Enhancement of Human Face Mask Detection Performance by Using Ensemble Learning Models

Xinyi Gao, Minh Nguyen, and Wei Qi Yan

Auckland University of Technology, Auckland 1010 New Zealand

Abstract. Given the prevalence of worldwide pandemics, the need of adhering to appropriate mask use becomes more paramount. Therefore, the importance of developing a human face mask detection model that is both efficient and accurate cannot be overstated. Nevertheless, there is a need for additional enhancement in the accuracy and efficiency of mask detection algorithms, particularly in dealing with increasingly complex scenarios. In this paper, we make a valuable contribution to the current literature by utilizing Swin Transformer model to address mask detection challenges. The Swin Transformer, an innovative deep learning architecture, has shown remarkable effectiveness in computer vision applications. The main aim of our research work is to assess the efficacy of the Swin Transformer in improving precision and efficiency of mask detection. Our methodology includes the careful selection of datasets, design of model architecture, and implementation of experimental settings. The test results show that our suggested model, Swin+YOLOv8, surpassed the baseline models in terms of accuracy and mean average precision (mAP). The research outcomes of this paper will facilitate the advancement of general object detection and make a valuable contribution to the improvement of public health and safety.

Keywords: Human Face Mask Detection · Swin Transformer · YOLOv8 · Deep Learning.

1 Introduction

In current worldwide pandemic, the use of masks has become imperative in public settings, serving a crucial function in safeguarding people's well-being. The development of accurate and efficient mask detection is a significant problem, particularly in diverse contexts such as surveillance recordings in crowded public areas, individual or group photographs, and social media platforms. Given the context, the primary objective of this paper is to construct a proficient mask detection model using deep learning [25].

Currently, there exists a number of conventional machine learning methods as well as deep learning approaches for the purpose of mask detection [5]. These include feature-based methods, convolutional neural networks (CNN) [9], and recurrent neural networks (RNN) [1]. Nevertheless, there is a need for enhancing

the accuracy and efficiency of mask detection algorithms. Recently, there has been a notable surge in the interest of academics towards Transformer models, particularly Swin Transformer [13], owing to its exceptional efficacy in a diversity of computer vision applications. The advent of Swin Transformer presents a novel approach to address visual problems, including image classification and object recognition. In contrast to conventional convolutional neural networks, it exhibits notable benefits in resolving long-range problems.

The primary contribution of this paper is in the use of Swin Transformer model for the purpose of conducting mask detection, leveraging its window partitioning and self-attention mechanism. The objective of this paper is to assess the effectiveness of the model in enhancing the accuracy and efficiency of mask detection. Additionally, this research project aims to provide a novel solution for the practical application of mask recognition. The outcomes of our study will facilitate the progress of mask detection technology and make a valuable contribution to enhancing public health safety.

The following sections provide a comprehensive account of our study methodologies and the detailed findings obtained from our experiments. In this paper, we will commence by conducting a comprehensive examination of pertinent studies on mask detection and the fundamental principles underlying the Swin Transformer. Subsequently, we will present our approach, encompassing the selection of datasets, model architecture, and experimental configurations. Following this, we will meticulously discuss our experimental findings, including a detailed comparison and analysis of model performance. Ultimately, we will summarise our contributions and deliberate on potential avenues for future research.

2 Related work

Deep learning algorithms [26] have been the dominant approaches in contemporary mask recognition challenges. There are two often adopted deep learning methods [6]. The two-stage target detection paradigm encompasses two distinct stages: feature extraction and feature classification, effectively partitioning the target detection process, such as Faster R-CNN [14]. Another approach is a single-stage object detection model. This model has the capability to immediately derive classification outcomes using the regression approach, hence enabling real-time detection, such as You Only Look Once (YOLO) [3].

Ren et al. reviewed the identification of mask-wearing using the YOLOv3 algorithm [16]. The work proposed an enhanced Face_Mask Net identification approach based on a convolutional neural network (CNN) to address the labor-intensive task of manually finding masks. In the work, enhancements were made to the non-maximum Suppression (NMS) module of YOLOv3 [10]. The Distance-IoU (DIoU) metric is proposed as a replacement for the popular Intersection over Union (IoU) metric in Non-Maximum Suppression (NMS) algorithms. The use of the K-Means algorithm aims to optimize anchor boxes and enhance the accuracy of object identification. The training and testing procedures are conducted by using the self-collected Face_Mask dataset. The experimental findings of the

Face_Mask Net model demonstrate its efficacy in detecting the presence of masks on individuals, with a great level of accuracy compared to the pre-trained networks.

