Response to Reviewer 3 Comments

Point 1: Comments and Suggestions for Authors

The paper presents an interesting subject, but the following aspects must be explained more clearly in order to increase the soundness of the paper:

section about related work must be included, also containing results of the described methods (existing work that are already described must be added in a separate section)

Response 1: Thanks to the reviewers for their valuable comments on this article. We have added a section "Related Work" on the third page of the article. In this section, we pointed out some of the problems in current object detection, and at the same time stated some algorithms related to face mask detection in the existing work. So as to make the content of the article more substantial and reliable. The "related work" is as follows:

2. Related Works

2.1. Problems Exists in Object Detection

There are two key points of face mask wearing detection. One is to locate the position of the face in the image; the other is to identify whether the face given by the data set is wearing a mask and if the mask is worn correctly. Problems of the present object detection algorithm can be attributed to face occlusion, variable face scale, uneven illumination, density etc., these problems seriously affect the performance of the algorithm. What's more, the traditional object detection algorithm adopts the selective search method **Error! Reference source not found.** in feature extraction would lead to problems such as poor generalization ability, redundant information, low accuracy and poor real-time performance.

2.2. Existing Work

Some researchers have used the extraction of RGB color information to perform face mask recognition [27]. However, the article does not consider the case of non-standard wearing of masks, so the adaptability of the algorithm needs to be further improved. Combined YOLO-v2 and ResNet50, authors in [28] realized face mask recognition whose backbone network is DarkNet-19. However, DarkNet-19 has been optimized by CSPDarkNet53. The ablation experiment in our paper shows that the CSP1_X module produces better results than CSPDarkNet53. In [29], authors pointed out that the combination of ResNet50 and SVM can realize face mask detection and its accuracy can reach to 99.64%. But the algorithm takes a lot of computational costs. What's more, the combination of SSD and MobileNetV2 for mask detection was proposed in paper [30], but its model structure is too complex and its performance is inferior to YOLO-v4.

Only two categories are used in the papers mentioned in the above paragraph and authors did not consider the influence of wearing mask irregularly on the algorithm. Therefore, the feature extraction ability and model practicability of these algorithms need to be improved. In this paper, based on improved YOLO-v4, face mask recognition is considered and three categories: face_mask, face and WMI are included. In addition, the feature extraction ability of this paper is improved by CSP1_X, and CSP2_X impels PANet to speed up the circulation of semantic feature and strengthen feature fusion, thus improves the robustness of the model.

Point 2: the chosen of the base network that is improved must be explained based on existing results

Response 2: In the second paragraph of Section 2.2, we explain why the improved YOLO-v4 is used for face mask detection based on the problems existing in the existing face mask detection algorithm. The contents are shown below.

Some researchers have used the extraction of RGB color information to perform face mask recognition [27]. However, the article does not consider the case of non-standard wearing of masks, so the adaptability of the algorithm needs to be further improved. Combined YOLO-v2 and ResNet50, authors in [28] realized face mask recognition whose backbone network is DarkNet-19. However, DarkNet-19 has been optimized by CSPDarkNet53. The ablation experiment in our paper shows that the CSP1_X module produces better results than CSPDarkNet53. In [29], authors pointed out that the combination of ResNet50 and SVM can realize face mask detection and its accuracy can reach to 99.64%. But the algorithm takes a lot of computational costs. What's more, the combination of SSD and MobileNetV2 for mask detection was proposed in paper [30], but its model structure is too complex and its performance is inferior to YOLO-v4.

Only two categories are used in the papers mentioned in the above paragraph and authors did not consider the influence of wearing mask irregularly on the algorithm. Therefore, the feature extraction ability and model practicability of these algorithms need to be improved. In this paper, based on improved YOLO-v4, face mask recognition is considered and three categories: face_mask, face and WMI are included. In addition, the feature extraction ability of this paper is improved by CSP1_X, and CSP2_X impels PANet to speed up the circulation of semantic feature and strengthen feature fusion, thus improves the robustness of the model.

Point 3: the novelty of the proposed method must be explained more clearly

Response 3: We explain the novelty of this paper in detail from line 76 to line 84. The novelty of this paper is as follows:

- Aiming at the problem of training time, this paper introduces the improved CSPDarkNet53 into the backbone to realize the rapid convergence of the model and reduce the time cost in training.
- An adaptive image scaling algorithm is introduced to reduce the use of redundant information in the model.
- To strengthen the fusion of multi-scale semantic information, the improved PANet is added into the Neck module.

The Hard-Swish activation function introduces in this paper can not only strengthen the nonlinear feature extraction ability of the network but also enable the detection results of the model more accurate.

Point 4: uniformise references: eg. line 189: "He et al."

Response 4: In view of this problem, we have carried on the detailed inspection to the full text and has carried on the unified adjustment and the correction.