International Journal of Gaming and Computer-Mediated Simulations
Volume 16 • Issue 1

An Improved Face Mask Detection Simulation Algorithm Based on YOLOv5 Model

Yue Qi, Taiyuan Open University, China
Yiqin Wang, Jinzhong University, China*
Yunyun Dong, Taiyuan University of Technology, China

ABSTRACT

This paper proposes an advanced approach for detecting faces with mask occlusion based on YOLOv5 to address various challenges encountered in face detection, including illumination blur and occlusion. The proposed methodology involves the integration of a convolutional block attention module into the backbone network and different network levels in the neck of the YOLOv5s. This approach enables the suppression of irrelevant features and emphasizes the identification of masked facial features. Replacing the conventional loss function with the Focal Loss function addresses the problem of sample imbalance. The enhanced YOLOv5s network was applied to detect mask-occluded faces. Empirical evaluations were conducted on the WIDER Face and AIZOO datasets. The simulation comparison results demonstrate that the proposed method achieves superior real-time detection performance, fulfilling the objective of developing a lightweight detection model.

KEYWORDS

CBAM, Face Recognition, Focal Loss Function, Mask Occlusion, Simulation, YOLOv5

INTRODUCTION

Face recognition technology has developed rapidly in recent decades, driven by advancements in hardware precision and computer processing power (Wang et al., 2022). Convolutional neural networks have significantly developed, enhancing the accuracy of face recognition (Li et al., 2023). However, the outbreak of COVID-19 in 2019 urged people to wear masks. The large area of covered faces made the existing face recognition systems unable to correctly recognize faces (Liu et al., 2022). For example, facial recognition for phone unlocking and high-speed railway station ticket verification becomes inoperable when masks are worn (Wang et al., 2023). In scenarios where facial payment is required, password payment is also necessary (Yan et al., 2021). Currently, there are two main difficulties in facial recognition caused by mask occlusion. Firstly, facial recognition simulation algorithms achieve identity recognition by comparing facial features. Due to the opacity of the mask, the camera cannot capture the nose and mouth of the face. Therefore, the simulation system cannot

DOI: 10.4018/IJGCMS.343517 *Corresponding Author

accurately detect the position of the mask-occluded face and extract sufficient information (Chen et al., 2021). In addition, facial recognition algorithms based on deep learning technology rely on a large amount of data for training. However, there is almost no dataset available for training on occluded faces (Vukovic et al., 2021).

In the past two years, significant research has been conducted on mask-wearing efficacy to enhance epidemic prevention and control measures. Overall, these detection algorithms are divided into two types based on different application scenarios and purposes (Melkani & Maggu, 2021). The first type is detection at checkpoints, access control, and other locations. Personnel are close to the collection equipment, and the image quality obtained through the camera is high. An image only contains one face target. This application scenario has certain requirements for precision. The second type is suitable for crowded places with heavy traffic. This application scenario has many interfering factors and high difficulty. An image contains multiple scales of mask face targets, and most of them are small-scale (Oualla et al., 2021; Zheng & Xu, 2021). To solve real-time pedestrian, mask-wearing detection in such scenarios, various improved methods based on universal object detection models have been proposed to adapt to mask occlusion facial object detection (Guo et al., 2021). However, most existing simulation algorithms cannot obtain reliable detection results when facing difficult problems such as small-scale and complex-scene occlusion. Therefore, an improved mask occlusion face detection algorithm based on YOLOv5 is proposed in this paper. The innovation points of the proposed algorithm are summarized as follows:

1) To focus on the key information of the face under mask occlusion, the proposed algorithm introduces the Convolutional Block Attention Module (CBAM). It is placed within the YOLOv5s network backbone on the P5 layer and within the neck section on both the P4 and P5 layers. This configuration aims to obtain more effective features for detection tasks and improve detection accuracy.
2) Since the binary cross entropy loss function has unbalanced positive and negative samples, the proposed method is improved by using the Focal Loss function to balance samples and ensure the detection effect.

