 3 × 3 × 512 + 512 × 1024 = 1053184, compared with the traditional convolution, it can reduce the parameter quantity by about 4.5 times. The number of times required is: = 512 × 1024 × 320 × 320 + 3 × 512 × 320 + 512 × 1024 × 320 × 320 = 107,846,041,600 times, which can save 4.5 times compared with the traditional convolution.

3.2. Improvement based on DC-PPM module

The purpose of this study is to detect the wearing of masks in crowded public places, but some of the targets to be detected are seriously occluded each other and the background is changeable. In order to integrate the local feature information and the whole image information deeply, this paper improves the YOLOV7 model by using DC-PPM module, ability to determine whether to wear a mask.

Due to the presence of pool layer and Bilinear interpolation upper sampling layer, it is inevitable that a large amount of feature information will be lost. In this paper, the convolutional nuclei are 3 × 3,5 × 5,7 × 7, the dilation rate D is 2, and the filling pixel p is 2,4, and 6 cavities with the same receptive field, respectively, the PPM module is improved by replacing the pooled cores of 5 × 5,9 × 9 and 13 × 13, respectively, so that it has the ability of feature extraction that fuses the whole and local information without losing the detail information, the improved PPM is dc-PPM. The structure of PPM module is shown in Figure 5, and the calculation formula of cavity convolution receptive field is shown in Formula (5).

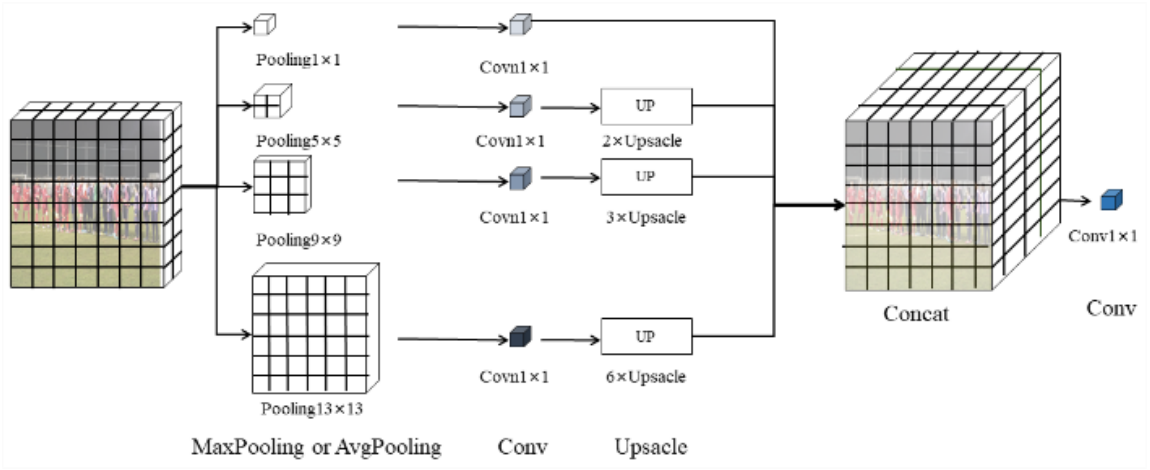


Fig. 4 PPM module flow

$$f\_w = (k-1)*d+1 \quad (5)$$

In formula (5), K is the receptive field of cavity convolution, K is the number of convolution kernel width or height direction parameter, d is the cavity rate parameter.

3.3. Improvement based on DC-PPM module

The loss function of Yolov7 is divided into three parts: target location loss Lloc, target confidence loss LCONF and target classification loss LCLA.

In this paper, EIOU loss is used as the target location loss function of Yolov7 instead of the original mean square error loss.

The goal of EIOU loss is to optimize the positioning accuracy of the bounding box. Compared with the traditional IOU loss, EIOU loss considers the center distance and the ratio of length to width of the bounding box, which can evaluate the accuracy of the bounding box more accurately and improve the precision of target location.

EIOU loss is more robust to changes in bounding box dimensions and aspect ratios. The traditional IOU loss may have some deviations when dealing with boundary boxes of different scales and aspect ratios. However, EIoU loss introduces penalty terms for center distance and aspect ratio, can better adapt to the changes of different bounding boxes.

EIOU loss has a numerical range closer to 1 than IOU loss, providing a more accurate representation of the loss value. The traditional IOU loss values range from 0 to 1, while the EIOU loss values range from 0 to 2, which more accurately reflects the location of the bounding box.

The overall improved model structure is shown in Figure 9.