Ye et al. discussed a mask-wearing detection algorithm based on an improved YOLOv4 network [27]. The improved algorithm integrates the CBAM attention mechanism and depthwise over-parameterized convolution (DO-Conv) to enhance accuracy and reduce the number of parameters. The experimental results using a dataset of approximately 4000 images show that the improved algorithm achieves significantly higher recognition accuracy than the original algorithm and algorithm outperforms current mainstream algorithms in terms of recognition accuracy.

Yu et al. utilized YOLOv4 model to accurately identify and classify face masks, as well as determine if they are being worn correctly according to established guidelines [29]. The study focuses on the challenges posed by intricate surroundings, including poor precision, real-time performance, and resilience. The experimental findings indicate that the algorithm attains a mean average accuracy (mAP) of 98.3% and exhibits a notable frame rate of 54.57 FPS.

The YOLOv5+CBD method was built on an enhanced iteration of the YOLOv5 model [8]. This approach aims to tackle several issues encountered in computer vision, including occlusion, dense targets, and small-scale objects. The proposed approach integrates many techniques to enhance the accuracy of object recognition. These techniques include the use of the Coordinate Attention mechanism, the incorporation of a weighted bidirectional feature pyramid network, and the implementation of Distance Intersection over Union with Non-Maximum Suppression. The experimental findings demonstrated that the YOLOv5+CBD model attains a detection accuracy of 96.7%, exhibiting a notable enhancement of 2.1% in comparison to the baseline model.

Wang et al. proposed a face mask-wearing detection model based on a loss function and attention mechanism [21]. An attention mechanism was integrated in the feature fusion process to improve feature utilization and explore different attention mechanisms to enhance deep network models. The impact of different bounding box loss functions was investigated on mask-wearing recognition. The model achieved a mean average precision (mAP) of 90.96% on a dataset of mask-wearing images, outperforming traditional deep learning methods.

Wang et al. introduced a new mask-wearing detection model called YOLOv7-CPCSDSA [20]. This model combined YOLOv7 base model with the CPC structure, SD structure, and SA mechanism. The CPC structure reduces computational redundancy and improves memory access, while the SD structure enhances the detection of small targets. The SA mechanism focuses on important local information, further improving accuracy. Comparative and ablation experiments using a mask dataset validate the effectiveness of the YOLOv7-CPCSDSA model. The results show that the model achieves higher mean average precision compared to YOLOv7 and meets real-time detection requirements [22].

Deng et al. proposed an enhanced mask-wearing inspection algorithm based on the single shot multibox detector (SSD) algorithm [2]. The algorithm incor-

porated with inverse convolution, feature fusion, and attention mechanisms to improve the accuracy of mask-wearing detection. A dataset with 3,656 manually labeled images was created for training the network. The experimental results demonstrate that the algorithm has good accuracy for mask-wearing inspection, with an average accuracy of 91.7%.

Jesús et al. examined the identification of improper face mask use via the utilization of convolutional neural networks (CNNs) with transfer learning. The machine learning methods, namely convolutional neural networks [17] were employed. A comprehensive analysis of the difficulties encountered in constructing a training dataset for the given task was conducted, while also providing a thorough examination of the existing literature pertaining to artificial intelligence (AI) technological approaches. A comprehensive overview of initiatives that have devised AI-enabled systems for the purpose of mask identification was taken, exhibiting diverse degrees of precision and efficacy.

Xue et al. presented an intelligent detection and recognition system for mask-wearing based on an improved RetinaFace algorithm [24]. It consists of a face mask detection algorithm, a mask standard wearing detection algorithm, and a face recognition algorithm. The system utilizes the improved RetinaFace algorithm for real-time detection of mask-wearing and identification of proper mask usage. It also incorporates a voice prompt module to assist in the functionality of the system. The system has been tested and proven effective in achieving its purpose of face mask detection and recognition.

Ullah et al. described a novel algorithm for mask detection and recognizing human actions during the COVID-19 pandemic [18]. The proposed method for detecting face masks uses the Mask R-CNN ROI wrapping with the Resnet-152 algorithm and evaluates the model using Apache MXNet. The article also emphasizes the need for developing AI, IoT, big data, and machine learning technologies.

