Faced-mask Detection based on YOLOv4

Ruohan Wen, Jiadun Qi, Yong Sun, Hao Tian, Cheng Huang and Kao-kin Zao
Department of Information Engineering
The Chinese University of Hong Kong
Hong Kong, China
johnkzao@cuhk.edu.hk

***Abstract*—With the tension raging around the world with the new coronavirus pneumonia, people have incorporated the wearing of masks into their daily lives. This paper is based on the existing epidemic situation and developed a model that can quickly detect masks. The prototype of the model is based on YOLOv4. We have improved the base of this model so that it can be quickly deployed on the production side through the container. At the same time, we introduce the feature enhancement module for effective feature enhancement. The experimental results also prove that our method has a certain improvement compared with the traditional target detection model, in which the speed is greatly improved.**

***Index Terms*—object detection, YOLOv4, container, real-time**

I. Introduction

Novel coronavirus pneumonia (COVID-19) refers to the pneumonia caused by the 2019 novel coronavirus infection. Due to the highly contagious and asymptomatic infection of the virus, it is often difficult to respond quickly to control the spread of COVID19 in the population. But people can slow the spread of COVID-19 by wearing masks. Some areas are urging people to wear masks through manpower inspections and counting the number. However, there are two problems in manual inspection. One is that the inspectors themselves are at risk of being infected, and the other is that it will cost huge human resources. If we can develop a set of applications or software that automatically recognize masks, we can deal with these two problems well.

In view of this situation, we can apply deep learning to this scene to assist people's mask wearing detection. There are many computer vision models that can provide us with a good technical reference, such as YOLO [1], Faster R-CNN [2], and SSD [3]. These models have cross-applications in many fields. Object tracking [4][5], medical image processing [6][7], instance segmentation [8][9] and other fields can see many applications of computer vision models. Among them, the YOLO series is a model that has performed very well in recent years. They perform better than the previous Faster R-CNN: faster, more real-time, smaller, and more accurate. Considering that the actual project requires a fast-deployable and lightweight model to complete mask detection, and there are certain requirements for detection speed, we chose YOLO. Among them, YOLO has a total of five versions YOLOv1 to YOLOv5 series[1][10][11][12][13]. We finally chose the YOLOv4. First, this model is more stable. At the same time, YOLOv5 itself has not improved much in this v4 version. For some special needs, we may have better optimizations.

We did find a more optimized method. First, we replaced and optimized a part of the convolution of YOLOv4 by GhostNet [14], and at the same time we use K-means [15] to get the best size of the anchor box to improve the accuracy and speed of mask detection. These two optimizations prove the correctness of our method in subsequent experiments. In the future after this, we will continue to explore this topic based on existing models and the latest methods, and develop new and more effective methods.

II. Related Work

A. YOLOv4

The current detection algorithm is relatively large, mainly divided into two types, one is single-step target detection, the other is two-step target detection, YOLOv4 is the one with better single-step target detection accuracy. The network model recently proposed by YOLOv4 not only has high detection speed but also very accurate detection accuracy.

As shown Fig. 1, its structural framework consists of a backbone feature extraction network Backbone, a feature fusion neck network Neck, and a head network YOLO Head for regression and classification. YOLOv4 adopts a large CSP structure in the entire DarkNet. The structure of CSPnet is not complicated, that is, the stack of the original residual blocks is split into two parts: the main part continues the stacking of the original residual blocks; the other part is like a residual edge, Connect directly to the end with a little processing. Therefore, it can be considered that there is a large residual edge in the CSP. The second is the addition of the SPP structure [16], which is mainly used to solve how feature maps of different sizes enter the fully connected layer. The feature maps of any size can be directly pooled with a fixed size to obtain a fixed

Ruohan Wen is the first author. She is currently pursuing a master's degree in Information Engineering at the Chinese University of Hong Kong. She will work towards a PhD degree in the future. Her E-mail is 1155166052@link.cuhk.edu.hk.

