Journal of Science and Technology
ISSN: 2362-1443 (Print) 3059-9490 (Online) 2026, Vol. 5, No. 1

Cascaded YOLO v8 and ResNet-50 based Masked Face Detection

Andolan Parajuli
Pulchowk Campus, Institute of Engineering
Tribhuvan University, Nepal
andolanp@gmail.com

Sanjeeb Prasad Panday*
Pulchowk Campus, Institute of Engineering
Tribhuvan University, Nepal
sanjeeb@ioe.edu.np

Santosh Giri
Pulchowk Campus, Institute of Engineering
Tribhuvan University, Nepal
santoshgiri@pcampus.edu.np

Sakar Pudasaini
Pulchowk Campus, Institute of Engineering
Tribhuvan University, Nepal
sakar.pudasaini@gmail.com

* Corresponding author

***Abstract*—Large population of the world have been affected due to respiratory infectious diseases. General recommendations for proper face masks usage, regular and proper hand hygiene as well as employing social distancing at such infection prone areas could help in managing the transmission of these illnesses. Masked face detection would be used as an essential tool for the monitoring as well as employment of different control measures. The works included in the article consists of two parts: bounding box estimation of the faces (with and without occlusions) that was accomplished by the implementation of modified YOLO v8. The second part was the implementation of ResNet-50 for classification of the faces. The models used was trained with a masked face dataset that included 853 images. The dataset consists of properly masked faces, improperly masked faces as well as unmasked faces. The modified YOLO v8 model with FPN was used for training of the available training set for bounding box estimation of the faces. Here, the neck of YOLO v8 (PANet) was replaced with the FPN. The mean average precision of the modified model was found to be better than that of the unmodified YOLO model. The mAP50 for the modified YOLO v8 was 0.877 which was better as compared to the unmodified YOLO v8 model with 0.847. Similarly, the use of ResNet-50 was done for classificaiton of the faces obtained from the bounding box estimated using YOLO models. The least training loss of the ResNet-50 was obtained to be 0.4744.**

***Keywords* —Modified YOLO v8, ResNet-50, FPN, Masked Face Detection**

I. Introduction

The prevalence of respiratory infectious diseases have affected the livelihood of individuals throughout the world. The respiratory virus could enter from hosts to hosts via droplets or aerosols contaminated with the virus of concern. For proper control and prevention of the spread of the virus, the mask mandates ought to be observed in public areas since, researches indicated that nonpharmaceutical interventions (NPIs) influence respiratory diseases, and easing NPIs could result in the resurgence of such respiratory conditions [1]. During both pandemic and non-pandemic scenarios, employing non-drug-related public health strategies like practicing hand hygiene, wearing face coverings, and maintaining social distance can provide accessible, cost-effective, and efficient ways to minimize the transmission and consequences of acute respiratory infections [2]. Similarly, new virus outbreaks in the future is more likely.

Face detection is one of the computer vision application that encompasses the identification of faces in a given digital image. Additionally, it serves as the initial step for different technologies related to processing of human faces. Human face detection is an easy and natural task for humans but executing the task for machines is complicated and requires different computationally intensive steps. The accuracy of the face detection depends upon various factors such as illumination, distance, orientation, face occlusion and so on. Due to the mandates of masks in the public, the the images obtained could contain masks making it difficult to detect the faces. Therefore, masked face detection is required [3].

The presence of occlusion in the face makes it difficult for face detection for the masked individuals. Different methods have been utilized for masked face detection. Presently, some deep learning models are popular that trains on the masked face dataset [4]. You Only Look Once (YOLO), ResNet, MobileNet, and so on are some of the algorithms that are used for masked/unmasked face detection.

DOI : https://doi.org/10.3126/jost.v5i1.85919
Article History: Received: 17 Nov. 2025, Revised: 17 Jan. 2026, Accepted: 26 Feb. 2026.

