Mask wearing detection based on YOLOv3

Xiaokang Ren*, Xingxing Liu*
School of Computer Science and Engineering, Northwest Normal University, Lanzhou 730070, China
*renxk@nwnu.edu.cn
*565898717@qq.com

Abstract: The CoVID-19 is still raging around the world, and the work of relying on manpower to detect masks in public places is time-consuming and laborious. To solve this problem, this paper proposes an improved Face_mask Net detection method for convolutional neural network based on YOLOv3. First, a new convolutional neural network Face_mask Net is designed based on YOLOv3 network structure. Second, before training, Face_mask Net uses the K-Means algorithm to cluster the labeled dataset and change its anchor value. Finally, the Face_mask Net loss function uses DIoU and the classifier uses DIoU-NMS. The combination of the two can further improve the detection accuracy of the target. To verify the effectiveness of the proposed algorithm, the Face_mask Dataset for mask detection was collected and annotated in this paper. The experimental results show that Face_mask Net can effectively detect the target of wearing mask and not wearing mask, and its performance can be close to real-time and the accuracy is higher than that of the network before the improvement.

1. Introduction
With the continuous improvement of target detection algorithms in deep learning, researchers have built a detection classifier based on convolutional neural network by studying different network depths, convolution kernel size and obtained excellent results. The target detection algorithm based on CNN is R-CNN[1], which proposes to combine selective search with CNN in the classification task. Fast R-CNN [2] reduces the operating cost by using the algorithm of region sharing deep feature mapping. Faster R-CNN[3]Regional Proposals Network (RPN), which share full image convolution features with the detection, have been introduced, thus achieving almost cost-free regional proposals (RPN). Instead of these neural networks, Redmon came up with a You Only Look Once [4](YOLO) framework, which treats object detection as a regression problem, separates bounding box and associated category probability spatially. In order to get faster and more accurate testing results, YOLOv2 adds the Batch Normalization layer after each volume layer. On the basis of YOLOv2 algorithm, the dataset is fused, and a real-time detection algorithm is proposed, which can identify more than 9000 classes of objects. The algorithm is called YOLO9000[5], YOLOv3[6] based on YOLOv9000, Darknet-53 was used as the backbone network and multi-scale fusion prediction was used to further improve the speed and skill level.

Although the epidemic in China is basically stable, the global epidemic is still severe, and it is imperative to guard against imported cases from abroad. Therefore, masks are still required when traveling and in crowded places. In order to reduce the workload of epidemic prevention personnel, the improved YOLOv3 network Face_mask Net was used to replace manual detection of whether pedestrians wore masks.
2. **YOLOv3 detection algorithm**

In order to improve the recall rate and accuracy of target detection by network, this paper compares four loss functions based on the deep convolutional neural network YOLOv3 of the single-step target detection algorithm. The YOLOv3 network model of four different loss functions was trained, and the model with high detection accuracy was selected through the comparison of mAP evaluation indexes and applied to mask detection.

2.1 **DIoU**

![Figure 1 Comparison of loss functions of IoU, GIoU and DIoU. Green and red represent the target box and the prediction box.](image)

As shown in Fig.1, when the target box completely encloses the prediction box, the values of Intersection over Union (IoU)[7] and Generalized Intersection over Union (GIoU)[8] are the same. At this time, GIoU degrades into IoU and its relative position relationship cannot be distinguished. Distance-IoU (DIoU)[7] can better optimize such problems by adding the normalized distance of the center point. In addition, compared with GIoU Loss, DIoU loss has a faster convergence speed. It takes into account the overlapping area and the distance from the center point, but does not take into account the aspect ratio.

\[
L_{DIoU} = 1 - IoU + \frac{\rho^2(b + b_{gt})}{c^2} (1)
\]

In this loss function, \(b, b_{gt}\) represents the center point of anchor frame and target frame respectively, \(\rho\) is for calculating the Euclidean distance between the two centers, and \(c\) represents the diagonal distance of the minimum rectangle that can cover both anchor frame and target frame. DIoU feature: Similar to GIoU loss, it can still provide the movement direction for the boundary box when it is not overlapped with the target box. DIoU Loss\((L_{DIoU} = 1 - DIoU)\) can directly minimize the distance between two target frames, so it converges much faster than GIoU loss. DIoU can also replace the common IoU evaluation strategy and be applied to non-maximum Suppression (NMS) to make the results obtained by NMS more reasonable and effective.

