



RESEARCH ARTICLE

Heterogeneous Ensemble Approaches for Robust Face Mask Detection in Crowd Scenes

Xufeng Hu¹ , Younghoon Jeon¹ and Jeonghwan Gwak^{1,2,3,4,*} ¹Department of Software, Korea National University of Transportation, Republic of Korea²Department of Biomedical Engineering, Korea National University of Transportation, Republic of Korea³Department of AI Robotics Engineering, Korea National University of Transportation, Republic of Korea⁴Department of IT & Energy Convergence, Korea National University of Transportation, Republic of Korea

Abstract: Face masks are one of the effective tools to slow the spread of disease and reduce medical overload by protecting people from infectious diseases including COVID-19. To prevent infection from respiratory droplets, it is imperative to wear a mask that covers the nose and mouth completely. However, it is difficult to make it mandatory for crowds to wear masks in public places where many people gather. For example, detecting incorrect mask-wearing in crowded scenes is a tedious and attention-grabbing task. Therefore, the success of deep learning in computer vision motivates automated monitoring systems. However, deep learning-based detection models are unstable if the domain task is changed and may have different strengths and weaknesses. Therefore, in this study, we propose a heterogeneous ensemble-based detection model for robust face mask detection in crowd scenes. First, independent detection models such as You look Only Ones (YOLO) v6, YOLO v7, and Faster R-CNN are employed for the model ensemble. Second, the prediction results obtained from the detection models are post-processed such as merging, non-maximum suppression, and weighted box fusion. The experimental results show that the classification performance of our proposed model has an F1 score of about 90.5% and that the improvement of the generalization ability due to the ensemble strategy contributed to the improvement of the classification performance.

Keywords: face mask detection, crowd scenes, deep learning, heterogeneous ensemble

1. Introduction

COVID-19 is one of the epidemics affecting millions of people worldwide, spreading exponentially across more than 150 countries from 2019 to 2022. The disease has spread exponentially across more than 150 countries, resulting in approximately 530 million confirmed cases of the virus and approximately 6 million deaths between 2019 and 2022. There are mainly two routes of transmission of viruses: (1) direct transmission and (2) indirect transmission. First, direct transmission is caused by droplets discharged through the respiratory tract of an infected patient. Second, indirect contact is diffusion through objects. Coronavirus can survive for a certain amount of time on the surface of objects such as doorknobs and tables. Therefore, contact with unsterilized objects increases the risk of infection. Although risks such as mortality have recently decreased due to the popularization of vaccines for the treatment of COVID-19, many people are still at risk of infection.

A face mask is one of the effective tools to mitigate the rising trend of infection by covering the nose and mouth with a cloth to prevent transmission. However, the duty to wear a mask may be neglected for

various reasons, such as a shortage of supplies, inconvenience, and a lack of safety. Along with the declining sense of duty, it is common for most institutions to hire additional staff to monitor the wearing of masks. However, mask-wearing surveillance is a tedious and attention-grabbing task for inspectors and can be less effective in crowded situations.

The recent success of deep learning in computer vision has motivated the development of automated mask detection systems. The performance of mask detection can be improved through tracking (Yang et al., 2017). Through compression of the deep learning-based models, it can be advanced so that mask detection can be performed even in the internet of things devices (Kang & Gwak, 2020). However, most of the detection algorithms are based on personal detection, which only detects one person at a time. In addition, the detection of incorrect wearing or non-wearing tends to be ignored because the focus is only on positive samples indicating the correct wearing of the mask. Monitoring systems based on single-target detection algorithms are inappropriate because they limit the movement of people in public places and can cause confusion. Furthermore, it must successfully detect targets that appear relatively small due to their distance from the visual sensor and those with seemingly limited information on the edge of the detection equipment. Algorithms for multi-object tracking in the

*Corresponding author: Jeonghwan Gwak, Department of Software, Korea National University of Transportation, Republic of Korea. Email: james.han.gwak@gmail.com

video have been proposed, including Faster R-CNN using convolutional neural network-based region proposals and You look Only Ones (YOLO) models using grid method and reliability. It can also be used for abnormal situation detection. Moreover, the process of finding an optimal strategy for parameter optimization can be utilized (Gwak et al., 2014). However, according to our empirical results, these detection models are sensitive to the characteristics of domain tasks and have different strengths and weaknesses.

Therefore, this study proposes a deep learning-based ensemble approach for robust mask-wearing detection in crowd scenes. First, independent detection models such as YOLO v5, YOLO v6, YOLO v7, and Faster R-CNN are employed for the construction of the ensemble model. The prediction results of each model are merged with post-processes such as non-maxima suppression and weighted box fusion (WBF). Our model is rigorously evaluated using statistical techniques, with two datasets for single-object detection and multi-object detection.

The rest of this paper consists of four parts. Section 2 gives a brief overview of related works on object detections, mask detections, and ensemble strategies. Our proposed detection method is introduced in Section 3. Section 4 shows the experimental results and discussion for the evaluation of our proposed model. Finally, Section 5 summarizes and concludes this paper.

2. Literature Review

Deep learning has been successfully applied in many fields recently (Fang et al., 2022; Yang, 2022; Yang & Song, 2022; Suresh et al., 2021). In this section, several literatures are presented to help understand our proposed work. First, representative deep learning-based object detection models are introduced in Section 2.1. Second, several studies using object detection algorithms for mask-wearing detection tasks are described in Section 2.2. Finally, the concept of ensemble approaches and various strategies are described in Section 2.3.