The remainder of this article is organized as follows. The next section, "Related Work," outlines the advantages and disadvantages of existing detection algorithms. The following section is the "Method" of this article and discusses in detail the YOLOv5s network and improvement measures, as well as the detection process of masks blocking faces. Next, the "Results and Analysis" section verifies the effectiveness of the proposed algorithm through the WIDER Face and the AIZOO datasets. Lastly, the conclusion summarizes the proposed algorithm and analyzes future research directions.

RELATED WORK

Various facial detection algorithms based on object detection have been developed (Zheng & Xu, 2021). Vinodini and Karnan (2022) proposed a method based on principal component analysis with good accuracy. However, the method was unable to achieve accurate recognition of faces in complex scenes. Verma et al. (2022) proposed a multi-angle face detection method combining polar harmonic transform with multi-block local binary patterns. This method effectively extracted and used image features from digital images, improving the face recognition rate. Zhou et al. (2022) proposed a new rotation invariant multi-task collaborative network, which improved detection performance through collaboration between face detection and alignment. However, ineffective image feature extraction algorithms resulted in poor detection performance. Akash et al. (2021) proposed a skin color matching method with unchanged lighting conditions in the face detection process. The method balanced high-intensity and low-intensity facial images through separate rules. The face detection algorithm based

on Haar features achieved good effectiveness and robustness; however, the detection method lacked consideration of the issue of facial occlusion, and thus requires further development to ensure its effectiveness in real-world applications.

In the category of deep learning-based face detection methods, Akash et al. (2021) proposed a method based on feature pyramid and triplet loss for training single-level deep neural networks for face detection and recognition. The computational complexity was reduced by sharing the weights, but the performance of a single deep neural network for analyzing small-scale and dense faces was not ideal. Pan et al. (2022) designed a fully convolutional network for face detection and facial landmark localization. They combined the network with progressive pseudo-label training to eliminate the harmful effects caused by inaccurate/noisy annotations. However, the fully convolutional network had a high computational complexity and long detection time. Guo et al. (2022) proposed a multi-face Memory Augmented Neural Network (MTNN) method for detecting and aligning original faces in overlapping facial scenes, ensuring the effectiveness of network training through data expansion. Abdessamad and Korichi (2021) proposed a new support vector machine (SVM) method incorporating Gabor wavelet and other good feature extraction techniques. This approach utilized the generalized Newton Descent direction based on the subgradient calculation to improve face recognition accuracy. However, SVMs struggle with handling multi-class classification problems. Hou et al. (2021) proposed a multi-scale hybrid pyramid convolutional network for face detection. Menaka and Yogameena (2021) proposed a blind deconvolution algorithm to improve the face detection rate and combined it with the discrete wavelet transform to suppress the ringing artifacts in videos, achieving good accuracy in face detection. However, this method did not consider factors such as occlusion and lighting and lacked an efficient detection network. Therefore, the reliability of face detection needs further improvement.

PROPOSED METHOD

Improved YOLOv5 Algorithm

Currently, the mainstream object detection algorithms are divided into single-stage and two-stage detection algorithms. The YOLO series belongs to the single-stage detection category, which has the characteristics of high computational efficiency, making it suitable for detecting faces occluded by masks in practical scenarios (Thome et al., 2021). Therefore, the proposed algorithm selects the YOLOv5s network as the benchmark model, which is a lightweight network with a high recognition rate.

On the Tesla P100 graphics card, YOLOv4 has a detection speed of 50 FPS. Compared to YOLOv4's Darknet architecture, clocking in at 245 MB, YOLOv5s offers significantly greater flexibility due to its compact size of just 27MB. This translates to a roughly 9-fold reduction in model size. YOLOv5 was developed while retaining YOLOv4, improving accuracy and timeliness, with a maximum speed of 140 FPS. The common feature of all target detection frameworks is that the backbone network compresses and extracts features from input, and then forwards information to the target detector. In the YOLO series, its target detectors are the neck section and the YOLO detection head.

The YOLOv5 algorithm has mainly released four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The benchmark network used in the proposed algorithm is YOLOv5s, which has high accuracy, fast speed, and small size. It can be used in embedded devices and is suitable for detecting facial targets occluded by masks. YOLOv5s has a flexible structure that can be easily modified. It has low computational requirements, such as replacing the loss function with Distance Intersection over Union (DIoU), and it introduces multiple data sets to increase experimental samples, which improves the detection effect.