Jiadun Qi, Yong Sun, Hao Tian and Cheng Huang are currently pursuing a master's degree at the Department of Information Engineering, Chinese University of Hong Kong. Their E-mails are respectively {1155161048, 1155169190, 1155169008, 1155152552} @link.cuhk.edu.hk.

Professor John (Kao-kin Zao) is currently a Professor of Practice in Information Engineering at the School of Engineering, Chinese University of Hong Kong, and the corresponding author of this paper. His research interests are Internet Engineering (IoT & Edge Computing); Wireless Communications and Networking.

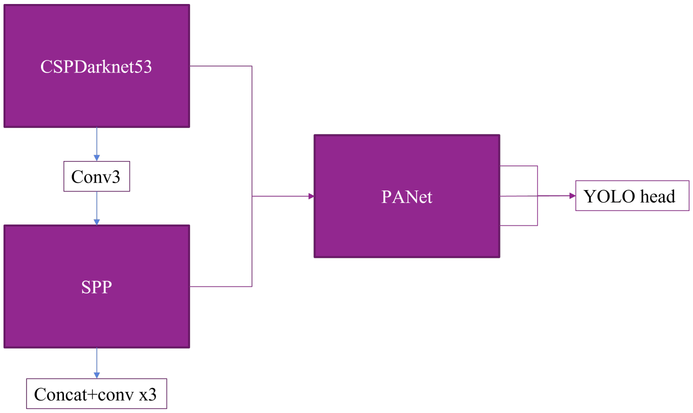


Fig. 1. The structure of YOLOv4.

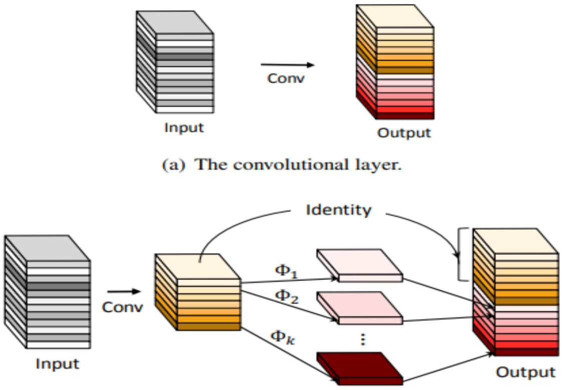
number of features. Its final features are transmitted to the PANs network [17], and the last three convolution branches output the head of YOLO respectively.

B. K-means

K-means is our most commonly used clustering algorithm based on Euclidean distance. It believes that the closer the distance between two targets, the greater the similarity. The algorithm first selects k objects at random, each object represents a cluster of initial average or center of each of the remaining objects with various clusters according to their distance from the center, then assigned to the nearest cluster [18]. Then recalculated the average of each cluster. This process is repeated until the criterion function converges.

C. GhostNet

Sufficient or redundant information in feature layers can always ensure a comprehensive understanding of the input data, and there are many similarities between feature layers, which are like ghosts of each other. This is also the origin of the Ghost network name. The core of it is that there is no need to generate these redundant feature maps with a large number of FLOPs and parameters. It is because the feature maps are partially similar, so there is no need to deconvolute them all, and directly use part of them as repeated redundant information to process [19].



(a) The convolutional layer.

Fig. 2. The Ghost module.

As shown in Fig. 2, the Module in GhostNet is divided into two steps to obtain the same number of feature maps as ordinary convolutions: i. smaller amount of convolution (For example, normally we use 16 convolution kernels, but here we use 8.) ii. cheap operations. In this way, the feature maps of the upper and lower parts are phantoms of each other. The operation of cheap in this article is: use the convolutional layer of group=c, which is, layer-by-layer convolution and is also the same as the first step in the depthwise separable convolution. They first use ordinary regular convolution to compress the input feature dimension c to m. Then they extended the m-dimensional feature to mxs=n by layer-by-layer convolution, and the layer-by-layer convolution obtained the ghost feature in the text, and finally concat the feature obtained by conventional convolution with the ghost feature as the output.