2.1. Object detection models

The deep learning-based object detection model mainly uses the following two strategies: (1) single-phase strategy and (2) multi-phase strategy. In the single-phase strategy, the object detection model obtains the category and location information of the object directly from the input image. Therefore, this strategy requires less computation and can be designed intuitively. On the contrary, a multi-phase strategy employs a candidate region extraction process that proposes regions in the input image where objects may be present before object detection. Therefore, the object detector uses the candidates to predict object information such as existence, class, and location. The YOLO series is one of the representatives of single-phase object detection models, which significantly reduces the inference time with little loss of performance. From the initial version of YOLO, seven versions have been developed. In the development of YOLO, seven versions have been updated; since YOLO v1-v4 belongs to the previous version, the performance of this model is weak. We only study YOLO v5 and its subsequent versions YOLO v6 and YOLO v7. The description of the YOLO series is limited to YOLO v5 (Redmon et al., 2016), YOLO v6 (Li et al., 2022), and YOLO v7 (Wang et al., 2022) employed as independent object detection models in this study. YOLO v5 is implemented using PyTorch instead of Darknet (Chen, 2019) to improve accessibility. Mosaic data enhancement (Zeng et al., 2021) is adopted at the main contribution input end of YOLO v5, with reference to the method of CutMix data enhancement (Yun et al., 2019). Data enhancement to improve the imbalance of small, medium, and large target data in

data concentration is used. If the dataset itself has many small targets, mosaic data enhancement will cause the originally small targets to become smaller, resulting in the poor generalization ability of the model. YOLO v6 has customized a series of networks of different scales for industrial applications in different scenarios. To achieve the best trade-off between speed and accuracy, architectures of different scales vary. The small model uses a simple single-path trunk, while the large model is built on an efficient multi-branch block. The main purpose of YOLO v6 is to promote its application in industrial applications. YOLO v6 injects a self-distillation strategy, performing both classification and regression tasks. At the same time, the author dynamically adjusts the knowledge from teachers and labels to help students learn more effectively at all training stages. Advanced label allocation detection technology, loss function, and data enhancement technology have also been widely used and verified, and they are selectively used to further improve performance. The author of YOLO v7 primarily focuses on optimizing the model from two key perspectives. First, the re-parameterization technique, as discussed by Ding et al. (2021), centers on optimizing the optimizer rather than the architecture itself. Second, the work conducted by Ding et al. (2022) involves the development of Convolutions into Fully-connected Layers for Image Recognition. In recent years, model structure re-parameterization and dynamic label allocation have become the main optimization direction in the field of target detection. For the structure re-parameterization, the structure re-parameters of different layers of the network are optimized by analyzing the propagation path of the gradient, and the structure re-parameterization of different planning models is proposed. In dynamic label assignment, because the model has multiple output layers, it is difficult to assign better dynamic targets to different branches during training. Another multi-stage strategy method for object detection is characterized by (1) generating a suggestion area, (2) extracting features from the suggestion area, and (3) classifying the extracted category features. Faster R-CNN (Ren et al., 2015) is a representative object detection model using a multi-stage strategy. The region recommendation network extracts regions of interest to reduce computing time. Compared with other methods based on depth learning, Faster R-CNN supports end-to-end learning at all layers while maintaining robustness (Akshatha et al., 2022; Cygert & Czyżewski, 2022). However, to avoid overlapping candidate boxes, the model uses the classification score as the standard, and then uses non-maximum suppression (NMS) (Neubeck & Van Gool, 2006) for post-processing. This method is easy to cause false negatives (FNs) and is not friendly to occluded objects.

2.2. Mask detection systems

Cygert and Czyżewski (2022) proposed an optimistic CNN based on MobileNet (Howard et al., 2017) for face mask detection. The proposed network was trained using a large dataset consisting of high-quality images, but incorrect mask-wearing cannot be detected because it only focuses on the presence of the mask. Ren & Liu (2020) propose the labeling of the training dataset through the K-means clustering algorithm and the use of distance-intersection over union (IoU) as a loss function for efficient training of YOLO v3. The comparison results of four different loss functions show that the proposed method contributes to the improvement of accuracy and acceleration of the convergence speed. However, the proposed model focuses only on wearing a mask, and the images used for training and validation are low in complexity, which may cause unstable performance in places with many connections. Degadwala et al. (2021) evaluated the performance of YOLO v4 for the classification of three classes, masked correctly, no mask, and masked incorrectly,

but the complexity of the image, such as a crowd scene, was not considered. Sharma (2020) has been and has become a model that shows superior performance in the detection of small objects in the task of detecting the wearing of a mask through a camera in real time. However, in the performance evaluation experiment of the proposed model, actual venues with a lot of people were not considered, and it is impossible to detect the case of wearing a mask incorrectly.

