UTM UNIVERSITI TEKNOLOGI MALAYSIA
INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Improve Accuracy and Response Time in Real-time Facemask Detection System

Zhong Baitong[1], Johan Mohamad Sharif[2], Farkhana Muchtar[3], Mohd KuFaisal Mohd Sidik[4]
Department Computer Science
Faculty of Computing, Universiti Teknologi Malaysia
81310, UTM Johor Bahru, Johor, Malaysia
Email: baitong@graduate.utm.my[1], johan@utm.my[2], farkhana@utm.my[3], mohdkufaisal@gmail.com[4]

Md Sah Salam
Deparment of Emergent Computing
Faculty of Computing, Universiti Teknologi Malaysia
Johor Bahru, Johor, Malaysia
Email: sah@utm.my

Submitted: 25/7/2024. Revised edition: 15/8/2025. Accepted: 25/8/2025. Published online: 30/11/2025
DOI: https://doi.org/10.11113/ijic.v15n2.477

***Abstract*—To improve people's health safety in public places and strengthen the government's epidemic prevention and control measures, the accuracy and response speed of mask detection in public areas need to be improved. The current real-time mask detection algorithm is developed based on a one-stage object detection algorithm in deep learning. The key to the algorithm research is how to improve the accuracy and response speed of the algorithm at the same time to meet the efficiency of real-time monitoring. Few of the current algorithms can achieve an accuracy of more than 90% under the same data set, and the FPS can also achieve the standard of real-time monitoring.**

***Keywords*—Real-time mask detection, deep learning, one-stage**

I. INTRODUCTION

Nowadays, with the progress of society, more and more people believe that human labor can be replaced by technology. However, the current mask testing technology still has many limitations. Currently, the mainstream algorithms are unable to accurately and promptly identify the mask-wearing situation. For instance, the YOLO-V4 algorithm can only achieve an accuracy rate of 89.2%, while YOLOV4-Large, although its accuracy rate reaches 94.2%, has a response speed of only 18 (FPS), which is 74% lower than the former. This study summarizes the current research on real-time mask detection and discusses the problems of the current research.

II. PROBLEM BACKGROUND

Mask detection is developed based on object detection technology. The task of object detection is to find out all the objects of interest in the image and determine their category and location, which is one of the core problems in the field of computer vision. Because of the different appearance, shape, and attitude of various objects, as well as the interference of illumination, occlusion, and other factors during imaging, object detection has always been the most challenging problem in the field of computer vision.

In terms of dataset selection, the mainstream approach is to choose the MS COCO (Microsoft Common Objects in Context) dataset. However, the MS COCO dataset itself does not include masks as a specific labeled category. Therefore, for mask detection, the MAFA dataset (Masked Face Dataset) is more suitable. The MAFA dataset is a large-scale dataset focused on research on masked face recognition, mainly applied to computer vision tasks such as mask detection and facial occlusion analysis, and contains approximately 88,000 instances of masked faces. The mainstream modification method is to improve the accuracy and response speed of one, sacrificing the other, according to the requirements.

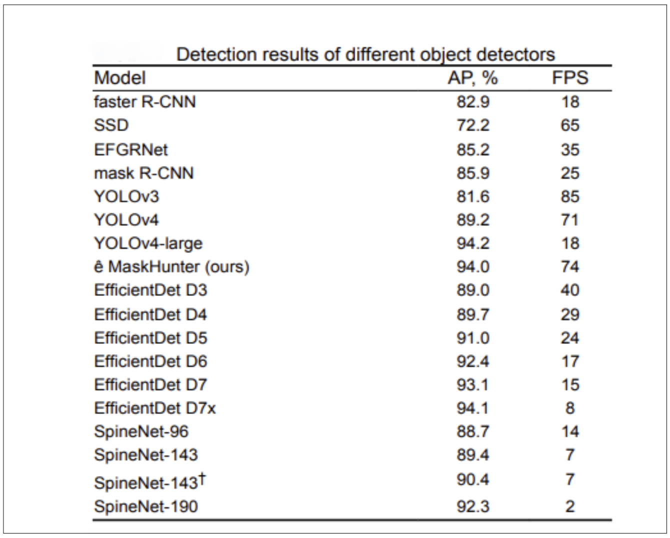
At present, real-time mask detection mainly focuses on the following two points.