Swin Transformer was introduced as a general-purpose backbone for computer vision in 2021 [13], which addresses the challenges of adapting the Transformer from language to vision by proposing a hierarchical Transformer with Shifted windows. The Swin Transformer outperforms previous state-of-the-art models in ImageNet 1K image classification, visual object detection, and semantic segmentation tasks. Due to the superiority of Swin transformer in the field of vision, more and more researchers apply Swin transformer to vision tasks. Ye et al proposed to solve the difficulty of simultaneously completing masked face detection [19] and recognition tasks by enhancing the performance of Swin Transformer in the field of face feature extraction [28].

Zeng et al. successfully proposed a novel framework called Swin+CasUNet based on Swin Transformer for the restoration of masked faces [31]. Previous studies have acknowledged pain expression recognition by considering the whole face, and Yuan et al. used the Swin Transformer model to recognize pain intensity by recognizing the whole face [30]. It is becoming more and more common to use Swin Transformer for vision tasks.

3 Methodology

In this paper, we present a novel deep learning model for human face mask detection, which is built on the integration of YOLOv8 [23] and Swin Transformer. The objective of our paper is to identify facial features in photographs and ascertain the presence or absence of masks on individuals. The model structure of the Swin Transformer [15] is adopted in our approach. The recognition step use the detection module of YOLOv8. The network architecture in our paper has two main components: A backbone network responsible for extracting features, and a head network dedicated to making predictions.

3.1 Backbone

The backbone component is employed for the purpose of feature extraction. It is comprised of many PatchEmbed, SwinStage, and PatchMerging modules [13]. The first step is the use of the PatchEmbed module to partition the input picture into patches with dimensions of 4×4. Subsequently, these patches are encoded. Subsequently, the SwinStage module proceeds to extract the characteristics of these compact units by using the self-attention process. The PatchMerging module is responsible for the consolidation of neighbouring tiny blocks, therefore reducing the computational complexity of future computations and enhancing the model's capacity to accurately identify bigger items. As shown in Figure 1, the aforementioned procedure undergoes numerous iterations inside the SwinStage and PatchMerging modules.

3.2 Head

The primary function of the head network is to convert the extracted characteristics obtained from the backbone network into the ultimate prediction outcome. Initially, by using a sequence of upsampling and connecting procedures, we integrate characteristics of varying sizes. The integration of these fusion operations allows our model to effectively process objects of varying sizes concurrently. Next, the Conv module is applied to do a convolution operation in order to extract more features. Ultimately, by use of the Detect module, the model generates the predicted outcomes for each category, along with the matching bounding box. The comprehensive network architecture is shown in Figure 1.

The YOLOv8 model was included in our methodology for the purpose of mask detection. Initially, the input picture undergoes a preprocessing stage. The input picture is transformed into a PyTorch tensor and the pixel values are normalised from the range of $0 \sim 255$ to the range of $0.0 \sim 1.0$. Subsequently, the YOLO model is proffered to provide predictions on the preprocessed images. The outcome of the prediction is a collection of object candidate boxes, with each box carrying the associated category and a measure of confidence.

Following the acquisition of the predicted outcomes, postprocessing was conducted based on the obtained findings. The Non-Maximum Suppression (NMS) method is propounded to eliminate object candidate boxes that overlap and to