2.3. Ensemble learning approaches

Ensemble learning aims to design one strong agent by combining several weak agents. A homogeneous class uses an array of independent classifiers with the same architecture, whereas a heterogeneous ensemble employs a set of independent classifiers of different types. The three main classes of ensemble learning methods are bagging, stacking, and boosting. Boosting and bagging, one of the ensemble-based model design strategies, merge several weak classifiers trained using an array of datasets obtained through repeated random sampling. Dataset diversification due to bagging techniques can contribute to mitigating overfitting. Recently, methodologies to adapt ensemble learning to object detection tasks have been proposed. In the bagging strategy, it is possible to design models in parallel because each weak classifier can be trained independently. Ensemble learning for object detection tasks is mainly about efficiently merging the prediction results of different independent models. Therefore, Vilhelm et al. (2022) analyzed the impact of ensemble methodology in designing object detection models, including finding desert vehicles in ultra-high-resolution images such as satellite images. Experimental results show that ensemble strategies such as bagging allow high reliability at the expense of additional training and increased prediction times. Casado-García et al. (2020) proposed several voting strategies and increased testing time for ensembles of models. The voting strategy consists of three parts, (1) affirmative, (2) consensus, and (3) unanimous. The proposed ensemble strategy

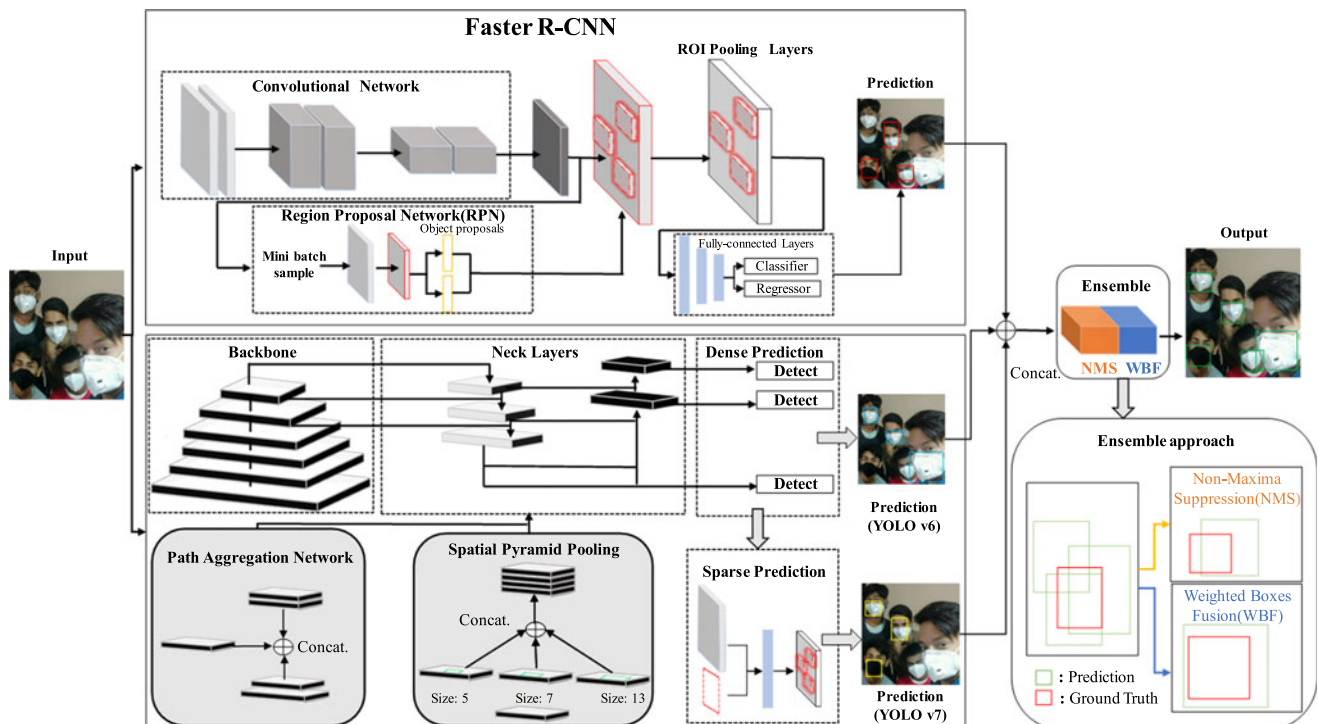
has higher classification accuracy than the independent model, and the positive strategy and the unanimous strategy are favorable for the task with many FNs and the task with many false positives (FPs), respectively. It is beneficial for many tasks. Consensus strategies also provide a compromise between affirmative and unanimous strategies. In stacking, an algorithm takes the outputs of sub-models as input and attempts to learn how to best combine the input predictions to make a better output prediction. Stacking is also known as a stacked generalization and is an extended form of the model averaging ensemble technique in which all sub-models equally participate as per their performance weights and build a new model with better predictions. This new model is stacked up on top of the others. The method of our proposed ensemble model is to use the same dataset to train different models. Methods such as data enhancement in the training project to improve the accuracy of each model and the difference between models are used, and finally fusion output based on the prediction results of each model is performed.

3. Research Methodology

3.1. Overview

In this study, we propose a strategy to detect mask-wearing in images. The proposed model predicts the position of a face in an image and classifies it into three classes, including wearing of mask correctly, no mask, and mask incorrectly. For robust model design, we employed an ensemble strategy and several post-processing. Heterogeneous object detectors are based on Faster R-CNN and YOLO series to complete the detection of mask-wearing. As shown in Figure 1, the pictures are input into different models. On the above Faster-RCNN object detector, through a region proposal network (RPN) and a region-based detection network make object predictions. The following is the

Figure 1 Architecture of ensemble learning-based object detector



prediction result of the model using the YOLO series. First, input the picture into the backbone network of the YOLO series model, and then pass through the Neck layer network. This layer includes Spatial Pyramid Pooling for extracting features of different scales, and a path aggregation network for these features is integrated. In this way, three layers of feature maps with different sizes are obtained, and finally, the features of different scales are aggregated to obtain the prediction results. Then we merge the prediction results from the Faster R-CNN and YOLO series models and filter the repeated and low-score prediction boxes through the ensemble module to obtain the output of the model.