The first point is that the accuracy rate of mask detection is too low, which will lead to false positives or missed positives in the system. If the alarm is not accurately reported, the safety of students will not be guaranteed. So, maintaining a high accuracy is a problem that needs to be studied. There are two problems in the accuracy of mask detection: the first one is what kind of feature extraction algorithms are used (one-stage or two-stage). Considering that this paper is monitored in real time on campus, one stage is more suitable. However, although one-stage feature extraction algorithms have a high response speed, in general, in order to meet the fixed-size input limit of the convolutional neural network classifier, the input image can be cropped, scaled, and other operations, but this will often distort the image.

The second point is the system response speed. The reason for improving the response speed is that many scenes need to meet the real-time performance, and the large flow of people in public places requires a high detection speed. At present, the mainstream practice is to use lightweight human object detection based on YOLOv3 [1], which can more quickly identify whether to wear masks and other issues. But there are still delays. Wang [2] introduced the improved spatial pyramid structure into the YOLOv3 algorithm and applied it to mask detection tasks, achieving a slight improvement in mAP and FPS indicators. However, the background of the algorithm data set is relatively simple, which makes it difficult to expand to complex multi-scene mask detection tasks, and the real-time detection speed is still not ideal. YOLOv4-tiny, published by Wang *et al*. in June 2020 [3], achieves 42.0% detection accuracy at 443FPS (on GeForce RTX 2080 Ti) real-time detection on the COCO [4] dataset. Although the speed is significantly improved, the accuracy is greatly reduced, which is not suitable for this scenario. Another reason for limiting reaction time is the transmission speed of the system. Even if cloud processing is adopted [5], transmission speed will affect the upper limit of system response speed.

III. Methodology

The classification system using deep neural networks can be considered the best approach to achieve high accuracy and give better results than other traditional approaches in terms of accuracy and loss functions. Cao [7] has made a corresponding study based on the MAFA dataset [8] and WIDER FACE data set [9] to improve the accuracy and response speed of mask detection and compared with the current mainstream algorithms. The specific data are shown in Fig. 1 as follows.



Detection results of different object detectors

Model	AP, %	FPS
faster R-CNN	82.9	18
SSD	72.2	65
EFGRNet	85.2	35
mask R-CNN	85.9	25
YOLOv3	81.6	85
YOLOv4	89.2	71
YOLOv4-large	94.2	18
ê MaskHunter (ours)	94.0	74
EfficientDet D3	89.0	40
EfficientDet D4	89.7	29
EfficientDet D5	91.0	24
EfficientDet D6	92.4	17
EfficientDet D7	93.1	15
EfficientDet D7x	94.1	8
SpineNet-96	88.7	14
SpineNet-143	89.4	7
SpineNet-143†	90.4	7
SpineNet-190	92.3	2

Fig. 1. Precision

It can be seen from the figure that only Mask R-CNN, EFGRNet, YOLOV4, and YOLOv4-Large have an accuracy of over 85. However, except for YOLOV4, the rest of the FPS are below 50. Too low a response speed will lead to too low efficiency of real-time monitoring. However, the accuracy of several algorithms with high FPS is lower than 85, which means that there are often false negatives and false positives in the detection. Therefore, the current deep learning algorithms for mask detection still need to be improved in detection accuracy and response speed.

