Response to Reviewer 1 Comments

Point 1: The English spelling and grammar must be significantly improved. The reviewer suggests proofreading the whole manuscript. Some examples include:

> Lines 29-30: *"They can enter the lungs as humans breathe and then causing pneumonia, nerve damage or even toxic reactions, which can be so serious that endanger lives."*

Response 1: Thank you for your valuable suggestions. We carefully follow the opinions of the reviewers and improve the spelling and grammar of the full text words.
In lines 24 to 25, we changed this sentence to "Things like, respiratory infection viruses, toxic and harmful gases and dust suspended in the air can enter the lungs of humans as breathe and then cause pneumonia, nerve damage and toxic reactions.".

Point 2: Figure 9: Incorrect spelling of the word "Mouth".

Response 2: Thank you for your valuable comments. In picture 10, the spelling of "mouse" is wrong, and then we changed it to the correct "mouth".

Point 3: Lines 597 to 599: "Because the objects in the data set used in this paper can be directly observed, and the wearing of face masks at night is not considered."

Response 3: In lines 623 to 624, we changed this sentence to "In addition, the case of wearing a mask when the light is insufficient is also not considered.".

Point 4: The background section is incomplete. It covers the development of object detection algorithms based on deep learning, but fails to explain the advantages of the proposed method versus other methods related to face mask detection. Further examination of the literature needs to be carried out and research gaps and study contributions need to be highlighted.

Response 4: We have refined the background in detail. In the "Related Work" section, we explain the advantages of the algorithm in this paper over other face mask detection algorithms, and at the same time the differences between the current algorithms and the research contributions brought by the algorithm in this paper.

Point 5: The evaluation parameters, as stated in section 5.5, are unclear and need to be further clarified: "FP (false positive) means that other categories are detected as being related to this category by mistake; FN (false negative) indicates that the real sample is detected as the opposite result or undetected category", the definitions of FP and FN seem to be similar.

Response 5: In section 6.5, line 489 to line 490, we define FP as " FP (false positive) represents the number of samples that the detected object category is inconsistent with the real object category " to prevent the definitions are similar.

Point 6: In Table 7, the reasoning time for Faster R-CNN is a lot higher than that of other models. However, Table 10 shows that the image size for this model (600*600) is larger than other models. Can this be the reason for the significantly slower performance of Faster R-CNN when compared with other models?

Response 6: The main reason for the long inference time of Faster R-CNN is that the full connection layer is used at the end of Faster RCNN, so that the model occupies a large part of the number of parameters. All retained ROI pooling layers will pass through this full connection layer after processing the ROI and will be calculated separately without sharing the calculations between them. Generally, Faster R-CNN limits the size of input images to 600 minimum edges and 1000 maximum edges. In order to consider the calculation cost, the size is selected as (600*600).

Point 7: In the Introduction section, it is stated that the aim of this paper is to develop a face mask detection method suitable for complex environments. However, according to results displayed in Figure 11, most test subjects were very close to the camera and do not actually represent a real time complex scenario. Moreover, the reviewer believes that a resolution of 316*316 is not good enough to accurately detect when the subject is far away from the camera as in a real complex environment. Do elaborate on that.

Response 7: The complex environment referred to in this paper means that the face is in occlusion, multi-scale, occlusion, etc. in figure 12, the number of faces in the image from left to right is increasing constantly, and there are occlusion, multi-scale problems in the face. At the same time, we added two sets of pictures in picture 12. In the rightmost image, the face is very far from the camera. The face in these two pictures is in a state of serious distortion, so the face is a difficult to detect sample. Through comparative experimental analysis, we can see that the algorithm in this paper has better detection results for such targets. At the same time, the distance from face to camera is getting farther and farther from left to right image. In the experiment of this paper, the size of the image input is 416*416, but the size of the image input can be changed. You can specify 416*416, 608*608, 1664*1664 etc. The image size only needs to be a multiple of 32. However, when people are far away from the camera, it means that there is less feature information, which puts forward higher requirements for network feature extraction ability. Therefore, in order to enable the network to extract deeper feature information and improve the detection ability of the model, we have improved the YOLO-v4 feature extraction network accordingly, and the effect of these improvements is very obvious. When the image size is 416*416, our algorithm achieves better detection results.