The structure of YOLOv5s consists of four parts. The input end controls the size of the input image, and the backbone part extracts features. The neck part fuses features with position information to obtain richer feature information for the model. The YOLO head part performs the final prediction

output. The backbone section mainly consists of a Focus network and a Cross Stage Partial Network (CSPNet) structure. The base layer of the CSPNet structure is further divided into two parts, the trunk stack dense block part and the residual edge part. The residual edge is directly connected to the main stack after simple processing. These two parts mainly perform the feature fusion process (Su et al., 2022). The backbone also uses a Focus network structure to efficiently extract features. The input image is downsampled by taking every other pixel value. These values are then stacked along the channel dimension to obtain four feature layers. Then, four independent feature layers are stacked, including width and height information. The Focus network effectively reduces the image resolution while expanding the feature channels (assuming that the input is 640×640×3, the shape of the feature layer becomes 320×320×12, with an expansion on the channel and compression on the width and height). Consequently, the concatenated feature layer becomes 12 channels compared to the original three channels. The neck part uses the Feature Pyramid Network (FPN) + Personal Area Network (PAN) structure. At the output end, the bounding box loss function and Non-Maximum Suppression (NMS) are used. The bounding box loss function is used to detect the fit between the real box and the predicted model output box. The NMS is used to eliminate redundant detection boxes.

Due to the large number of useless features in the face covered by a mask, the proposed algorithm introduces an attention mechanism and places it in the P5 layer of the backbone and P4 and P5 layers of the neck part of the YOLOv5s network. The Focal Loss function is used to improve detection accuracy and suppress other useless features, focusing more on the feature information of wearing a mask on the face. The improved YOLOv5s network structure is shown in Figure 1.

The process of detecting faces occluded by masks using the YOLOv5s model embedded with the attention module is described in Figure 2.

Figure 1. YOLOv5s Model Embedded With Attention Module

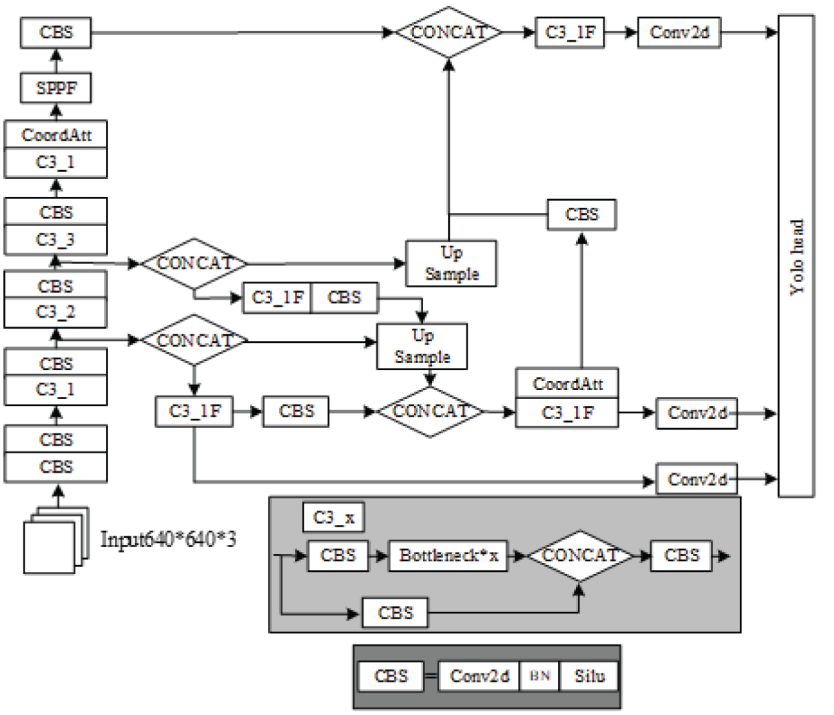
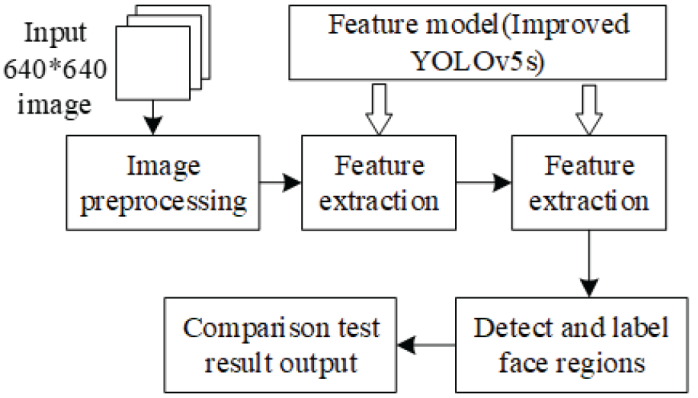