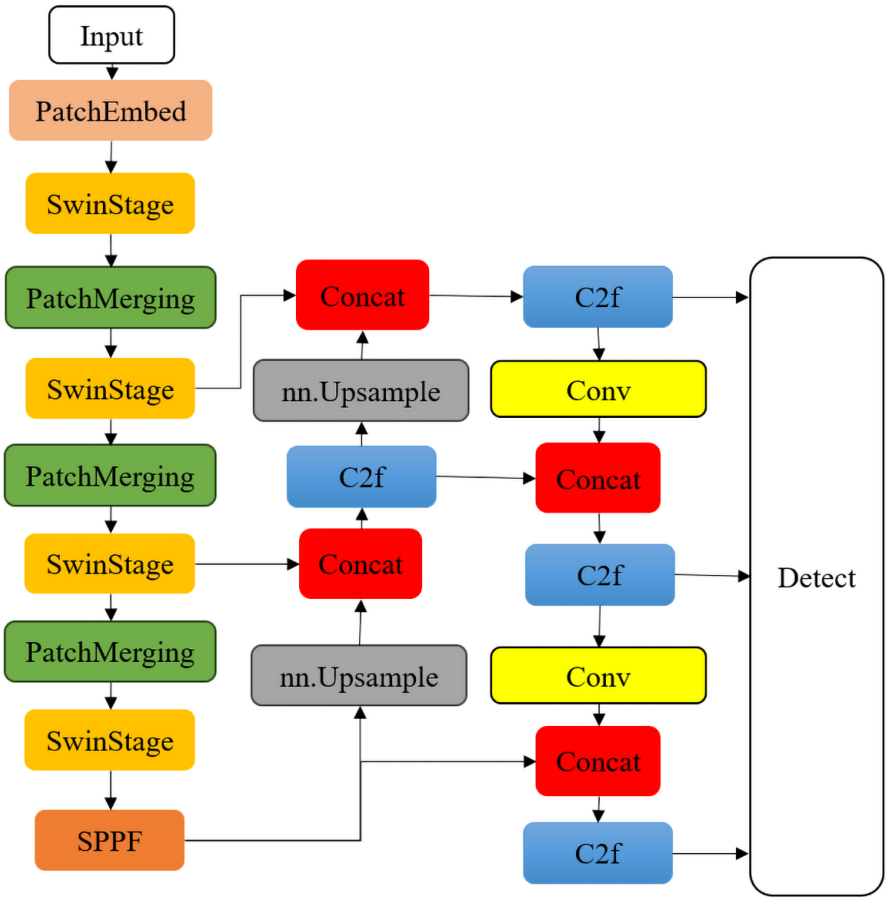

Fig. 1. Structure of Swin+YOLO model

maintain the box with the greatest confidence [7]. Subsequently, the coordinates of the remaining object recommendations are converted from the scale of the input picture to the scale of the original image.

NMS is a populat post-processing approach in the field of object detection. The primary objective of NMS is to address the problem of redundant detection boxes that arise from the prediction of numerous bounding boxes for the same item, frequently exhibiting substantial overlap. If left unattended, there is a possibility that the same item might be identified many times, which would subsequently lead to a decline in the quality of the detection results. Therefore, it is essential to develop a methodology that can effectively remove these intersecting candidate boxes, while preserving just the most ideal one.

The use of NMS offers a viable resolution to this issue. The system functions by arranging the candidate bounding boxes in accordance with their respective

confidence ratings. The bounding box exhibiting the largest confidence score is selected and any other bounding boxes with substantial overlap are removed. Confidence scores are determined from Intersection over Union (IoU) metrics that exceed a certain threshold. The aforementioned procedure is iterated until all candidate boxes have been processed. The calculation of the IoU between two bounding boxes, denoted as A and B, is listed as follows.

$$IoU(A, B) = \frac{|A \cap B|}{|A \cup B|} \tag{1}$$

where $|A \cap B|$ denotes the area of overlap between A and B, and $|A \cup B|$ denotes the area encompassed by A and B. As a consequence of this procedure, a singular bounding box is preserved for each item, namely the one with the greatest confidence score out of all potential options. This greatly improves the accuracy and reliability of the detection outcomes.

In summary, we instantiate a Results object that encompasses the original picture, image URL, category name, and information about the object contender box. In general, our approach utilises the robust detection capabilities of the YOLO model and incorporates preprocessing and postprocessing techniques, facilitating the direct application of the model to mask detection problems.

3.3 Model Training

In this paper, the mask detection dataset was harnessed, including 853 images encompassing a collective count of 4072 faces [11]. The dataset contains facial images that have been labelled with one of three distinct labels: "with_mask", "without_mask", or "mask_weared_incorrect". The dataset presented below offers an extensive collection of instances suitable for training our models, including a wide range of scenarios that depict the many ways in which masks are either worn or not worn.

The face mask detection dataset was partitioned into training, validation, and test sets by using a random splitting method. The training set comprises 80% of the whole dataset, while the validation set is 5% and the test set represents 15%. During the training process, it is essential to continuously check the loss and accuracy metrics on the validation set in order to mitigate the risk of overfitting. Overfitting is a phenomenon that arises when a model exhibits high performance on the training dataset, but fails to generalise well to unknown data. This discrepancy suggests that the model has excessively focused on memorising the training data, rather than acquiring the ability to generalise from it.

During the course of the studies, the models are trained, validated, and tested on Google Colab using Tesla T4 GPU. The model under consideration has been trained for a total of 100 epochs using the dataset specifically designed for mask detection. The training process was conducted using a batch size of 8. An epoch signifies a whole iteration throughout the entirety of the dataset. Every epoch is comprised of a forwards pass and a backwards pass. After the training was completed, we evaluated the model's performance on a separate test set that the

model had not seen during training. This allowed us to assess the model's ability to generalize to unseen data, which is crucial for its practical application.