3.2. Heterogeneous object detectors

The corresponding class is the index of the highest value in the array of the corresponding probability where the instance is the probability value for each class. Our ensemble strategy employs a list of different object detectors $L = [D_1, \dots, D_m]$, where m is the number of object detectors. Each detector generates a list of detections $D = [d_1, \dots, d_N]$, where each d_i is given by a triple $[b_i, c_i, p_i]$ that consists of a bounding box (b_i), the corresponding class (c_i), and the corresponding probability (p_i). The bounding box indicating the position of the object is expressed in the form of $[x_1, y_1, x_2, y_2]$. The bounding box indicating the position of the object is expressed in the form of $[x_1, y_1, x_2, y_2]$, where x_1 and y_1 are upper left coordinates and x_2 and y_2 are lower right coordinates. The corresponding class is the index of the highest value in the array of the corresponding probability. In this experiment, four representative deep learning-based object detectors, YOLO v5, YOLO v6, YOLO v7, and Faster R-CNN described in Section 2, are employed.

3.3. Ensemble approaches

The purpose of the ensemble model is to design a robust detector in crowded scenes. Combining prediction results from different models can improve performance by preventing overfitting. For the final fusion of prediction results, it is proposed to use NMS and WBF (Solovyev et al., 2021) simultaneously.

NMS is one of the strategies to solve the problem of generating multiple bounding boxes for each object in the prediction of the detector. In the NMS process, the number of overlapping bounding boxes is minimized by repeating the removal of boxes that overlap a lot with the box with the highest confidence score. The formula for IoU is the criteria for overlapping bounding boxes. The NMS algorithm is effective for detecting a single object, but it has problems such as having multiple bounding boxes in the detection result of one object in a multi-object detection task. The pseudo-code of NMS is described in Algorithm 1.

Algorithm 1. Non-maximum suppression (NMS)

input: $B = \{b_1, \dots, b_N\}$, $C = \{c_1, \dots, c_N\}$, Th

B is the list of initial bounding boxes

C is corresponding confidence scores

Th is threshold

output: Prediction results

Begin:

$D \leftarrow \emptyset$

```

while  $B \neq \emptyset$  do
     $m \leftarrow \text{argmax } S$ 
     $M \leftarrow b_m$ 
     $D \leftarrow D \cup M, B \leftarrow B - M$ 
    for  $b_i$  in  $B$  do
        if  $\text{IoU}(M, b_i) \geq Th$  then
             $B \leftarrow B - b_i$ 
             $C \leftarrow C - c_i$ 
        end
    end
end
return  $D, C$ 
end

```

Algorithm 2. Weighted box fusion (WBF)

input: $B = \{b_1, \dots, b_N\}$, $C = \{c_1, \dots, c_N\}$, Th

B is the list of initial bounding boxes

C is corresponding confidence scores

Th is threshold

output: Prediction results

Begin:

$D \leftarrow \emptyset$

```

while  $B \neq \emptyset$  do
     $m \leftarrow \text{argmax } S$ 
     $M \leftarrow b_m$ 
     $D \leftarrow D \cup M, B \leftarrow B - M$ 
    for  $b_i$  in  $B$  do
        if  $\text{IoU}(M, b_i) \geq Th$  then
             $B \leftarrow \text{weighted\_sum}(B, b_i)$ 
             $C \leftarrow \text{Average}(C - c_i)$ 
        end
    end
end
return  $D, C$ 
end

```

WBF is a strategy that ensembles multiple bounding boxes. If the IoU of two randomly selected bounding boxes exceeds the threshold, the coordinates of the bounding boxes are weighted,

and the optimal bounding box is the output. The pseudo-code of WBF is described in Algorithm 2.

The mixed use of NMS and WBF is suggested in this experiment. Specifically, the NMS and WBF are sequentially applied to the merging of the prediction results of each independent object detector.

4. Experiment

4.1. Dataset

We obtained totally 5763 images of mask correctly, no mask, and mask incorrectly, from the internet. The dataset is divided into two categories including the simple dataset (with less than four persons) and the complex dataset (with more than or equal to four persons). The simple dataset consists of 4748 images with 3 or fewer people in each image, while the complex dataset has 1015 images with 4 or more people in each image. The dataset is further divided into three categories, mask correct, no mask, and mask incorrectly. A mask correctly covers the entire face, including the mouth and nose as shown by the green border box in Figure 2, while a mask incorrectly may leave the nose or mouth exposed as shown by the blue border box in Figure 2. A no-mask category includes images of people without any type of mask. We divide into a training set, validation set, and test set according to the ratio 70%, 20%, and 10%. The simple dataset is used to evaluate simple mask-wearing detection. On the other hand, the complex dataset evaluates the detection ability of crowded scenes and the total of simple and complex datasets. Table 1 shows the details of the simple and

complex datasets. Table 2 shows the number of objects in simple, complex, and total datasets.

4.2. Experimental implementation

An ensemble learning algorithm is an algorithm that efficiently merges the prediction results of each trained independent model. Therefore, in all experiments, we compare the prediction results of the ensemble model with the performance of different detection models trained individually for each model. Our proposed model was rigorously evaluated through a series of experiments. The first is to evaluate the position prediction performance in object detection. How well the model detects objects and is evaluated using test samples not used for training. The second is the evaluation of the classification performance of the detected object. The metrics of classification performance, precision (PR), recall (RC), F1-score (F1), and average precision (AP) are used.