Based on advantages of Ghost Modules, the author introduces the Ghost bottleneck (G-bneck) specially designed for small CNNs, as shown in Fig. 3 [19]. Borrowing from MobileNetV2, ReLU is not used after the second Ghost module, and other layers apply BN and ReLU after each layer.

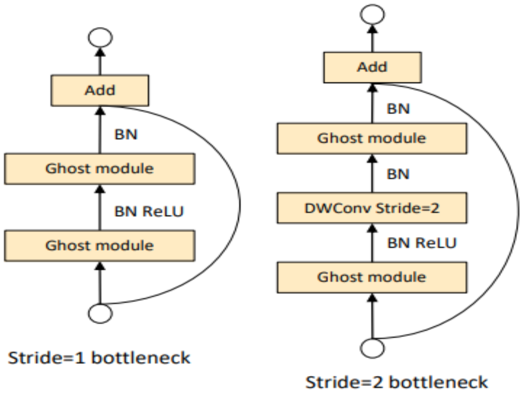


Fig. 3. Ghost bottleneck with different parameters.

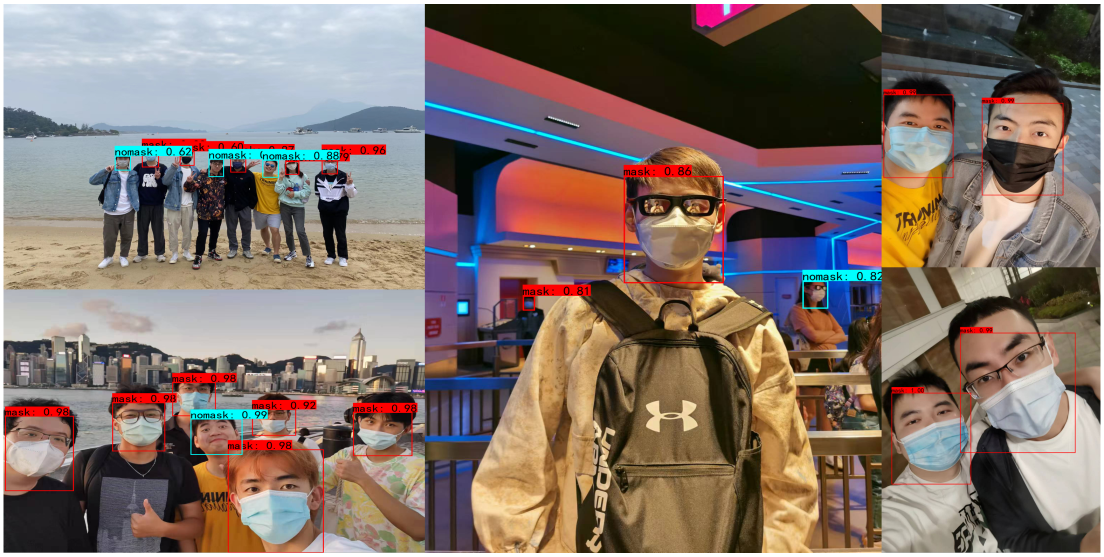


Fig. 4. Some interesting miscalculations.

For efficiency reasons, the initial convolution in the Ghost module is a point convolution.

D. Our Method

For our method and its improvemnr, because of the different shapes and sizes of human faces and masks, like some people like girls have small face, some men and boys may wear big mask, so we need to use K-means to generates a new anchor frame size, so that the anchor frame of the Yolov4 network model is more suitable for detecting different sizes of masks on different faces.

Simultaneously, the volumetric layer of YOLOv4 is dimensionally reduced using GhostNet model to reduce the number of model parameters and the amount of convolution computations, thereby improving the detection rate. We replaced the ”residual unit” unit in YOLOv4 with Ghost Bottleneck to achieve the effect of high-dimensional convolution, while reducing the parameters and computation of the convolution itself, improving detection efficiency. In terms of speed, because we reduce the number of parameters, we get a good improvement.