II. Related Works

Different works has been carried out for the purpose of detection of faces and objects with different occlusions, complex backgrounds as well as variable range of illuminations. The works for efficient pedestrian detection was carried out by [5] (2018) based on a deep detection model by adjusting the algorithm with respect to the pedestrians characteristics. Another works by [6] (2022) used a model based on YOLO v5 for the detection as well as location of humans with high success rate. The dataset of around one thousand images for the training as well as test of the model. The authors in [7] (2022) used a face detection algorithm based on the YOLO v4-tiny making it a lightweight model that they named SMD-YOLO. The use of the detection model on small and medium faces showed improved mAP values and showed to be able to perform in real time environment. The works of ResNet-50 was done by [8] (2021) for the masked face recognition. Different optimizer (SGD and adam) were used for masked and unmasked faces to improve accuracy of the model. It was found that convolutional neural networks could be used for recognition of unmasked faces but it would struggle for faces with occlusions. Cascaded models based on YOLO v2 and ResNet-50 was shown to be efficient for masked face detection as shown by the works of [9] (2021). The use of IOU and data augmentation was done for improvement of the results. Similarly, the authors of [10] (2021) used a cascaded model based on ResNet-50, FPN and ResNet-18 for masked face detection. The networks consisted of a Dilation RetinaNet Face Locaiton network (with ResNet-50 based backbone and FPN for feature fusion) and SRNet20 based on ResNet-18. Similarly, the experimental use of modified YOLO v5 with the replacement of the backbone with ResNet-50 and DenseNet and neck with biFPN was done for small object detection. The authors of the works named the model YOLO-z in [11] (2021).

Also, it could be concluded that the object detection performance of the YOLO models under different occlusions, objects in complex backgrounds, different sizes of the object to be detected as well as for different illuminations could be improved according to the works by [12] (2021), [13] (2022), [14] (2023) [15] (2023). The works and findings of the works was the motivation of carrying the required tasks for the proposed model.

III. Methodology

The works for the detection of masked faces would require the execution of different steps that is shown in the figure 1.

As shown in the figure 1, the steps that were required for the execution of detection model for masked faces were theoretical study, data acquisition, training of models, bounding box estimation, classification of faces and evaluation of the models. After required theoretical study was done, the dataset was acquired from kaggle and required organization of the dataset was done according to the requirements of the models used. The organized dataset was used by the unmodified YOLO v8 model first for the bounding box estimation. Then, the estimated faces were used for the training of the ResNet-50 model that would be used for the classification of the faces into masked, unmasked and improperly masked faces. Then the overall performance of the model were evaluated.

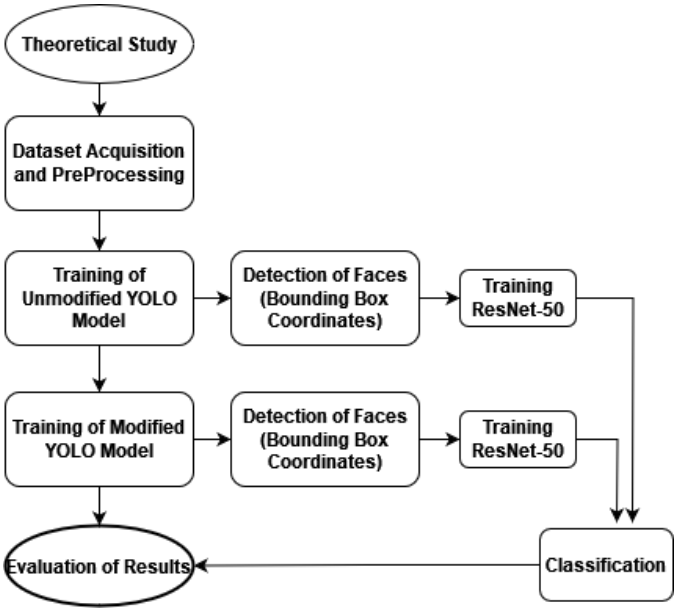

Fig. 1. Proposed Framework

A. Dataset

The dataset used for the models was obtained from kaggle and was publicly available. The dataset comprises 853 images, classified into three categories: unmasked, properly masked, and improperly masked faces (MakeML Dataset, 2020). Within the dataset, there are 3,232 instances of masked faces, 113 instances of improperly masked faces, and 717 instances of properly unmasked faces. The recommended architecture for training on this dataset is PyTorch Faster R-CNN. Consequently, annotations and dataset organization should conform to the specifications of the YOLO and RESNET-50 architectures. The distribution of the dataset is outlined in Table I. The dataset was divided in to training set of 767 images, validation set of 61 images and test set of 25 images.