The performance evaluation indices adopted by the training network in this experiment are mainly mAP (Mean Average Precision), Precision (P), and Recall (R), which are calculated by the following formula. In this paper, to detect the target face-mask wearing a mask, TP (True Positive) represents the number of faces wearing a mask correctly identified by the mask-wearing detection algorithm as a face-mask. FP (False Positive) represents the number of faces without masks that are incorrectly identified by the mask-wearing detection algorithm as faces with masks. FN(False Negative) represents the number of faces that the mask-wearing detection algorithm recognizes the mask as not wearing a mask. The average accuracy of all categories is the average accuracy mAP.

$$Precision = \frac{TP}{TP + FP} \times 100\% \hookleftarrow$$

$$Recall = \frac{TP}{TP + TN} \hookleftarrow$$

$$mAP = \frac{\sum_{i=1}^{\varepsilon} AP_i}{TP + FP} \times 100\% \hookleftarrow$$

This study will improve the algorithm based on YOLOv4-tiny and use the MAFA dataset and WIDER FACE dataset to compare with Mask Hunter and mainstream algorithms to verify the improvement effect of the algorithm. The main means of improvement include introducing SPP and improving it. Another way is using the path aggregation network (PAN) as the feature enhancement network.

A. Improving the Space Pyramid Pool

Inspired by SPP, this paper optimizes SPP accordingly, and the improved SPP is shown in Fig. 2. First, three pooling layers with a step size of 1 and a convolution kernel size of 5 × 5, 9 × 9, and 13 × 13 were used to maximize the input features, and three local features of different scales were obtained. Then, the obtained multi-scale local features are stacked with the input global features to achieve feature enhancement, and a richer feature table is obtained to further improve the detection accuracy of the network.

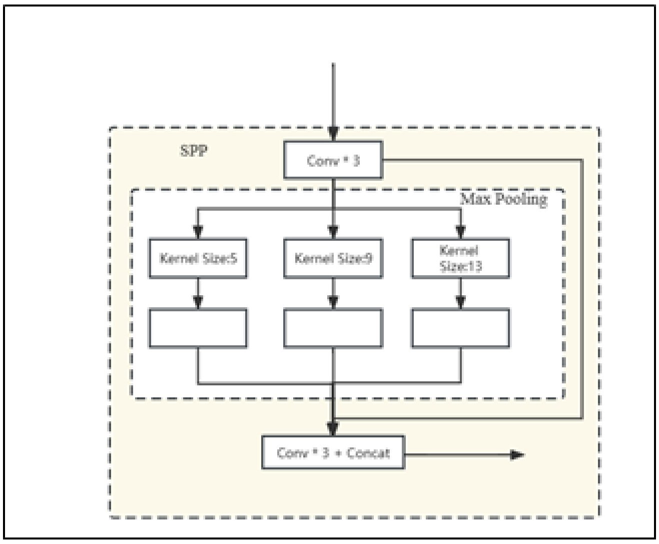


Fig. 2. Improved space pyramid pool

B. Feature Enhancement Network PAN

Because the feature information fusion of YOLO V4-tiny is processed by FPN, which is too simple, the overall receptive field is low, and the utilization rate of details is also low, the performance of YOLO V4-TINY in the face of small targets or blocked targets in complex scenes is not ideal. Therefore, this paper uses the path aggregation network (PAN) as the feature enhancement network, and the structure is shown in Fig. 3.

There are two feature input layers in the figure: F1 is from the backbone network, and F2 is from the SPP module. F2, as a bottom-up fusion path input, needs to undergo convolution and Upsampling processing, then is stacked with F1, and is convolved three times again without getting the first output. Similarly, F1, after convolution and Downsampling, is stacked with F2 and convolved three more times to get a second output. In this strategy, because the features from different scales are repeatedly enhanced and fused, the feature information is fully utilized, so that the expression ability of the prediction layer to the target is greatly enhanced.