2.2 **CIoU**

Based on DIoU, Complete IoU (CIoU)[7] is defined as follows:

\[
L_{CIoU} = 1 - IoU + \frac{\rho^2(b + b_{gt})}{c^2} + \alpha \nu (2)
\]

In formula (4), CIoU has two more parameters \(\alpha\) and \(\nu\) than DIoU. Where \(\alpha\) is the coefficient used to balance the proportion, and \(\nu\) is used to measure the proportional consistency between the anchor box and the target box. Their formula is as follows:

\[
\nu = \frac{4\pi^2}{\pi^2 (arctan \frac{\omega_{gt}}{h_{gt}} - arctan \frac{\omega_{}}{h})} (3)
\]

\[
\alpha = \frac{\frac{\nu}{(1-IoU) + \nu}} (4)
\]

3. **Algorithm in this paper**

3.1 **Face_mask Net**

Face_mask Net's main job is to detect whether people wear masks in public places. In order to improve the detection accuracy, on the basis of YOLOv3, the loss function in the original network structure was replaced, that is, IoU was replaced with DIoU. Secondly, in order to further improve the
accuracy of the network, the NMS function in yolo module in Fig.2 is replaced by DIou-NMS. The network structure is shown in Fig.2.

![Figure.2 Face_mask Net network structure](image)

### 3.2 optimize anchor frame

YOLOv3 introduced FPN[9]. The idea is to detect objects on three different scales by combining the high resolution features of the shallow layer and the high level features of the high semantic information. According to the target characteristics under the background of this paper, the dataset is clustered before network training, and K-means[10] is adopted. The algorithm can obtain the anchor value with high accuracy, so as to assign more accurate boxes to the target on the larger feature map. K-means clustering was used to obtain the size of anchors, and nine anchor point frames were taken, which were respectively (11,14), (18,23), (27,33), (39,50), (57,72), (80,106), (118,154), (182,235), (305,388). Therefore, each cell in each scale should use three anchor points to predict three boundary boxes.

### 3.3 DIoU-NMS algorithm

In traditional NMS, the IoU index is often used to suppress the redundant prediction box, and often produces error suppression for occlusion. DIoU-NMS regards DIoU as the criterion of NMS, because not only overlapping areas but also the distance between the center points between the two borders should be considered in the inhibition criteria, and DIoU considers both. For the prediction box $M$ with the highest score, the update formula of DIoU-NMS can be formally defined as:

$$ s_i = \begin{cases} s_i, & \text{IoU} - R_{DIoU}(M, B_i) < \epsilon \\ 0, & \text{IoU} - R_{DIoU}(M, B_i) \geq \epsilon \\ \end{cases} $$

Among them, box $B_i$ is deleted by considering the distance between IoU and the center points of the two boxes at the same time, $s_i$ is the classification score, and $\epsilon$ is the NMS threshold. DIoU-NMS suggests that borders with far central points may be on different objects and should not be deleted. This is the biggest difference between DIoU-NMS and NMS.

### 4. Mask detection experiment and results

In this paper, all network model training environments are Ubuntu 18.04.3LTS under Linux system, and Tesla P4 8GB memory video card is used. By means of transfer learning, first obtained in VOC[11], then the subdivisions were 10,400 times of training combined with the Face_mask Dataset training. The Subdivisions were 16, the Batch size was 64, the initial learning rate was 0.0005 and the Momentum was 0.9. At the beginning of the experiment, based on the YOLOv3 network, the original loss function was replaced with IoU, GIoU, DIoU and CIoU respectively. Four models with different loss functions were trained on Face_mask Dataset and named as IoU-NMS, GIoU-NMS and
CIoU-NMS respectively. In order to verify the effect of anchor after optimization, this paper independently trains the loss function as DIoU, and the classification function is the model of DIoU-NMS, and it is named DIoU-DIoU_NMS.

4.1 Experimental dataset
The Dataset in this article is the self-collected and summarized mask Dataset and is named Face_Mask Dataset. The dataset mainly selected 9056 photos of people from different regions (such as pedestrians wearing masks and pedestrians without masks), which were labeled with LabelImg. The dataset tag divides the data into two categories, one named Face (people who don't wear a mask and cover their face with their hands, etc.) and the other named Face_mask (people who wear a mask correctly). The data of pedestrians without masks was mainly selected from WIDERFace[12] and MAFA[13]Dataset, from which about 3114 images were selected and re-annotated. Data of pedestrians wearing masks are from web crawlers, video screenshots of pedestrians wearing masks and RWMFD[14]Dataset, also about 6312 images were re-annotated. Among them, 8242 pieces were randomly selected as training sets, 814 pieces as verification sets and 91 pieces as test sets.