During the process of training the model, the cross-entropy loss function was created. The cross-entropy loss function is widely employed as a means of measuring the predictive performance of classification issues [4]. The metric quantifies the disparity between the probability distribution projected by the model and the observed probability distribution. The mathematical formulation for classification issues may be expressed as follows.

$$Loss_{CEL} = -\frac{1}{N}\sum_{i=1}^{N} y_i \log(p_i) + (1 - y_i) \log(1 - p_i) \tag{2}$$

where N represents the total number of samples, y_i is the actual label of the i sample (either 0 or 1), and p_i represents the probability assigned by the model to the i sample belonging to the positive class.

The cross-entropy loss function is adopted in our mask detection job to optimise the predictive performance of the model. The model provides a probability value for each picture, indicating the likelihood of the individual in the image wearing a mask. The projected probability value is compared to the actual label, which indicates whether the individual is wearing a mask, in order to compute the cross-entropy loss. During the training phase, the objective is to minimise the aforementioned loss.

In order to enhance the performance of the model, we take use of the Adam optimizer, which is a kind of adaptive learning rate optimisation technique specifically developed to mitigate the issues of gradient sparsity and noise that may arise during the training phase. Furthermore, we implemented an early stopping technique, whereby the training process is halted when the loss on the validation dataset fails to exhibit a drop across a consecutive number of epochs. Upon the completion of the training process, we proceeded to save the model that exhibited the highest performance on the validation set.

By using this approach, our model demonstrates proficiency in accurately detecting the presence of individuals in the picture and determining if they are wearing facial masks. The performance of our model on the test set demonstrates its superiority over other current mask detection methods.

4 Experimental Results and Analysis

4.1 Evaluation Index

Mean Average Precision Mean Average Precision (mAP) [12] is a metric to determine the performance of an object detection algorithm. mAP is the average of multiple class average precision (AP), where the AP for each class is calculated from the detection results of that class. The specific formula is as follows,

$$mAP \ = \ \frac{1}{C}\sum_{i=1}^{N} \mathrm{AP}_i \tag{3}$$

where C represents the total number of classes. where mAP is one of the most popular indicators in object detection, and it is usually utilized to evaluate the accuracy and reliability of object detection algorithms. In order to better evaluate the performance, we use mAP50 and mAP50-95 for evaluation. mAP50 indicates the mAP value when IoU is 0.5. $mAP50-95$ indicates the mAP value when IoU is 0.5$\sim$ 0.95.

Precision-Recall Curve The precision-recall curve (PR-cure) is the curve of the model for the confidence score threshold element. The horizontal axis is the recall rate, indicating the proportion of detected positive samples to all positive samples. The recall formula is,

$$Precision = \frac{TruePositives}{TruePositives + FalsePositives} \tag{4}$$

The vertical axis is precision, which represents the proportion of the true correct number of detected positive samples to the total number of detected samples. The exact formula is,

$$Recall = \frac{TruePositives}{TruePositives + FalseNegatives} \tag{5}$$

In visual object detection, mAP is precisely determined by calculating the area under the precision-recall curve. That is to say, the larger the area under PR-cure, the larger the mAP value, and the better the performance of the model.

4.2 Experimental results and analysis

Experimental results To accurately assess the variation in detection speed across various models, we conducted a comparative analysis by subjecting two baseline models, namely YOLOv7 and YOLOv8, to identical data set conditions for detection purposes. The results are shown in Table 1.

Table 1. Comparison of detection speed between YOLOv7 and YOLOv8

Model	Total time	Average time	Number of detected images	GPU
YOLOv8n	464.2ms	42.2ms	11	Tesla T4
YOLOv7	12060ms	1096ms	11	Tesla T4

Based on the facts shown in Table 1, it is evident that the experimental conditions and inference dataset remain consistent. The YOLOv8n model exhibits much quicker reasoning speed compared to the YOLOv7 model. The inference speed of YOLOv8n is 11596ms greater than that of YOLOv7. On average, there is a speed improvement of 1053 milliseconds per image.