TABLE I
Dataset Distribution

Classes	Training	Validation	Test
Properly Masked Faces	3116	776	389
Improperly Masked Faces	117	36	21
Unmasked Faces	691	175	76

Some of the sample images of the dataset could be observed in the figure 2.

Fig. 2. Sample Images of the Dataset Used

B. YOLO v8

YOLOv8 represents the most recent advancement in YOLO models, offering capabilities in tasks like object detection, image classification, and instance segmentation. Developed by Ultralytics, YOLOv8 surpasses its predecessor, YOLOv5, through innovations such as anchor-free detection, the introduction of a new convolution block (C2f), and mosaic augmentation. The implementation of anchor-free detection reduces the necessity for bounding box predictions, thereby enhancing the speed of Non-Maximum Suppression (NMS). Additionally, the substitution of the 6x6 convolution block with a 3x3 block in the new convolution scheme further contributes to the model's improved performance. [16]

1) Training: To train the dataset with unmodified YOLOv8s, it was partitioned into training, validation, and test sets. Specifically, 767 images were allocated for training, 61 for validation, and 25 for testing. The training process involved utilizing 3 batches across 50 epochs. The trained model architecture had the total layers of 168 layers, 11,126,745 parameters and 28.4GFLOPS. The training information for the unmodified model (YOLO v8s) is shown in the table II.

The training process for 767 images from the training set using the standard YOLOv8s model concluded within 0.840 hours over 50 epochs. Following training, the best and final weights of the model were saved in their designated directories. The model utilized a total of 168 layers, containing 11,126,745 parameters, with a computational requirement of 28.4 GFLOPS. The model's pre-processing speed was 0.3 milliseconds, inference speed was 6.7 milliseconds, and post-processing speed was 2.9 milliseconds per image.

TABLE II
TRAINING INFORMATION OF YOLO V8S

Class	Instances	Box (P	R	mAP50	mAP50-95)
all	3577	0.808	0.794	0.847	0.589
with_mask	2820	0.904	0.917	0.953	0.689
mask_weared _incorrect	109	0.692	0.679	0.746	0.516
without_ mask	648	0.829	0.787	0.843	0.562

The training result of unmodified YOLOv8s could be observed in the figure 3. The figure depicts the losses and precision throughout both the training and validation phases. Notably, a consistent reduction in losses is evident during both training and validation. Similarly, there is an overall upward trend in precision observed across.

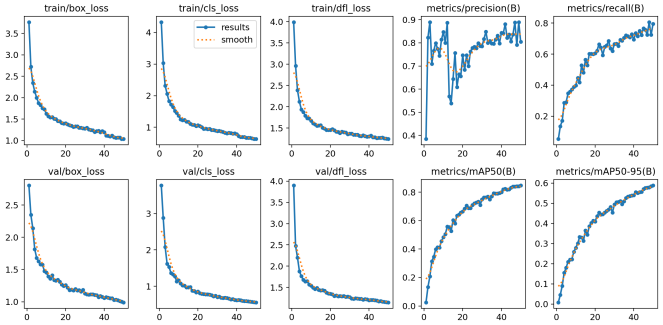

Fig. 3. Results of Training with unmodified YOLOv8

C. Modification in YOLO v8

The modification in the original YOLO v8 architecture involves replacing the neck architecture with FPN. The differences between the original and modified architectures include variations in the total layers utilized for training, the number of parameters involved, the GFLOPs required, as well as the training duration, among other parameters. The modification in the YOLO architecture could be observed in the figure 4.