III. Dataset Preprocess and Hardware Equipment

A. Dataset Preprocess

For the dataset, we collected some images from our students themselves and some local passers-by in Hong Kong with their permission. Among them, 350 images were selected as the training set and 150 images were used as the test images. The pictures for the test also include pictures of our students themselves or taken.

The production process of the data set is to input these pictures into Labelme software for labeling to generate jsn files, and then convert them into xml files for training and testing.

B. Hardware Equipment

On the hardware equipment, we use the server provided by the Department of Information Engineering, School of Engineering, Chinese University of Hong Kong. Some parameters of its server are as follows: Ubuntu18.04, 16GB memory, and 2080Ti graphics card.

IV. Experimental Results

As shown in Fig. 4, these images are results of our method. The images we tested included long-range, close-up, face-decorated, and a variety of perspectives. Experimental results show that our method can detect objects well without the limitation of obvious occlusion.

As shown in Table I, Compared with the traditional YOLOv4, our method has more than 1% improvement, and even surpasses YOLOv5 in some aspects. Since YOLOv4 is combined with GhostNet, its speed is very small compared to YOLOv5, but it is more stable and easier to deploy on the production side than YOLOv5. Accuracy is calculated as follows:

$$precision = \frac{TP}{TP + FP}$$

But at the same time, we also found some exceptions. As shown in Fig. 5, If some behaviors, such as half covering the face with the mask, or blocking the hand near the mouth, will cause some misjudgments. At the same time, we also found some very interesting derivative problems, such as if black people wear black masks, or if the skin color is the same as the mask color, and some hunting masks (with funny patterns such as mouths and noses on them) will cause a certain degree of misjudgment. At the same time, when we tracked this problem, we found that these problems are the same as the black face recognition in face recognition. It is not obvious to mention a few features, which is also the direction of our future research and exploration.

TABLE I
COMPARISON OF DETECTION EFFECTS OF DIFFERENT MODELS

Model	mask	un_mask
YOLOv4	95.12%	96.03%
YOLOv5	96.62%	97.87%
SSD	94.78%	95.31%
Faster R-CNN	95.98%	95.04%
Our method	97.08%	97.85%

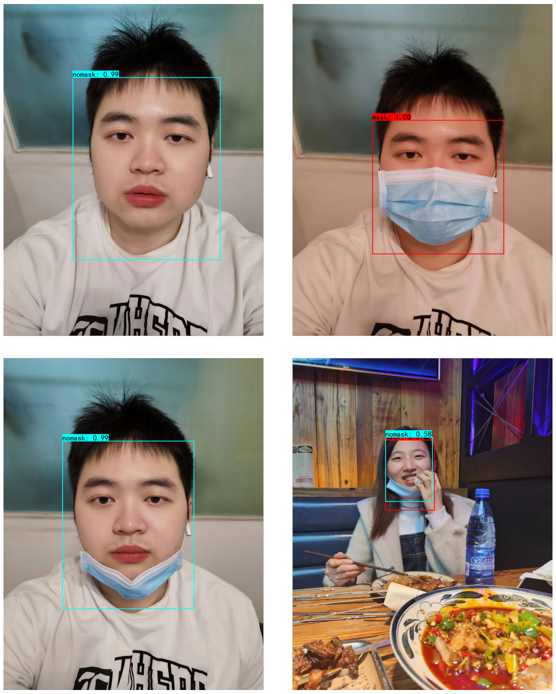


Fig. 5. Some interesting miscalculations.

V. CONCLUSION

In this paper, we combine the YOLOv4 and GhostNet networks, and use K-means as an auxiliary frame to implement a method that can automatically detect whether a mask is worn. Compared with the previous deep learning method, this method has a good improvement in accuracy, and its detection speed can be compared with YOLOv5. At the same time, in the process of exploring the subject, we have also discovered the direction and areas that we need to follow up and explore in the future. We will also do better research work in the future to help people get rid of the epidemic.