While it has been shown that the model using YOLOv7 is comparatively less efficient than the model employing YOLOv8, for the purpose of enhancing

the reliability of the experimental findings, we have opted to choose a model based on YOLOv7 for the comparative experiment. To enhance the assessment of the model's performance, we conducted sequential training on the dataset using YOLOv7+CBAM, YOLOV8n, YOLOV8n+DCNv2, and Swin+YOLOv8. Table 2 displays the parameters pertaining to the four sets of models.

Table 2. Comparison of four models

Model	Layers	Parameters	FLOPs	Gradients
YOLOv7+CBAM	415	37207344	105.1	37207344
YOLOv8n	225	3011433	8.2	3011417
YOLOV8n+DCNv2	225	3167863	7.7	3167847
Swin+YOLOv8	348	51382394	455.5	51382378

Based on the data presented in Table 2, it is evident that there are significant variations in several parameters across different models. In terms of the network layer count, YOLOv8 exhibits a smaller number of network layers compared to YOLOv7. The YOLOv8 model using SwinTransformer exhibits a greater number of layers compared to the YOLOv8 model without SwinTransformer. From a parameterization standpoint, it can be seen that the SwinTransformer model has a greater overall parameter count compared to other models. Based on the number of parameters, it can be initially deduced that the training duration of the SwinTransformer model is expected to be the longest.

Based on the obtained training outcomes, the mask detection accuracy achieved using the YOLOv7+CBAM model is 92%. The recall rate, which measures the proportion of true positive instances correctly identified, is 0.84. The mean average precision at a threshold of 50% (mAP50) is 0.90, while the mean average precision throughout a range of thresholds from 50% to 95% (mAP50-95) is 0.609. The total duration of the training process amounts to 2.260 hours. The mask detection accuracy achieved by using the YOLOv8n model is 91.7% in terms of precision. The recall rate is measured at 0.823, while the mean Average Precision at mAP50 stands at 0.903. Furthermore, the mean Average Precision throughout the range of mAP50-95 is calculated to be 0.663. Lastly, the training process for this model requires 2.341 hours. The YOLOV8n+DCNv2 model achieves a precision of 94.4% in detecting masks. The recall rate is 0.843, indicating the model's ability to accurately identify positive instances. The mean Average Precision at mAP50 is 0.909, reflecting the model's performance in object identification. The mAP50-95, which measures the average accuracy across different overlap thresholds, is 0.668. The training process for this model is 2.478 hours. The Swin+YOLOv8 model achieved a mask detection accuracy of 96.1%. The recall rate is measured at 0.906, while the mAP50 and mAP50-95 scores are reported as 0.962 and 0.727, respectively. The training process for this model takes around 3.999 hours. Table 3 is derived from the training outcomes of each model.

Table 3. The training results for YOLOv8n, YOLOV8n+DCNv2, YOLOv8s, DSOTAs

Model Name	Precision	Recall	mAP50	mAP50-95	Training Time (Hour)
YOLOv7+CBAM	0.920	0.840	0.903	0.609	2.260
YOLOV8n	0.917	0.823	0.903	0.663	2.341
YOLOV8n+DCNv2	0.944	0.824	0.909	0.668	2.478
Swin+YOLOv8	0.961	0.906	0.962	0.727	3.999

The disparity in accuracy often elicits an intuitive perception of the variance in model performance. The data illustrates that the training speed of the three models using YOLOv8n is comparatively slower in comparison to YOLOv7. Nevertheless, the detection accuracy of the three models using YOLOv8 is consistently high. In particular, the YOLOV8n+DCNv2 model demonstrates a 2.4% increase in accuracy compared to the YOLOv7+CBAM model. The precision of Swin+YOLOv8 surpassed that of YOLOv7+CBAM by 4.1%. Moreover, it has been seen that models using Swin+YOLOv8 or YOLOV8n+DCNv2 exhibit superior performance compared to models that do not use the Transformer architecture. The model's accuracy, while using YOLOV8n+DCNv2, achieves a level of 94.4%. The Swin+YOLOv8 model achieves a model accuracy of 96.1%, surpassing the other three models and attaining the greatest accuracy.

The duration required for training a model using a Swin Transformer architecture exceeds that of a model without Swin Transformer. However, the performance of this model in terms of accuracy is superior when utilizing a Transformer. The training duration of YOLOv8n is 1.5 hours shorter compared to Swin+YOLOv8. The precision of YOLOv8n is observed to be 1.7% inferior in comparison to Swin+YOLOv8. The model using Transformer exhibits higher accuracy compared to YOLOv8n. The mean Average Precision (mAP) value of Swin+YOLOv8 is seen to be greater compared to the model that does not include other models. Frequently, a higher mAP number is indicative of a more pronounced impact of the model. In contrast to the model without Swin Transformer, Swin+YOLOv8 has the highest mAP value of 0.961.