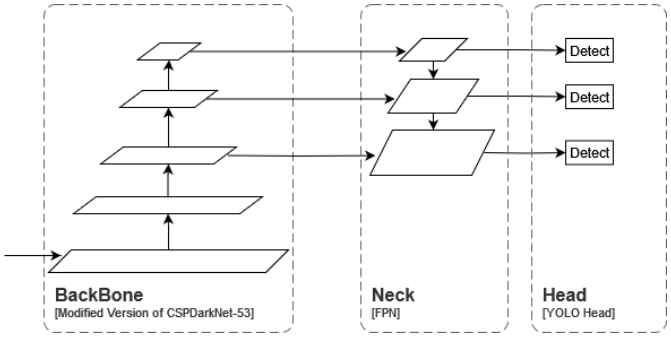

Fig. 4. Modified YOLO Model

1) Training: The modified YOLO v8 model underwent training using 767 training images, 61 for validation, and 25 for testing. The training process involved three batches across 50 epochs. Upon completion, the architecture yielded a total of 200 layers, consisting of 22,301,017 parameters, with a computational requirement of 71.4 GFLOPS. The training information for the modified model is shown in the table III.

TABLE III
TRAINING INFORMATION OF MODIFIED YOLO V8

Class	Instances	Box (P	R	mAP50	mAP50-95)
all	3577	0.89	0.815	0.877	0.624
with_mask	2820	0.937	0.915	0.96	0.704
mask_weared_ incorrect	109	0.866	0.743	0.808	0.58
without_ mask	648	0.867	0.785	0.863	0.588

Training for the modified YOLO v8 with 767 images from the training set concluded within 0.916 hours across 50 epochs. Following training, the best and final weights were saved in their respective directories. The training utilized a total of 200 layers, comprising 22,301,017 parameters, with a computational requirement of 71.4 GFLOPs. Pre-processing speed averaged 0.4 milliseconds, inference speed was approximately 8.1 milliseconds, and post-processing required around 2.2 milliseconds per image.

The training result of the modified YOLOv8 is depicted in Figure 5. This figure illustrates the losses and precision across both training and validation phases. Notably, a consistent decline in losses is evident during both training and validation. Likewise, there is a noticeable overall upward trend in observed precision.

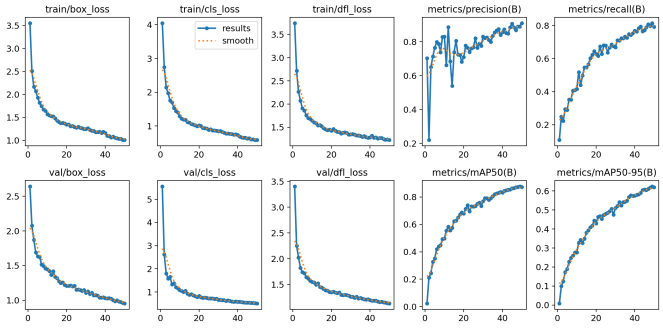

Fig. 5. Results of Training with Modified YOLOv8 with FPN as the Neck

D. ResNet-50

The ResNet-50 model has been employed to classify masked/unmasked faces and detect incorrectly worn masks on individuals. To prepare the faces dataset for classification, the data from the masked face dataset underwent cropping over and then trained for five epochs, and the final weights of the algorithm were obtained for classification purposes. Cross-entropy loss served as the training criterion, and the Adam optimizer was utilized during the training epoch. The outcomes of the ResNet50 training are displayed in Table IV.

TABLE IV
TRAINING RESULT FOR RESNET-50

Epoch		Loss	Accuracy
2*1	Training	0.6805	0.7321
	Validation	0.4775	0.819
2*2	Training	0.5343	0.8085
	Validation	0.3869	0.8675
2*3	Training	0.4744	0.8279
	Validation	0.3407	0.8795
2*4	Training	0.5044	0.7930
	Validation	0.3556	0.8675

E. Experimental Setup

The works for detection of masked faces were conducted using Google Colaboratory, which employs Python 3 within a Jupyter Notebook environment for executing Python code directly in the browser. Google Colaboratory was chosen for its accessibility, as it offers hosted Jupyter Notebooks without the need for setup, provides free access to computational resources, and is well-suited for implementing machine learning and data science projects. The computational resources available on Google Colaboratory included a T4 GPU with 15102MB of GPU RAM, 12.7GB of system RAM, and 78.2GB of system storage.

F. Sample Detection Output

Some of the samples of the output obtained after trained unmodifed YOLO v8, modified YOLO v8 as well as ResNet-50 architecture could be observed in the figures below.

1) Bounding Box Estimation using unmodified YOLO v8s: Some sample images for the detected faces using YOLOv8 is shown in the figure 6.