4.2 mAP assessment under IoU index
In order to verify the accuracy of the model, all the five models were trained in the same environment, and the images were processed into 416×416 size before training. The performance comparison results are shown in Table.1. IoU is used as an indicator to evaluate the mAP table of five models. The maximum value in a column is marked in red. The higher the mAP value is, the higher the accuracy of the model will be. As can be seen from Table 1, on AP50 and 55, Face_mask Net is only 0.8 and 0.7 percentage points lower than CIoU-NMS, but it is superior to other models on AP60-75.

| network models       | AP50  | AP55  | AP60  | AP65  | AP70  | AP75  |
|----------------------|-------|-------|-------|-------|-------|-------|
| IoU-NMS              | 0.797 | 0.777 | 0.752 | 0.693 | 0.613 | 0.491 |
| GIoU-NMS             | 0.793 | 0.769 | 0.746 | 0.691 | 0.626 | 0.527 |
| CIoU-NMS             | 0.813 | 0.808 | 0.787 | 0.739 | 0.69  | 0.561 |
| DIoU-DIoU_NMS        | 0.806 | 0.799 | 0.784 | 0.7577| 0.703 | 0.591 |
| Face_mask Net        | 0.807 | 0.801 | 0.787 | 0.7579| 0.706 | 0.624 |

DIoU-DIoU_NMS and Face_mask Net are compared separately. Face_mask Net has a higher accuracy than DIoU-NMS network model on the whole. It can be seen that optimizing anchor helps to improve the model accuracy.

4.3 Face_mask Net test mask experiment
Fig3 shows the effect when different loss function models detect the three targets simultaneously. It can be seen that the accuracy of the five network models for the mask wearer on the left reaches over 99%. However, the detection accuracy of the mask wearer on the right remains above 99% except for IoU-NMS and Face_mask Net. IoU-NMS is lower than Face_mask Net in the detection of the middle mask wearer, and Face_mask Net maintains the accuracy of more than 99%. It can be seen that Face_mask Net is slightly better than the other four models in multi-target detection.

(a)IoU test results  (b)GIoU test results  (c)CIoU test results
5. Conclusion
In this paper, four different loss functions are compared mainly based on YOLOv3 network, and anchor value of the network model is optimized to further improve the accuracy and accelerate the convergence speed. The optimized network structure can detect the target of wearing masks in real time and realize the tracking effect. Secondly, the training, verification and test set of pedestrian targets wearing masks is designed and marked by ourselves. Experimental results show that the new convolutional neural network Face_mask Net proposed in this paper has good robustness for pedestrian target detection with masks, and the accuracy can reach 80%--99%.

References
[1] Girshick R, Donahue J, Darrell T, et al. Rich feature hierarchies for accurate object detection and semantic segmentation [C] // Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014:580-587.
[2] Girshick R. Fast r-cnn[C]//Proceedings of the IEEE international conference on computer vision. 2015: 1440-1448.
[3] Ren S, He K, Girshick R, et al. Faster R-CNN: towards real-time object detection with region proposal networks[C]. neural information processing systems, 2015: 91-99.
[4] Redmon J, Divvala S, Girshick R, et al. You only look once: Unified, Real-time Object Detection [C]//Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition, 2016;779 -- 788.
[5] Redmon J , Farhadi A . YOLO9000: Better, Faster, Stronger[C]// IEEE Conference on Computer Vision & Pattern Recognition. IEEE, 2017:6517-6525.
[6] Redmon J , Farhadi A . YOLOv3: An Incremental Improvement[J]. arXiv e-prints, 2018.
[7] Zheng Z , Wang P , Liu W , et al. Distance-IoU Loss: Faster and Better Learning for Bounding Box Regression[C]// AAAI Conference on Artificial Intelligence. 2020.
[8] Rezatofighi H , Tsoi N , Gwak J Y , et al. Generalized Intersection Over Union: A Metric and a Loss for Bounding Box Regression[C]// 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2020.
[9] Ding L, Goshtasby A. On the Canny edge detector[J]. Pattern Recognition, 2001, 34(3): 721-725.
[10] Divya V, Deepika R, Yamini C, et al. An Efficient K-Means Clustering Initialization Using Optimization Algorithm[C]//2019 International Conference on Advances in Computing and Communication Engineering (ICACCE). IEEE, 2019: 1-7.
[11] Yang, Shuo et al. "WIDER FACE: A FACE Detection Benchmark." 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016): N. Pag. Crossref.web.
[12] Everingham M , Gool L V , Williams C K I , et al. The Pascal Visual Object Classes (VOC) Challenge[J]. International Journal of Computer Vision, 2010, 88(2):303-338.
[13] Ge S, Li J, Ye Q, et al. Detecting masked faces in the wild with lle-cnns[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017: 2682-2690.
[14] Wang Z, Wang G, Huang B, et al. Masked Face Recognition dataset and Application [J]. ArXiv Preprint arXiv:2003.09093, 2020.