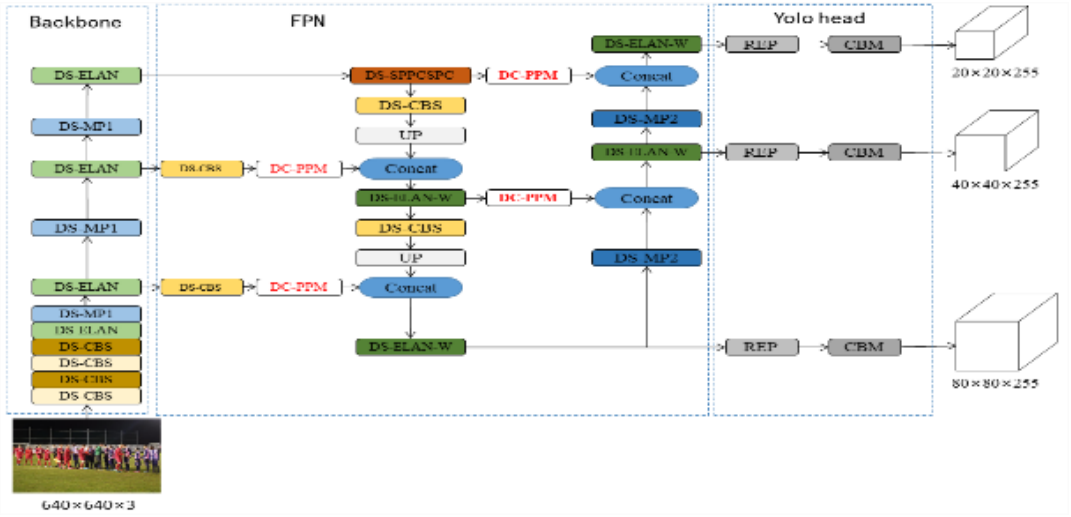


Fig. 8 Synthesizes the improved YOLOV7 network architecture

4. Experimental Analysis

4.1. Data set

The data set used in this experiment was the public MAFA (MAsked FAces) dataset and the WIDER FACE dataset, from which 1000 data sets were extracted, and the Internet, which collected 4000 samples, the object classification is marked by using LABELIMG software. A total of 5000 images of the dataset were collated and labeled as mask and nomask, with 4000 for the training set, 500 for the test set and 500 for the validation set.

Fig. 9 Sample dataset

4.2. Experimental environment

The hardware configuration of the experiment is: Intel (R) Xeon (r) CPU E5-2695 V 4@2.10 ghz 2.10 ghz (2 processors) , memory: (RAM)256GB, GPU (NVIDA Titan XP) a total of 4 blocks, graphics card memory 12GB, the computer language uses Python 3.8 and the Deep Learning Framework is Python 1.12.0.

Deep learning model hyperparameters: Total iteration cycles epoch set to 120; batch number of images set to 8; initial learning rate set to 1e-2; minimum learning rate set to 0.01 * 1e-2; Momentum is set to 0.937.

4.3. Convergence of loss function

The modified Yolov7 changes in the loss function of the training set and the validation set during training, as well as changes in the smoothing loss function of both, are shown in Figure 11. With the increase of training cycles, the loss function is stable and the model converges gradually when training about 9 Epoch.