The horizontal direction of the confusion matrix in Table 3 represents the model predictions, and the vertical columns represents the ground truth. The combination of the predicted and true classes allows us to compute four values, namely, true positives, false positives, false negatives, and true negatives, which are key metrics for evaluating model accuracy and performance. True Positive (TP) refers to the number of target objects correctly detected by the model. If the model correctly predicts the object in the image, then this prediction is considered TP. False positive (FP) is wrongly divided into the number of positive cases, that is, the number of actual negative cases but predicted by the model as positive cases. False negative (FN) is wrongly divided into the number of negative cases, that is, the number of actual positive cases but predicted by the model as negative cases. True negative (TN) is correctly divided into the number of negative cases, that is, the actual number of negative cases and the number of cases predicted by the model as negative cases.

PR is a measure of the classifier and is the ability to correctly identify samples. It represents the proportion of correctly predicted samples among the samples identified as positive samples. It is calculated as follows:

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

RC is the ratio of positive cases judged as true in the model to the total positive cases. It is calculated as follows:

$$RC = \frac{TP}{TP+FN}$$

F1 represents the harmonic average evaluation index of precision and recall, which is calculated as follows:

$$F1 = 2 \times \frac{PR \times RC}{P+RC}$$

AP measures the average accuracy of a single category model, which is calculated as follows:

Figure 2
Mask detection data samples



Table 1
The composition of the dataset

Data	Train	Validation	Test	Total
Simple	3323	949	476	4748
Complex	710	203	102	1015
Total	4033	1152	578	5763

Table 2
Number of objects in the dataset

Data	Train			Validation			Test			Total		
	Mask correctly	No mask	Mask incorrectly	Mask correctly	No mask	Mask incorrectly	Mask correctly	No mask	Mask incorrectly	Mask correctly	No mask	Mask incorrectly
Simple	3328	1259	26	914	393	6	470	190	3	4712	1842	35
Complex	3554	1825	67	936	544	18	562	278	5	5052	2647	90
Total	6882	3084	93	1850	937	24	1032	468	8	9764	4489	125

Table 3
Confusion matrix

		True class	
		Positive	Negative
Predicted class	Positive	True positive	False positive
	Negative	False negative	True negative

Table 4
Summary of system environment

Item	Value
CPU	Intel i5-10400
Memory	Memory 32GB
GPU	NVIDIA GeForce RTX 3080
CUDA version	10.2
Python version	3.8
Pytorch version	1.7.1

$$AP = \int_0^1 PR(RC)dr$$

YOLO v5, YOLO v6, YOLO v7, and Faster R-CNN are employed as independent object detectors. All models are pre-trained (Tan et al., 2018) with ImageNet (He et al., 2019). For the simple dataset, complex dataset, and total dataset, we use 70% ratio for fine-tuning, while the remaining 30% of the data is used for validating 20% and testing 10% of the performance of the model. Our image dataset is resized to 640 × 640 for a fixed size. Three hundred epochs are used for training all the models, whereas the learning rate is set to 0.001 for the optimizers. Table 4 is a summary of the experimental environment.

4.3. Result

The proposed methodology is evaluated according to a series of experiments. Four independent object detectors were employed, YOLO v5, YOLO v6, YOLO v7, and Faster R-CNN. We have

presented the results of each model in Tables 5 and 6 for the sample and complex dataset, respectively. For Ensemble 0, we have combined YOLO v6 and YOLO v7 while for Ensemble 1, we have merged Faster R-CNN, YOLO v6, and YOLO v7. Figure 3 shows independent classifier-based prediction results and ensemble-based prediction results. According to the detection results shown in Figure 3, we found that the Faster R-CNN model works well for edge detection. The red box indicates that the mask is not used, the green box indicates that the mask is used correctly, and the blue box indicates that the mask is used incorrectly. Independent object detectors are showing plausible detectability and also contain several misclassifications. In Fast R-CNN, the human forehead is highlighted as a bounding box, and more than one bounding box exists for the same object. YOLO v6 and YOLO v7 have objects that they do not detect, and YOLO v6 does not detect edge objects and blurry objects. YOLO v7 did not detect objects without masks. And YOLO v5 has not detected most of the objects at the edges, including the top and bottom of the image. Ensemble-based object detectors show a high level of detectability. All objects located at the edges are successfully detected.

Figure 4 shows a comparison of detection capabilities according to the ensemble method. In this experiment, the simultaneous use of the proposed ensemble technique, NMS and WBF, is evaluated using Ensemble 1. The merging of the prediction results of each independent model results in the production of many bounding boxes. If only NMS is used, FPs are not deleted, and if only WBF is used, there are still several bounding boxes for one object. On the other hand, a methodology using both algorithms can successfully detect any object.

Most object detectors in performance evaluation experiments using simple datasets tend to have relatively high classification performance (Table 5). This is because it is a relatively simple task, as fewer than four people are in the sample image of the simple dataset. However, in most of the performance perimeters, ensemble-based models outperform independent models. Ensemble 1 is state of the art in classification for all classes.