Figure 2. The Detection Process of Mask Blocking the Face



In the detection process, first, a 640×640-size image is obtained, and the specific coordinate position of the facial target in the image is determined. Then, the facial recognition module extracts relevant feature vectors for image classification and feature comparison.

Channel and Spatial Attention Module CBAM

Attention mechanisms can mimic human visual habits, focusing on more important parts when processing images and ignoring invalid information (Kumar et al., 2022). The general attention mechanism can be divided into channel attention and spatial attention, focusing on channel information and spatial information of the image, respectively. Figure 3 shows the CBAM attention mechanism structure consisting of two modules.

Attention mechanisms are widely used in various neural networks to obtain better features. CBAM aims to simultaneously focus on more influential elements from both channel and spatial dimensions. Given an input feature $G \in R^{C\times H\times W}$, $F_C \in R^{C\times 1\times 1}$, and $F_S \in R^{1\times H\times W}$ are channel features and spatial features, respectively. The calculation process is as follows:

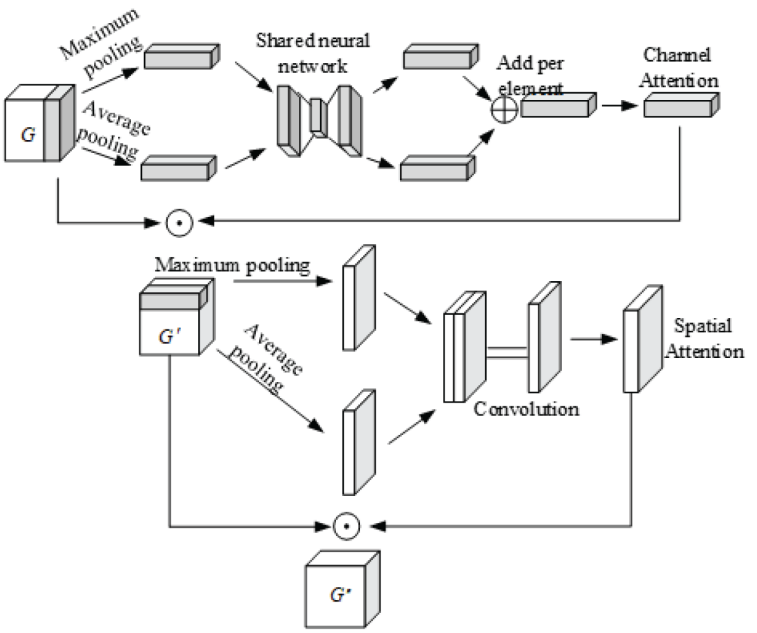
$$G' = F_C \odot G, G'' = F_S \odot G' \tag{1}$$

where, $\odot$ represents the dot product of the element.

The channel attention module mainly focuses on the relationship between channel features. Figure 3 shows that the input feature G undergoes global maximum pooling and global average pooling first, eliminating feature space information. Both resulting outputs are then fed into two shared neural networks. These networks use the activation function ReLU. By sharing the pooled features of neural networks and performing element sum operations, the feature F_C is obtained. After obtaining the feature F_C, it is combined with the earliest input feature G to obtain G', which serves as an input for the spatial attention module. Afterwards, the same maximum pooling and average pooling are performed on the feature to eliminate channel information and obtain spatial features. These two spatial features are joined in channel dimensions, and their dimensionality is reduced to obtain the features F_S. The features F_S are multiplied with G' to obtain G''.