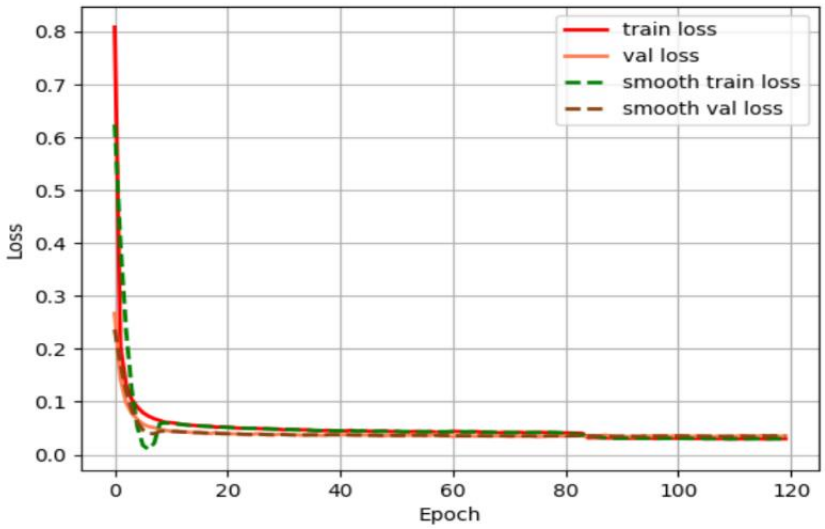


Fig. 10 convergence of loss function

4.4. Comparative experiment and analysis

In order to ensure the rigor of the experiment, the typical depth learning object detection model and the detection algorithm in this paper are tested under the condition of single variable experiment, its model performance metrics are shown in Table 1. From Table 1, the detection algorithm in this paper is higher than Faster R-CNN and SSD in mAP and FPS. Compared with Yolov3, mAP improved 11.79% , FPS improved 34 frames/s; compared with Yolov5, mAP improved 9.33% ; compared with original YOLOV7, mAP improved 3.4% , FPS improved 14 frames/s. In a word, the improved detection algorithm based on Yolov7 has more advantages in mask detection.

Table 1 comparison of target detection model accuracy and reasoning speed

Algorithm	FPS/(frames/sec)	mAP/%
Faster R-CNN	7	89.92
SSD	54	84.46
YOLOv3	45	83.28
YOLOv5	89	85.74
YOLOv7	65	91.67
Ours	79	95.07

(a) the original YOLOV7 algorithm results

(b) the detection results of the proposed algorithm

Fig. 11 Comparison of detection results

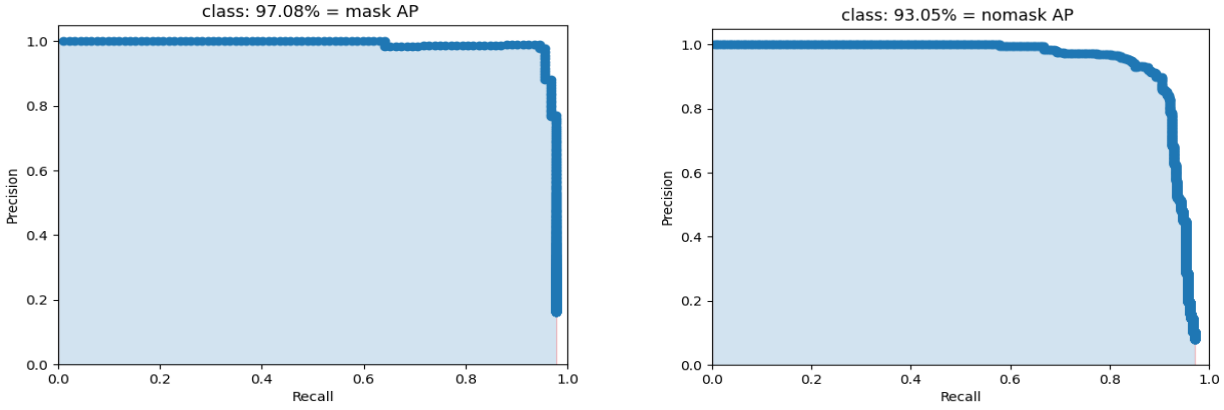
4.5. Ablation experiment and analysis

In order to verify the validity of the proposed Method, Method 1 uses depth-separable convolution to replace the traditional convolution layers in CBS, Elan, MP1, MP2, ELAN-w and SPPCSPC modules under the same experimental conditions Method 2 is based on Method 1 to improve pyramid cistern module, and Method 3 is based on Method 2 to evaluate the impact of the improved Method on the performance of detection algorithm by using target location loss as EIOU loss. The final result is shown in Table 2, figure 12.

Table 2. Improved YOLOV7 ablation experiment

Algorithm	DSC	DC-PPM	EIoU loss	Mask (AP)/%	Nomake(AP)/%	mAP/%	FPS/(frames/sec)
YOLOv7	×	×	×	91.83	91.51	91.67	65
Method 1	√	×	×	95.86	92.05	93.96	82
Method 2	√	√	×	96.78	92.65	94.72	77
Method 3	√	√	√	97.08	93.05	95.07	79

As can be seen from Table 2, using Method 1 or Method 2 alone, the mAP and FPS were significantly improved compared with the original YOLOV7, mAP was improved by 2.29% and 3.05% , respectively, fPS improved by 17 frames/s and 12 frames/s, respectively. The mAP and FPS of the improved model are 95.07% and 79 frames/s respectively. Although the FPS of the improved model is 3 frames/s smaller than that of Method 1, the mAP is improved by 1.11% . In conclusion, the proposed method has good performance on mAP and FPS, which verifies the effectiveness of the improvement.