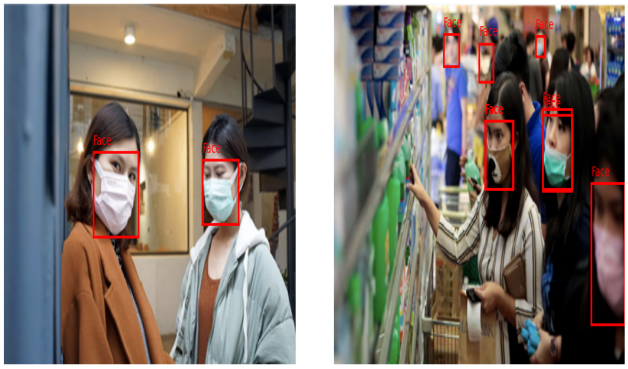

Fig. 6. Output from unmodified YOLOv8

2) Bounding Box Estimation using modified YOLO v8: Some sample images for the detected faces using modified YOLOv8 is shown 7.

3) Classification using ResNet-50: Some of the output obtained by using trained ResNet50 for classification is given below.

The output of the trained ResNet50 model's predictions showcases detected faces along with the probabilities assigned to each class the faces belong to. The classification is based on the class with the highest probability.

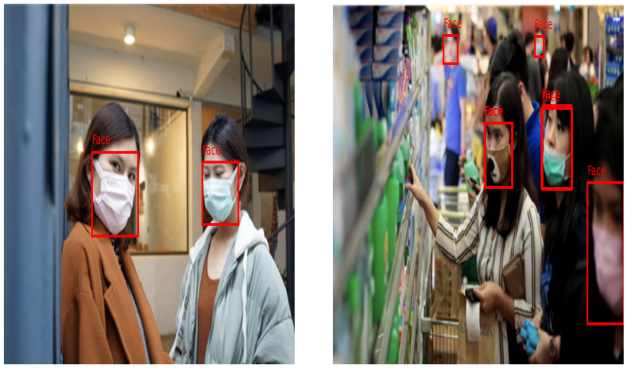

Fig. 7. Output from modified YOLOv8

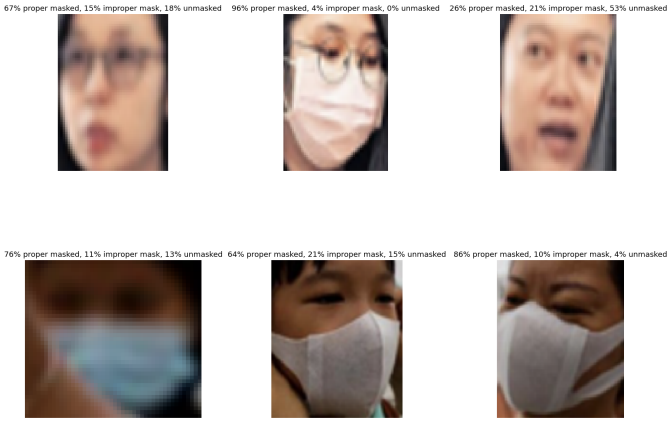

Fig. 8. Classification Output for ResNet50

IV. Results Analysis

The comparison between the unmodified YOLO v8s and the modified YOLO v8 is presented in Table V. This table displays the mean average precision of the models, including precision values at 50% and across the range of 50% to 95%.

It is evident that the modified YOLO v8 demonstrates superior mean average precision compared to the unmodified YOLOv5 model across instances of masked faces, unmasked faces, and incorrectly masked faces. Specifically, the overall mAP50 for YOLOv8s was determined to be 0.847, whereas for the modified YOLO v8 (with FPN as the neck), it was calculated to be 0.877. Similarly, the mAP50-95 for YOLOv8s (overall) was recorded as 0.589, whereas for the modified YOLOv8, it stood at 0.624. Further analysis of the data can be conducted using Figure .