The CBAM is placed in the P5 layer of the YOLOv5s network backbone and both P4 and P5 layers of the neck. In object detection models like YOLOv5s, feature maps extracted at different layers are crucial for detecting targets of different scales. The P5 layer is usually responsible for detecting larger

Figure 3. Structure of CBAM



targets, while the P4 layer is responsible for detecting medium-sized targets. Therefore, introducing CBAM in these layers can help the model better focus on target features at these scales and improve detection accuracy. The CBAM module consists of two attention submodules: channel attention and spatial attention. In channel attention, typically, global average pooling and global maximum pooling are used to generate channel attention maps. These maps are then transformed through a shared multi-layer perceptron (MLP). In spatial attention, a spatial attention map is generated through a convolutional layer. The output dimensions of the CBAM module must match the original feature map to ensure proper application of the attention weights.

The strength of the CBAM module lies in its ability to not only extract important spatial and channel information from images but to also effectively integrate with various neural networks. The network structure of the attention mechanism can enable the feature extraction network to obtain more effective features for detection tasks, which is precisely what is needed for face recognition in mask occlusion mode (Moujahid et al., 2021). When recognizing masked individuals, the human brain prioritizes the visible facial features over the occluded areas. This prioritization leverages prior knowledge and experience to focus on the remaining information to make an identification attempt. However, ordinary feature extraction networks may extract global feature information, which is also the main reason for the decrease in face recognition rate under mask occlusion.

Confidence Loss Function Focal Loss

The confidence loss function in YOLO uses the binary cross entropy loss function. However, the loss function will have problems of unbalanced positive and negative samples (Kumar et al., 2021). The problem of imbalanced positive and negative samples may be the reason for the low accuracy of first-order algorithms compared to second-order algorithms. Therefore, the proposed algorithm improves the confidence loss function in the YOLO algorithm by transforming the binary cross entropy function into the Focal Loss function. The loss function can simultaneously adjust the balance of different samples (Arora et al., 2021; Jia et al., 2021).

Entropy is originally a concept in Shannon's information theory, which is used to measure the probability of things happening. Cross-entropy can be used to measure the gap between the target and the predicted values. It is suitable for loss function. To solve the imbalance between different samples, the Focal-Loss loss function is proposed. The mathematical expression is as follows:

$$FL\left(\rho_t\right) = -\alpha_t \left(1-\rho_t\right)^{\gamma} \log\left(\rho_t\right) \tag{2}$$

Equation (2) shows that weights α_t are introduced to solve the problem of imbalanced samples, while another weight parameter, $\left(1-\rho_t\right)^{\gamma}$, is also introduced to solve the problem of imbalanced difficult samples. The working principle of $\left(1-\rho_t\right)^{\gamma}$ is that a small value of ρ_t indicates the prediction box is misclassified, and the sample is not easy to predict, meaning it may be a difficult sample. At this point, the value of $\left(1-\rho_t\right)^{\gamma}$ is close to 1, and its loss value is not affected. However, when ρ_t approaches 1, it indicates that classification prediction is easy. The sample may be a simple sample, and $\left(1-\rho_t\right)^{\gamma}$ approaches 0, which lowers the loss value of the simple sample. Overall, it is not affected by simple samples. Among them, γ is the modulation factor. When γ is larger, the contribution of simple sample loss will be lower. In practice, adjusting appropriate values can improve the final effect. α_t is the weight parameter that balances the influence of positive and negative samples. A typical value for α_t is 0.25. ρ_t is the factor that adjusts the weights of the difficult samples. Atypical value for ρ_t is 2. These two parameters need to be adjusted for the specific datasets and tasks. When implementing Focal Loss, it is necessary to calculate the category probability of each prediction box and the cross-entropy loss based on the true labels. Then, the loss of each sample is weighted. Finally, the losses of all samples are added to obtain the total loss value. Experiments demonstrated that Focal Loss excels in solving the problem of class imbalance in object detection, especially in the presence of a large number of simple negative samples. Its main characteristics and contributions are as follows:

1. Solving the problem of imbalanced categories: The object detection tasks usually have a problem of imbalanced positive and negative samples, which can affect the recognition accuracy of positive samples. Focal Loss increases the adjustment factor for cross-entropy loss, making the model pay more attention to difficult-to-classify samples during training. As a result, the model focuses more on learning from these challenging examples, ultimately improving the recognition accuracy of positive samples.
2. Improve model performance: Focal Loss helps the model prioritize learning from learning discriminative features by reducing the weight of a large number of simple negative samples in training, ultimately improving model performance.
3. Flexible application to different tasks: Focal Loss is not only suitable for object detection tasks but can also be applied to other tasks that need to solve class imbalance problems, such as image segmentation and face detection. It is possible to adapt to the characteristics and data distribution of different tasks by simply adjusting the form of hyperparameters and loss functions.

SIMULATION RESULTS AND ANALYSIS

Simulation Environment and Datasets

In the experiment, the computer operating system was Windows 10 Professional Edition, the CPU model was Intel (R) Core (TM) i9-10900F CPU @ 2.8 GHz, and the graphics card model was GeForce

RTX 3090, with 24 GB of graphics memory and 32 GB of memory. The model was based on the PyTorch 1.10 framework and accelerated on the GPU using CUDA 11.3.

The data used in the experiment was from WIDER Face and AIZOO datasets. The images from the two datasets were screened to remove incorrect and unreasonable labels and add missing labels. Finally, 3,984 images from the WIDER Face dataset and 4,628 images of faces wearing masks from the AIZOO dataset were selected. The dataset was divided into the training set, validation set, and testing set according to 5:3:2. The experiment was a three-class experiment dividing the images into three labeled classes: mask face class (mask), face class (face), and other occluded faces (face mask). Preprocessing is crucial to enhance the data when dealing with the challenge of mask occlusion. In terms of data preprocessing, firstly, it was necessary to carefully examine each image in the dataset to identify and label facial regions with mask occlusion. The existing facial detection algorithms were used to locate the face, and then, image segmentation or edge detection algorithms were employed to identify occluded areas. The images with severe occlusion or low quality were removed because such images may not have a positive impact on model training. For partially occluded images, only the unobstructed parts should be used for training. In terms of data augmentation, during the training process, random occlusion was applied to the facial regions of the image to simulate different types and degrees of mask occlusion. This approach increases the model's generalization ability. Rotating and flipping the image can further enhance the robustness of the model. This is because it can simulate occlusion situations at different angles and directions. The occlusion effect in different environments, such as indoor, outdoor, cloudy, and sunny, can be simulated by adjusting the brightness and lighting conditions of the image.

When detecting faces with masks, the type of mask, its color, and how much it covers the face are all important factors. The description of three types of masks follows.

(1) Mask type: N95/KN95 Mask – This type of mask usually has thick multi-layer materials and a tight-fitting design, which can cause significant occlusion of facial features.
(2) Medical masks: These masks are usually thin and may only cover the mouth area, having less impact on facial features.
(3) Fabric masks: These types of masks are usually loose, and the degree of coverage depends on their fit and material.

Color interference: Light or white masks may cause blurry boundaries between the face and background increasing detection difficulty. Each image in the dataset has different sizes and labels. The dataset also contains other occlusions blocking the face to prevent the occurrence of non-mask occlusions in practical application scenarios. Figure 4 shows some sample images and annotations.