Error Analysis Our model produces erroneous results on real mask detection. For example, in the detection results of Figure 2, the person at the top of the image wears the mask correctly. But the reality is that the person at the top of the image is wearing a mask incorrectly. We found multiple similar errors in our detection results. Through the analysis of the error results, it is found that errors are prone to occur when detecting people wearing masks on the side face. Inspection of the dataset shows that the dataset has only a small number of images of people wearing masks in profile. We will collect more profiles of people wearing masks in future work.

Fig. 2. One instance of a human face mask image that has been erroneously categorised.

5 Conclusion

In this paper, we introduce a fresh use of Swin Transformer model in the context of mask detection challenges. The proposed model, referred to as Swin+YOLOv8, combines Swin Transformer with the detection module YOLOv8. This integration showcases the enhanced performance compared to the existing mask detection models. The findings indicated that Swin+YOLOv8 model exhibited a mask detection accuracy 96.1% and a mean average precision (mAP) of 0.962, therefore exceeding the performance of traditional models. The use of Swin Transformer inside our model enabled the management of complicated situations and complex data, demonstrated its efficacy in augmenting the precision and efficiency of mask identification jobs.

Nevertheless, the research outcomes also revealed that the Swin+YOLOv8 model required a lengthier training period compared to the other models. This observation underscores the possibility of a compromise between the performance of the model and its computing efficiency. Subsequent investigations may prioritise the refinement of the training procedure for Swin Transformer-based

models, with the aim of achieving an optimal equilibrium between computational expenditure and performance outcomes.

In conclusion, this research work makes a valuable contribution to the existing body of knowledge on mask recognition technology by showcasing the considerable potential of the Swin Transformer in effectively identifying masks. The Swin+YOLOv8 has the potential in practical scenarios to augment public health security, particularly in times of pandemics.

References

1. Addagarla, S.K., Chakravarthi, G.K., Anitha, P.: Real time multi-scale facial mask detection and classification using deep transfer learning techniques. International Journal **9**(4), 4402–4408 (2020)
2. Deng, H., Zhang, J., Chen, L., Cai, M.: Improved mask wearing detection algorithm for SSD. In: Journal of Physics: Conference Series. vol. 1757, p. 012140. IOP Publishing (2021)
3. Diwan, T., Anirudh, G., Tembhurne, J.V.: Object detection using YOLO: Challenges, architectural successors, datasets and applications. Multimedia Tools and Applications **82**(6), 9243–9275 (2023)
4. Du, Z., Su, J., Ding, J., Liu, Z.: Research on YOLO-v3 road target detection based on the combination of K-means++ algorithm and cross-entropy loss function. In: International Conference on Electronic Information Technology (EIT 2022). vol. 12254, pp. 756–760. SPIE (2022)
5. Gao, X., Nguyen, M., Yan, W.Q.: Face image inpainting based on generative adversarial network. In: International Conference on Image and Vision Computing New Zealand (IVCNZ). pp. 1–6. IEEE (2021)
6. Gao, X., Nguyen, M., Yan, W.Q.: A method for face image inpainting based on autoencoder and generative adversarial network. In: Pacific-Rim Symposium on Image and Video Technology. pp. 24–36. Springer (2022)
7. Gong, M., Wang, D., Zhao, X., Guo, H., Luo, D., Song, M.: A review of non-maximum suppression algorithms for deep learning target detection. In: The Symposium on Novel Photoelectronic Detection Technology and Applications. vol. 11763, pp. 821–828. SPIE (2021)
8. Guo, S., Li, L., Guo, T., Cao, Y., Li, Y.: Research on mask-wearing detection algorithm based on improved YOLOv5. Sensors **22**(13), 4933 (2022)
9. Kattenborn, T., Leitloff, J., Schiefer, F., Hinz, S.: Review on convolutional neural networks (CNN) in vegetation remote sensing. ISPRS Journal of Photogrammetry and Remote Sensing **173**, 24–49 (2021)
10. Le, H., Nguyen, M., Yan, W.Q., Nguyen, H.: Augmented reality and machine learning incorporation using YOLOv3 and ARKit. Applied Sciences **11**(13), 6006 (2021)
11. Li, X., Wang, W., Wu, L., Chen, S., Hu, X., Li, J., Tang, J., Yang, J.: Mask dataset, https://makeml.app/datasets/mask
12. Li, Y., Li, S., Du, H., Chen, L., Zhang, D., Li, Y.: Yolo-acn: Focusing on small target and occluded object detection. IEEE Access **8**, 227288–227303 (2020)
13. Liu, Z., Lin, Y., Cao, Y., Hu, H., Wei, Y., Zhang, Z., Lin, S., Guo, B.: Swin transformer: Hierarchical vision transformer using shifted windows. In: IEEE/CVF International Conference on Computer Vision. pp. 10012–10022 (2021)