TABLE V
Comparison between unmodified YOLO v8 and Modified YOLO v8

2*Class	2*Instances	YOLOv8s		Modified YOLO v8	
		mAP50	mAP50 -95	mAP50	mAP50 -95
All	3577	0.847	0.589	0.877	0.624
With mask	2820	0.953	0.689	0.96	0.704
Incorrectly masked	109	0.746	0.516	0.808	0.58
Without Masks	648	0.843	0.562	0.863	0.588

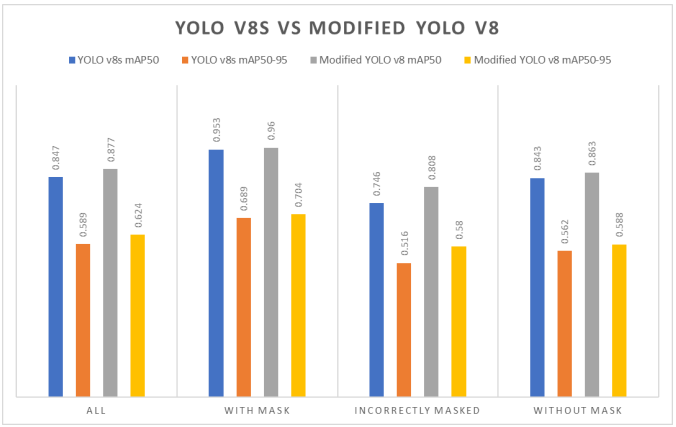

Fig. 9. YOLO v8s vs Modified YOLO v8

As per the presented data, the modified version of YOLO v8 exhibits superior mean average precision when compared to the original YOLO v8 across the entirety of the training dataset, encompassing instances of masked faces, unmasked faces, and improperly masked faces.

A. Differences Observed in the Output of the Models

Figure 8 visually demonstrates the disparities observed in the detection process between the trained YOLO v8 and the modified YOLO v8, utilizing the provided dataset. The images illustrate instances where the modified YOLO v8 displayed improvements in face detection compared to the unmodified YOLO v8. Additional faces detected by the modified YOLO v8 are highlighted in the images on the right side, indicated by arrows.

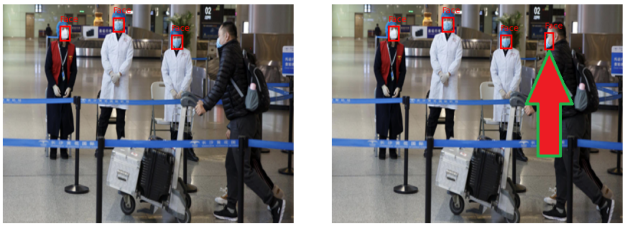

Fig. 10. Differences observed during detection using unmodified YOLO v8 and modified YOLO v8

V. Conclusion

Research on masked face detection holds potential for observing and implementing non-pharmaceutical interventions to manage respiratory infectious diseases. This study on masked face detection, employing modified YOLO and ResNet models, was conducted through theoretical studies, literature reviews, planned methodologies, and diverse experimental setups to achieve set objectives. Modifications were made to YOLO v8 by replacing the neck with the Feature Pyramid Network. The modified models were trained using a masked face dataset organized to meet model requirements, and their precision was evaluated. YOLO models were employed for bounding box estimation of faces in images. Unmodified ResNet-50 model training focused on classifying faces with masks, without masks, and with improperly worn masks. Experimental results indicated improved precision for the modified models. Likewise, ResNet-50 was utilized to classify faces derived from the bounding boxes estimated by YOLO models, achieving the highest training accuracy.

VI. Future Works

The modified models utilized in the execution and analysis of the model also exhibited certain limitations in their bounding box estimations. It was noted that the bounding box estimation showed inconsistencies for faces within complex backgrounds, small objects, and similar scenarios. This indicates the potential for further enhancements in the models to achieve improved results.

To optimize feature extraction, adjustments can be made to the YOLO v8 backbone. One approach involves substituting the CSPDarknet backbone with ResNet or DenseNet models. These models leverage feature reuse and gradient flow mechanisms, potentially enhancing the overall performance of the model.

Additional refinement of the utilized models could lead to improved performance in face detection. This involves steps such as hyperparameter tuning and data augmentation. Increasing the diversity of the training dataset can be achieved through various augmentation methods, including random rotations, flips, adjustments in brightness and contrast, among others.