As shown in Table 6, the performance of the basic model and ensemble-based models shows a slight downward trend. This is

Table 5
Comparison of classification performance of each model on the simple dataset

Model	Mask correctly				No mask				Mask incorrectly				Average			
	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1
Faster R-CNN	0.910	0.923	0.957	0.940	0.881	0.827	0.815	0.821	0.615	0.620	0.781	0.705	0.802	0.790	0.851	0.822
YOLO v5	0.987	0.974	0.953	0.963	0.960	0.973	0.940	0.956	0.303	0.472	0.333	0.402	0.750	0.806	0.742	0.773
YOLO v6	0.974	0.967	0.960	0.964	0.958	0.972	0.917	0.944	0.500	0.436	0.642	0.539	0.810	0.788	0.839	0.815
YOLO v7	0.989	0.977	0.970	0.973	0.968	0.948	0.952	0.950	0.262	0.471	0.333	0.402	0.740	0.793	0.752	0.754
Ensemble 0	0.973	0.989	0.985	0.987	0.986	0.974	0.993	0.983	0.333	0.500	0.333	0.446	0.764	0.821	0.770	0.805
Ensemble 1	0.995	0.996	0.999	0.997	0.995	0.997	0.995	0.996	0.610	0.997	0.562	0.779	0.866	0.996	0.852	0.925

Table 6
Comparison of classification performance of each model on the complex dataset

Model	Mask correctly				No mask				Mask incorrectly				Average			
	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1
Faster R-CNN	0.910	0.923	0.957	0.940	0.886	0.714	0.909	0.811	0.408	0.727	0.500	0.613	0.734	0.787	0.786	0.786
YOLO v5	0.987	0.974	0.953	0.963	0.973	0.958	0.917	0.937	0.597	0.674	0.600	0.637	0.852	0.865	0.823	0.845
YOLO v6	0.945	0.967	0.939	0.953	0.986	0.928	0.925	0.926	0.580	0.680	0.500	0.590	0.837	0.858	0.787	0.823
YOLO v7	0.989	0.977	0.970	0.973	0.978	0.966	0.931	0.948	0.716	0.648	0.600	0.624	0.890	0.863	0.833	0.848
Ensemble 0	0.975	0.913	0.982	0.947	0.953	0.901	0.964	0.932	0.600	0.600	0.600	0.600	0.842	0.804	0.848	0.826
Ensemble 1	0.973	0.929	0.940	0.934	0.968	0.933	0.941	0.937	0.610	1.000	0.564	0.782	0.850	0.954	0.815	0.884

Figure 3

Each independent object detector uses total data training, as follows: prediction results of each independent object detectors for a crowded scene

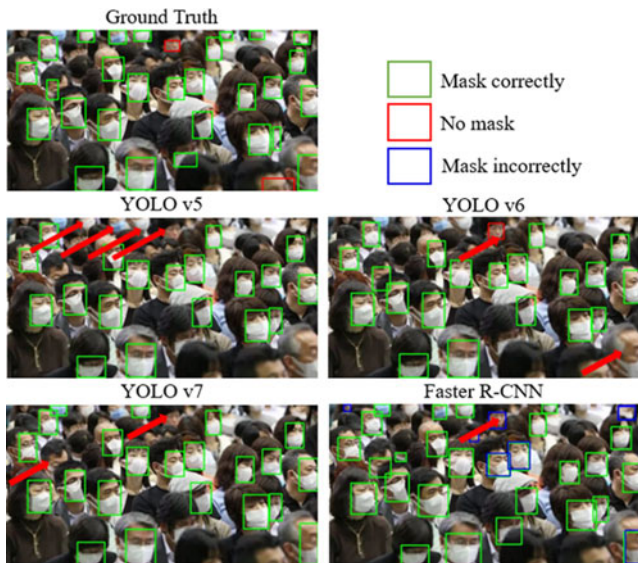


Figure 4

Comparison of detection capabilities of different integration methods after Ensemble 1 is trained with total data

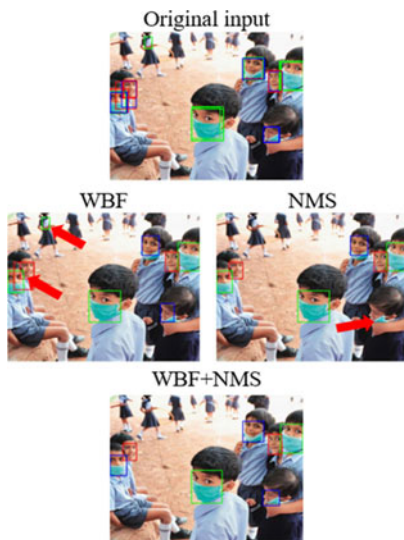
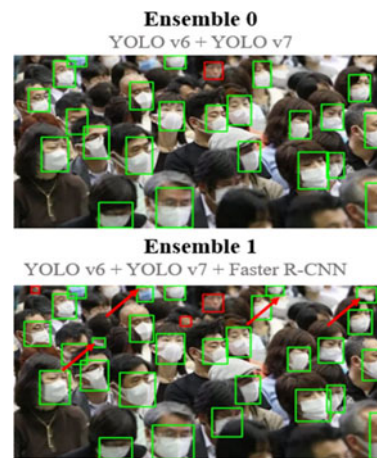


Figure 5

Prediction results of each object detector based on integrated learning in the crowd scene in the total dataset



caused by relatively complex datasets compared to simple datasets. The ensemble model we proposed in complex datasets is not as effective as YOLO v7. It may be that the ensemble model has a lot of image edge information and some blurred objects because these objects were not labeled when the dataset was labeled. Nonetheless, our proposed ensemble-based model outperforms standalone models on total datasets.

Table 7 shows the performance of the Ensemble 1 model on the total dataset. Compared with the basic model, it achieved the highest F1 score in each category.