(a) mask category AP value

(b) Nomask category AP value

Fig. 12 Yolov7 synthetic improved detection accuracy

5. Conclusion

This paper presents an effective mask-wearing detection method based on improved Yolov7. In order to reduce the parameters and computation of the model, the depth separable convolution is proposed to replace other common convolution. Secondly, in order to integrate the local feature information and the whole image information deeply and make the model have the ability to judge whether to wear the mask according to part of the face area, an improved module based on DC-PPM module is proposed Third, EIOU loss is used as the target location loss function to replace the original mean square error loss, which further improves the positioning accuracy and robustness, and has more accurate loss value.

The experimental results show that the mAP and FPS of the improved YOLOv7-based mask detection method are higher than those of the original YOLOV7, and it can perform better mask detection in crowded public places. This algorithm has good accuracy and robustness, and can meet the needs of practical applications.

Acknowledgements

This work was supported in part by the 2022 Graduate Innovation Fund Project of Sichuan University of Science and Engineering (Y2022162). The authors express their acknowledgement for the anonymous review.

References

[1] YISX, YANGZJ, ZHOU L Q, et al. Intelligent localization sampling system based on deep learning and image processing technology [J] Sensors, 2022, DOI: 10.3390/s22052021.

[2] LIU W, ANGUELOV D, ERHAN D, et al. SSD: single shot multibox detector [C]//Proceedings of the European Conference on Computer Vision, Netherlands, Oct 10-162016 Berlin, Heidelberg: Springer Verlag, 2016:21-37.

[3] REDMON J, DIVVALA S, GIRSHICK R, et al. You only look once: unified, real-time obj ect detection [C]//Processed ings of the IEEE Computer Vision&Pattern Recognition, 2016:779-788.

[4] REDMON J, FARHADI A. Yolo9000: better, faster ter, stronger [C]//Proceedings of the IE-EE Conference on Computer Vision and P-attern Recognition, 2017:7263-7271

[5] REDMON J, FARHADI A. YOLOv3: an incremental improvement [J]. arXiv: 1804.02767.2018.

[6] BOCHKOVSKIY A, WANG C Y, LIAO H.YOLOv4: o-ptimal speed and a.ccuracy of object de section [J]. arXiv: 204.109342020.

[7] WANG C Y, BOCHKOVSKIY A, LIAO H.YOLOv7: t-trainable bag of freebies sets new stat-e-of-the-art for real-time object dead actors [EB/OL] (2022-07-02) [2022-10-26].

[8] Duan Zhongjing, Li Shaobo, Hu Jianjun, et al. Review of Deep Learning Object Detection Methods and Mainstream Frameworks [J]. Progress in Laser and Optoelectronics, 2020,57 (12): 59-74.

[9] REN S Q, HE K M, GIRSHICK R, et al. Faster R-CNN: rewards real-time object detection with region proposal networks [J] I-EEE Transactions on Pattern Analysis a-nd Machine Intelligence, 2017,39 (6): 1137-1149.

[10] Zhaowei Cai, and Nuno Vasconcelos, "Casc-ade R-CNN: Delving into High Quality O-object Detection," in IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.

[11] C. Fu, W. Liu, A. Ranga, A. Tyagi, and A.C. Berg. DSSD: Deconvolutional single shot detector CoRR, abs/1701.066592017.

[12] Bo Jingwen, Zhang Chuntang. Lightweight mask wearing detection algorithm based on YOLOv3 [J]. Electronic Measurement Technology, 2021,44 (23): 105-110. DOI: 10.19651/j.cnki. emt. 2107568.

[13] Zhang Lieping, Li Zhihao, Tang Yuliang. A lightweight YOLOv2 mask wearing detection method based on transfer learning [J]. Electronic Measurement Technology, 2022, 45 (10): 112-117. DOI: 10.19651/j.cnki.emt.2108620.

[14] Xiang Rongrong, Li Bo, Zhao Qiao. Mask wearing detection algorithm based on improved YOLOv5s [J]. Foreign Electronic Measurement Technology, 2022, 41 (07): 39-44. DOI: 10.19652/j.cnki. femt2203765.

[15] Li Liangfu, Wang Nan, Wu Biao, et al. Bridge crack image segmentation algorithm based on improved PSPNet [J]. Progress in Laser and Optoelectronics, 2021,58 (22): 101-109

[16] Ge S, Li J, Ye Q, et al. Detecting Masked Faces in the Wild with LLE-CNNs [C]//IEEE Conference on Computer Vision&P-pattern Recognition IEEE, 2017.

[17] Yang S, Luo P, et al. WIDER FACE: A Fa ce Detection Benchmark [C]//IEEE Conference on Computer Vision and Pattern Re cognition, IEEE, 2016.