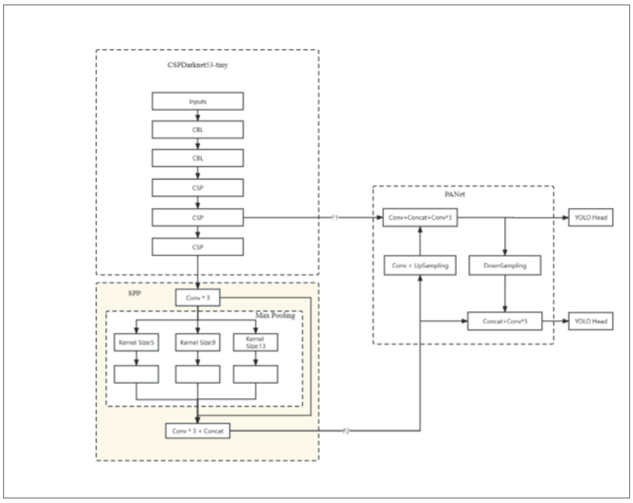


Fig. 3. PAN

IV. RESULT

Table I shows the comparison of mAP and FPS indicators between the proposed algorithm, YOLOv4-tiny, and other mainstream target detection algorithms. For face targets and face targets wearing masks, the mAP values of the proposed algorithm reach 95.7%, respectively, which increases by 1.7% and 2.1% compared with MaskHunter and YOLOv4-tiny, achieving higher detection accuracy. Due to the introduction of the SPP enhanced feature extraction network, the network size increases, and the speed is slightly lower than YOLOV4-TINY, but it is still faster than other algorithms.

TABLE I. MASK DETECTION RESULT

Algorithm	*mAP(%)*	*FPS*
SSD-512	***73.19***	***19.36***
YOLOV4	***95.5***	***22.5***
YOLOV3	***85.85***	***17.35***
YOLOV4-TINY	***93.6***	***112.34***
MaskHunter	***94.0***	***74***
Our Algorithm	***95.7***	***98.67***

V. CONCLUSION

By adding and strengthening the SPP and PAN modules, the accuracy is improved within the standard range. The R value and P value of the proposed algorithm reach 95.32% and 90.21% respectively, which are 1.61%and 4.14%， higher than those of YOLOv4-tiny. Compared with MaskHunter, the mAP is improved by 1.7%, and the response speed is improved by 24.67 FPS.

In practical applications, there are still many problems with mask detection. For instance, changes in visibility can affect the detection efficiency. Moreover, mask detection in densely populated areas often misses certain situations. Therefore, in subsequent research, more stress tests need to be conducted, and the algorithm needs to be upgraded to adapt to low visibility and large populations.

ACKNOWLEDGMENT

The authors would like to express their sincere appreciation to all individuals and institutions who contributed to this study. We are grateful for the constructive comments, support, and assistance received throughout the research and manuscript preparation. The authors also acknowledge any facilities, resources, or collaborations that have made this work possible.

CONFLICTS OF INTEREST

The authors declare that there is no conflict of interest regarding the publication of this paper.

REFERENCES

[1] Abed, A. (2022). Real-time multiple face mask and fever detection using YOLOv3 and TensorFlow Lite platforms. *Bulletin of Electrical Engineering and Informatics, 12*, 922–929. https://doi.org/10.11591/eei.v12i2.4227.

[2] Wang, Y. H. (2020). Mask-wearing detection algorithm based on improved YOLOv3 in complex scenes. *Computer Engineering*, 12–22.

[3] Wang, C.-Y., Bochkovskiy, A., & Liao, H.-Y. M. (2020). *Scaled-YOLOv4: Scaling cross stage partial network* (arXiv:2011.08036). arXiv. https://arxiv.org/abs/2011.08036.

[4] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014). Microsoft COCO: Common objects in context. In *European Conference on Computer Vision* (pp. 740–755). Springer.

[5] Ding, Y. (2022). JMDC: A joint model and data compression system for deep neural networks collaborative computing in edge-cloud networks. *Journal of Parallel and Distributed Computing, 173*, 83–93. https://doi.org/10.1016/j.jpdc.2022.11.008.

[6] Cao, Z., Shao, M., Xu, L., Mu, S., & Qu, H. (2020). MaskHunter: Real-time object detection of face masks during the COVID-19 pandemic. *IET Image Processing, 14*(16). https://doi.org/10.1049/iet-ipr.2020.1119.

[7] Ge, S., Li, J., Ye, Q., & *et al.* (2017). Detecting masked faces in the wild with LLE-CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

[8] Yang, S., Luo, P., Loy, C. C., & Tang, X. (2016). WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 5525–5533).