Figure 5 shows the prediction results of our ensemble model. In Ensemble 0, YOLO v6 and YOLO v7 are used as the prediction results of the basic model. For edge objects, there is a FN. Ensemble 1 added Faster R-CNN model. The result shows that there are detection errors, but the detection performance of the model for edge objects is improved. Figure 6 shows the prediction results obtained by the Ensemble 1 model after training with the total dataset. From the results, it can be concluded that our proposed ensemble model can perform well in the mask-wearing detection task for blurred objects (a), head occlusion area <50% (b), head occlusion area >50% (c), for small object detection (d), a great improvement.

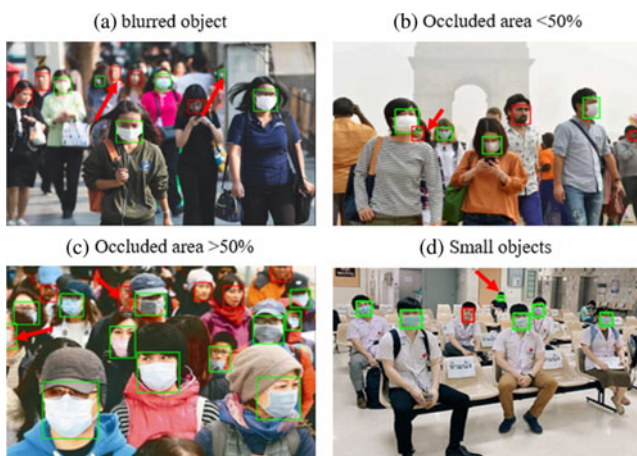
5. Discussion

This paper proposes a simple and efficient ensemble model for mask-wearing detection tasks. Different from previous mask-wearing detection tasks, we not only include the samples of the

Table 7 Comparison of classification performance of each model on the total dataset

Model	Mask correctly				No mask				Mask incorrectly				Mean			
	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1	AP	PR	RC	F1
Faster R-CNN	0.957	0.907	0.956	0.930	0.807	0.740	0.800	0.770	0.167	0.340	0.167	0.253	0.643	0.662	0.641	0.651
YOLO v5	0.974	0.954	0.923	0.938	0.967	0.954	0.923	0.938	0.406	0.625	0.500	0.562	0.782	0.844	0.784	0.815
YOLO v6	0.976	0.951	0.955	0.953	0.913	0.978	0.961	0.960	0.400	0.667	0.500	0.570	0.763	0.865	0.805	0.827
YOLO v7	0.969	0.962	0.931	0.946	0.969	0.956	0.936	0.846	0.523	0.606	0.500	0.553	0.820	0.841	0.789	0.781
Ensemble 0	0.973	0.911	0.985	0.950	0.975	0.926	0.978	0.950	0.450	0.571	0.500	0.530	0.799	0.802	0.821	0.810
Ensemble 1	0.964	0.957	0.959	0.958	0.978	0.916	0.974	0.968	0.603	0.682	0.664	0.790	0.848	0.850	0.865	0.905

Figure 6
Problems in mask wearing detection predicted by Ensemble 1 model



mask-wearing and not wearing but also added the cases of whether the mask is correctly or incorrectly worn. In addition, we use NMS and WBF in the merged segment. NMS is responsible for screening out predictions with low prediction scores, and the WBF algorithm gives different weights to different models. Finally, the performance of the model is tested using two different datasets. The experimental results show that the ensemble model has achieved significant results on both datasets. In the complex dataset, performance of both independent detectors and ensemble-based models showed a slightly decreasing trend. This is caused by a relatively complex dataset compared to the simple dataset. Nevertheless, our proposed ensemble-based models generally outperform independent models. When total data are used, it still performs best. However, the model is not satisfied with the detection result of the mask worn incorrectly category, which may be due to the low prediction score of the mask incorrectly category using the basic model.

6. Conclusion

In summary, we proposed an ensemble-based object detector for robust mask-wearing detection in crowd scenes. The fusion detection algorithm designed in this paper is based on the bounding box and confidence score of each model. The prediction results of independent object detectors are merged by NMS and WBF. Through the collected dataset and statistical techniques, our model is strictly flattened. The empirical results show that our model fusion method successfully performed mask-wearing even on high-complexity images, and the detection ability for closure and edges was also improved. In the future, we would like to explore that it cannot only detect the wearing of the pedestrian mask but also track the movement of pedestrians.

Acknowledgement

This research was supported by “Regional Innovation Strategy (RIS)” through the National Research Foundation of Korea (NRF), funded by the Ministry of Education [2021RIS-001(1345341783)].

Conflicts of Interest

The authors declare that they have no conflicts of interest to this work.