ACKNOWLEDGE

Thank you very much for your wonderful cooperation in this class of IEMS 5709. We not only completed intra-group cooperation, but also achieved good cross-group cooperation between groups. We also congratulate us on the success of this project!

REFERENCES

[1] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, ”You Only Look Once: Unified, Real-Time Object Detection,” 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 779-788.

[2] S. Ren, K. He, R. Girshick and J. Sun, ”Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[3] W. Liu, D. Anguelov, D. Erhan, S. Christian, S. Reed, C.-Y. Fu, et al., ”SSD: single shot multibox detector”, ECCV, 2016.

[4] Y. Zhang, Z. Chen and B. Wei, ”A Sport Athlete Object Tracking Based on Deep Sort and Yolo V4 in Case of Camera Movement,” 2020 IEEE 6th International Conference on Computer and Communications (ICCC), 2020, pp. 1312-1316.

[5] D. Mohanapriya and K. Mahesh, ”Multi object tracking using gradient-based learning model in video-surveillance,” in China Communications, vol. 18, no. 10, pp. 169-180, Oct. 2021.

[6] S. Zhang et al., ”MRI Information-Based Correction and Restoration of Photoacoustic Tomography,” in IEEE Transactions on Medical Imaging.

[7] A. S. Panayides et al., ”AI in Medical Imaging Informatics: Current Challenges and Future Directions,” in IEEE Journal of Biomedical and Health Informatics, vol. 24, no. 7, pp. 1837-1857, July 2020.

[8] K. He, G. Gkioxari, P. Dollár and R. Girshick, ”Mask R-CNN,” 2017 IEEE International Conference on Computer Vision (ICCV), 2017, pp. 2980-2988.

[9] Q. Feng, Z. Yang, P. Li, Y. Wei and Y. Yang, ”Dual Embedding Learning for Video Instance Segmentation,” 2019 IEEE/CVF International Conference on Computer Vision Workshop (ICCVW), 2019, pp. 717-720.

[10] J. Redmon and A Farhadi, ”YOL09000: better faster stronger[C]”, Proceedings of the IEEE conference on computer vision and pattern recognition., pp. 7263-7271, 2017.

[11] J. Redmon and A. Farhadi, ”Yolov3: an incremental improvement”, Computer Science, 2018.

[12] Alexey Bochkovskiy, Chien-Yao Wang and Hong-Yuan Mark Liao, ”YOLOv4: Optimal speed and accuracy of object detection”, 2020.

[13] T. F. Dima and M. E. Ahmed, ”Using YOLOv5 Algorithm to Detect and Recognize American Sign Language,” 2021 International Conference on Information Technology (ICIT), 2021, pp. 603-607.

[14] K Han, Y Wang, Q Tian et al., ”Ghostnet: More features from cheap operations[C]”, Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 1580-1589, 2020.

[15] R. M. Esteves, T. Hacker and C. Rong, ”Competitive K-Means, a New Accurate and Distributed K-Means Algorithm for Large Datasets,” 2013 IEEE 5th International Conference on Cloud Computing Technology and Science, 2013, pp. 17-24.

[16] K. He, X. Zhang, S. Ren and J. Sun, ”Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition,” in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 37, no. 9, pp. 1904-1916, 1 Sept. 2015.

[17] S. Liu, L. Qi, H. Qin, J. Shi and J. Jia, ”Path Aggregation Network for Instance Segmentation,” 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2018, pp. 8759-8768.

[18] L. Yang and M. Deng, ”Based on k-Means and Fuzzy k-Means Algorithm Classification of Precipitation,” 2010 International Symposium on Computational Intelligence and Design, 2010, pp. 218-221.

[19] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu and C. Xu, ”GhostNet: More Features From Cheap Operations,” 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 1577-1586.