14. Maity, M., Banerjee, S., Chaudhuri, S.S.: Faster R-CNN and YOLO based vehicle detection: A survey. In: International Conference on Computing Methodologies and Communication (ICCMC). pp. 1442–1447. IEEE (2021)
15. Qi, J., Nguyen, M., Yan, W.Q.: Small visual object detection in smart waste classification using transformers with deep learning. In: International Conference on Image and Vision Computing New Zealand. pp. 301–314. Springer (2022)
16. Ren, X., Liu, X.: Mask wearing detection based on YOLOv3. In: Journal of Physics: Conference Series. vol. 1678, p. 012089. IOP Publishing (2020)
17. Tomás, J., Rego, A., Viciano-Tudela, S., Lloret, J.: Incorrect facemask-wearing detection using convolutional neural networks with transfer learning. In: Healthcare. vol. 9, p. 1050. MDPI (2021)
18. Ullah, N., Javed, A., Ghazanfar, M.A., Alsufyani, A., Bourouis, S.: A novel deepmasknet model for face mask detection and masked facial recognition. Journal of King Saud University-Computer and Information Sciences **34**(10), 9905–9914 (2022)
19. Wang, H., Yan, W.Q.: Face detection and recognition from distance based on deep learning. In: Aiding Forensic Investigation Through Deep Learning and Machine Learning Frameworks, pp. 144–160. IGI Global (2022)
20. Wang, J., Wang, J., Zhang, X., Yu, N.: A mask-wearing detection model in complex scenarios based on YOLOv7-CPCSDSA. Electronics **12**(14), 3128 (2023)
21. Wang, Z., Sun, W., Zhu, Q., Shi, P.: Face mask-wearing detection model based on loss function and attention mechanism. Computational Intelligence and Neuroscience **2022** (2022)
22. Xia, Y., Nguyen, M., Yan, W.Q.: A real-time kiwifruit detection based on improved YOLOv7. In: International Conference on Image and Vision Computing New Zealand. pp. 48–61. Springer (2022)
23. Xiao, B., Nguyen, M., Yan, W.Q.: Fruit ripeness identification using YOLOv8 model. Multimedia Tools and Applications pp. 1–18 (2023)
24. Xue, B., Hu, J., Zhang, P.: Intelligent detection and recognition system for mask wearing based on improved retinaface algorithm. In: International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). pp. 474–479. IEEE (2020)
25. Yan, W.Q.: Introduction to Intelligent Surveillance: Surveillance Data Capture, Transmission, and Analytics. Springer (2019)
26. Yan, W.Q.: Computational Methods for Deep Learning: Theory, Algorithms, and Implementations. Springer Nature (2023)
27. Ye, Q., Zhao, Y.: Mask wearing detection algorithm based on improved yolov4. In: Journal of Physics: Conference Series. vol. 2258, p. 012013. IOP Publishing (2022)
28. Ye, Z., Zhang, H., Liu, Q.: Swtface: A multi-branch network for masked face detection and recognition. In: International Conference on Pattern Recognition and Artificial Intelligence (PRAI). pp. 381–387. IEEE (2022)
29. Yu, J., Zhang, W.: Face mask wearing detection algorithm based on improved YOLO-v4. Sensors **21**(9), 3263 (2021)
30. Yuan, X., Zhang, S., Zhao, C., He, X., Ouyang, B., Yang, S.: Pain intensity recognition from masked facial expressions using swin-transformer. In: IEEE International Conference on Robotics and Biomimetics (ROBIO). pp. 723–728. IEEE (2022)
31. Zeng, C., Liu, Y., Song, C.: Swin-CasUNet: Cascaded U-Net with Swin Transformer for masked face restoration. In: International Conference on Pattern Recognition (ICPR). pp. 386–392. IEEE (2022)