References

- Akshatha, K. R., Karunakar, A. K., Shenoy, S. B., Pai, A. K., Nagaraj, N. H., & Rohatgi, S. S. (2022). Human detection in aerial thermal images using faster R-CNN and SSD algorithms. *Electronics*, *11*(7), 1151. <https://doi.org/10.3390/electronics11071151>
- Casado-García, Á., & Heras, J. (2020). *Ensemble methods for object detection*. Netherlands: IOS Press.
- Chen, R. C. (2019). Automatic license plate recognition via sliding-window darknet-YOLO deep learning. *Image and Vision Computing*, *87*, 47–56. <https://doi.org/10.1016/j.imavis.2019.04.007>
- Cybert, S., & Czyżewski, A. (2022). Robust object detection with multi-input multi-output faster R-CNN. In *International Conference on Image Analysis and Processing*, 572–583. https://doi.org/10.1007/978-3-031-06427-2_48
- Degadwala, S., Vyas, D., Chakraborty, U., Dider, A. R., & Biswas, H. (2021). Yolo-v4 deep learning model for medical face mask detection. In *2021 International Conference on Artificial Intelligence and Smart Systems*, 209–213. <https://doi.org/10.1109/ICAIS50930.2021.9395857>
- Ding, X., Chen, H., Zhang, X., Huang, K., Han, J., & Ding, G. (2022). Re-parameterizing your optimizers rather than architectures. *arXiv Preprint: 2205.15242*. <https://doi.org/10.48550/arXiv.2205.15242>
- Ding, X., Xia, C., Zhang, X., Chu, X., Han, J., & Ding, G. (2021). Repmlp: Re-parameterizing convolutions into fully-connected layers for image recognition. *arXiv Preprint: 2105.01883*. <https://doi.org/10.48550/arXiv.2105.01883>
- Fang, B., Jiang, M., Shen, J., & Stenger, B. (2022). Deep generative inpainting with comparative sample augmentation. *Journal of Computational and Cognitive Engineering*, *1*(4), 174–180. <https://doi.org/10.47852/bonviewJCCE2202319>
- Gwak, J., Sim, K. M., & Jeon, M. (2014). Novel dynamic diversity controlling EAs for coevolving optimal negotiation strategies. *Information Sciences*, *273*, 1–32. <https://doi.org/10.1016/j.ins.2014.02.153>
- He, K., Girshick, R., & Dollár, P. (2019). Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4918–4927.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . , & Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv Preprint: 1704.04861*. <https://doi.org/10.48550/arXiv.1704.04861>
- Kang, J., & Gwak, J. (2020). Ensemble learning of lightweight deep learning models using knowledge distillation for image classification. *Mathematics*, *8*(10), 1652. <https://doi.org/10.3390/math8101652>
- Li, C., Li, L., Jiang, H., Weng, K., Geng, Y., Li, L., . . . , & Wei, X. (2022). YOLOv6: A single-stage object detection framework for industrial applications. *arXiv Preprint: 2209.02976*. <https://doi.org/10.48550/arXiv.2209.02976>
- Neubeck, A., & Van Gool, L. (2006). Efficient non-maximum suppression. In *18th International Conference on Pattern Recognition*, 850–855. <https://doi.org/10.1109/ICPR.2006.479>
- Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 779–788. <https://doi.org/10.1109/CVPR.2016.91>

- Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster R-NN: Towards real-time object detection with region proposal networks. *Advances in Neural Information Processing Systems*, 28.
- Ren, X., & Liu, X. (2020). Mask wearing detection based on YOLOv3. In *Journal of Physics: Conference Series*, 1678(1), 12089. <https://doi.org/10.1088/1742-6596/1678/1/012089>
- Sharma, V. (2020). *Face mask detection using yolov5 for COVID-19*. Doctoral Dissertation, California State University San Marcos.
- Solovyev, R., Wang, W., & Gabruseva, T. (2021). Weighted boxes fusion: Ensembling boxes from different object detection models. *Image and Vision Computing*, 107, 104117. <https://doi.org/10.1016/j.imavis.2021.104117>
- Suresh, K., Palangappa, M. B., & Bhuvan, S. (2021). Face mask detection by using optimistic convolutional neural network. In *2021 6th International Conference on Inventive Computation Technologies*, 1084–1089. <https://doi.org/10.1109/ICICT50816.2021.9358653>
- Tan, C., Sun, F., Kong, T., Zhang, W., Yang, C., & Liu, C. (2018). A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, 270–279. https://doi.org/10.1007/978-3-030-01424-7_27
- Vilhelm, A., Limbert, M., Audebert, C., & Ceillier, T. (2022). Ensemble Learning techniques for object detection in high-resolution satellite images. *arXiv Preprint: 2202.10554*. <https://doi.org/10.48550/arXiv.2202.10554>
- Wang, C. Y., Bochkovskiy, A., & Liao, H. Y. M. (2022). YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv Preprint: 2207.02696*. <https://doi.org/10.3390/s17030617>
- Yang, E., Gwak, J., & Jeon, M. (2017). Conditional random field (CRF)-boosting: Constructing a robust online hybrid boosting multiple object tracker facilitated by CRF learning. *Sensors*, 17(3), 617.
- Yang, M. (2022). Research on vehicle automatic driving target perception technology based on improved MSRPN algorithm. *Journal of Computational and Cognitive Engineering*, 1(3), 147–151. <https://doi.org/10.47852/bonviewJCCE20514>
- Yang, Y., & Song, X. (2022). Research on face intelligent perception technology integrating deep learning under different illumination intensities. *Journal of Computational and Cognitive Engineering*, 1(1), 32–36. <https://doi.org/10.47852/bonviewJCCE19919>
- Yun, S., Han, D., Oh, S. J., Chun, S., Choe, J., & Yoo, Y. (2019). Cutmix: Regularization strategy to train strong classifiers with localizable features. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 6023–6032. <https://doi.org/10.1109/ICCV.2019.00612>
- Zeng, G., Yu, W., Wang, R., & Lin, A. (2021). Research on mosaic image data enhancement for overlapping ship targets. *arXiv Preprint: 2105.05090*. <https://doi.org/10.48550/arXiv.2105.05090>

How to Cite: Hu, X., Jeon, Y., & Gwak, J. (2023). Heterogeneous Ensemble Approaches for Robust Face Mask Detection in Crowd Scenes. *Journal of Computational and Cognitive Engineering*, 2(4), 343–351, <https://doi.org/10.47852/bonviewJCCE3202478>