References

[1] Z. Zuo, C. Yang, F. Ye, M. Wang, J. Wu, C. Tao, Y. Xun, Z. Li, S. Liu, J. Huang, and A. Xu, "Trends in respiratory diseases before and after the covid-19 pandemic in china from 2010 to 2021," *BMC Public Health*, vol. 23, p. 217, 02 2023.

[2] J. Alqahtani, A. Aldhahir, S. Alrabeeah, M. Alnasser, I. AlDraiwiesh, S. Alghamdi, R. Siraj, J. Sreedharan, A. Alqahtani, and E. Alzahrani, "Future acceptability of respiratory virus infection control interventions in general population to prevent respiratory infections," *Medicina*, vol. 58, p. 838, 06 2022.

[3] A. Kumar, A. Kaur, and M. Kumar, "Face detection techniques: A review," *Artificial Intelligence Review*, vol. 52, 08 2019.

[4] X. Li, "Masked face detection and calibration with deep learning models," *Journal of Physics: Conference Series*, vol. 2196, no. 1, p. 012011, feb 2022. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2196/1/012011

[5] Z. Liu, Z. Chen, Z. Li, and W. Hu, "An efficient pedestrian detection method based on yolov2," *Mathematical Problems in Engineering*, vol. 2018, pp. 1–10, 2018.

[6] C. Hou, "The application of human detection based on yolov5," *Highlights in Science, Engineering and Technology*, vol. 34, pp. 203–208, 2023.

[7] Z. Han, H. Huang, Q. Fan, Y. Li, Y. Li, and X. Chen, "Smd-yolo: An efficient and lightweight detection method for mask wearing status during the covid-19 pandemic," *Comput. Methods Prog. Biomed.*, vol. 221, no. C, jun 2022. [Online]. Available: https://doi.org/10.1016/j.cmpb.2022.106888

[8] B. Mandal, A. Okeukwu, and Y. Theis, "Masked face recognition using resnet-50," *arXiv preprint arXiv:2104.08997*, 2021.

[9] M. Loey, G. Manogaran, M. H. N. Taha, and N. E. M. Khalifa, "Fighting against covid-19: A novel deep learning model based on yolo-v2 with resnet-50 for medical face mask detection," *Sustainable Cities and Society*, vol. 65, p. 102600, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2210670720308179

[10] R. Zhu, K. Yin, H. Xiong, H. Tang, G. Yin, and Y. Huang, "Masked face detection algorithm in the dense crowd based on federated learning," *Wirel. Commun. Mob. Comput.*, vol. 2021, jan 2021. [Online]. Available: https://doi.org/10.1155/2021/8586016

[11] A. Benjumea, I. Teeti, F. Cuzzolin, and A. Bradley, "Yolo-z: Improving small object detection in yolov5 for autonomous vehicles," *arXiv preprint arXiv:2112.11798*, 2021.

[12] Y. Zhao and S. Geng, "Face occlusion detection algorithm based on yolov5," *Journal of Physics: Conference Series*, vol. 2031, no. 1, p. 012053, sep 2021. [Online]. Available: https://dx.doi.org/10.1088/1742-6596/2031/1/012053

[13] J. Qi, X. Liu, K. Liu, F. Xu, H. Guo, X. Tian, M. Li, Z. Bao, and Y. Li, "An improved yolov5 model based on visual attention mechanism: Application to recognition of tomato virus disease," *Computers and Electronics in Agriculture*, vol. 194, p. 106780, 2022. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0168169922000977

[14] J. Zhang, J. Zhang, K. Zhou, Y. Zhang, H. Chen, and X. Yan, "An improved yolov5-based underwater object-detection framework," *Sensors*, vol. 23, no. 7, 2023. [Online]. Available: https://www.mdpi.com/1424-8220/23/7/3693

[15] J. Lai, Y. Liang, Y. Kuang, Z. Xie, H. He, Y. Zhuo, Z. Huang, S. Zhu, and Z. Huang, "Io-yolov5: Improved pig detection under various illuminations and heavy occlusion," *Agriculture*, vol. 13, no. 7, 2023. [Online]. Available: https://www.mdpi.com/2077-0472/13/7/1349

[16] RangeKing, "Brief summary of yolov8 model structure." [Online]. Available: https://github.com/ultralytics/ultralytics/issues/189