Evaluating Index

To verify the performance of the proposed algorithm, the commonly used evaluation indices including mAP50 and mAP0.5:0.95, F1 Value, precision P, and recall R, were selected. The P and R are calculated as follows:

$$\begin{aligned} P &= TP \,/\, \left(TP + FP\right) \\ R &= TP \,/\, \left(TP + FN\right) \end{aligned} \tag{3}$$

where TP represents the number of correctly classified positive examples in the sample, and FP is the number of incorrectly classified negative examples in the sample. The average precision mean (mAP) is which is calculated as follows:

Figure 4. Partial Dataset Images and Annotations

$$mAP = \sum_{C=1}^{n} AP_i \Big/ n \tag{4}$$

where n represents the number of categories in the dataset, and APi represents the average precision (AP) of class i. MAP50 refers to the mAP value at the IoU threshold of 0.5. Starting from 0.5, gradually increasing the IoU threshold to a high threshold of 0.95 in small increments of 0.05, the mAP values at different thresholds can be obtained. These mAP values are then averaged to obtain mAP0.5:0.95.

F1 Value is the harmonic mean of precision and recall. It is used to comprehensively evaluate the performance of both. The calculation formula for F1 value is:

$$F1 = 2 \times \frac{precision \times recall}{precision + recall} \tag{5}$$

Model Training

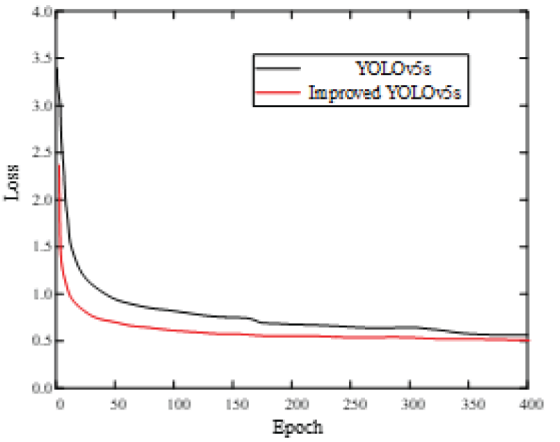
In the experiment, the improved YOLOv5s model was trained in two stages, with the Epoch range set to 0 to 200 in the first stage and 200 to 400 in the second stage. Figure 5 shows the variation in the loss function values of the YOLOv5s network and the proposed model on the WIDER Face and AIZOO datasets during the training process.

Figure 5 shows that as training progresses, the loss function values for both models steadily decrease. However, the improved YOLOv5s network in the proposed algorithm starts with a lower error compared to the original YOLOv5s, allowing the algorithm to converge faster. Taking the WIDER Face dataset as an example, the convergence is achieved when Epoch is about 150. The overall loss value of the proposed model is smaller than that of the original YOLOv5s network. Taking AIZOO dataset as an example, its loss value is less than 0.5.

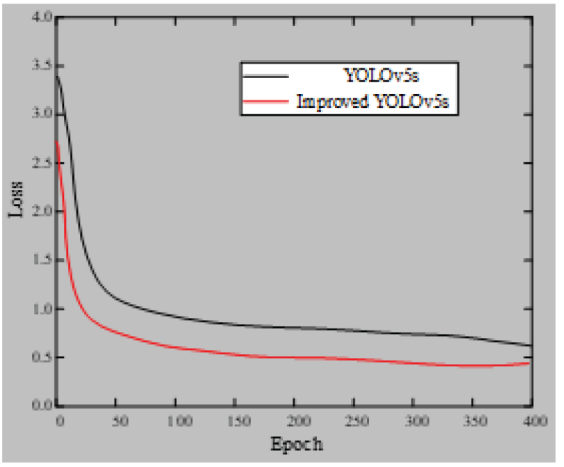
Comparative Experiment of Different Algorithms

To demonstrate the detection performance of the proposed algorithm, comparative experiments were conducted with Verma et al. (2022), Akash et al. (2021), Tsai and Chi (2022), and Hou et al. (2021).

Figure 5. Loss Function Variation Curves for Different Models: (a) WIDER Face; (b) AIZOO



(a)-WIDER-Face



(b)-AIZOO

Comparison of Different Algorithms on the Dataset WIDER Face

The proposed algorithm is applied to image detection in the WIDER Face dataset. Figure 6 shows some image detection results. The green and red boxes represent people with and without masks, respectively. To show the effect of cleanliness and tidiness, the confidence level is not directly drawn on the picture.

It can be observed from Figure 6 that the proposed algorithm performs better in detecting small-sized faces. However, there are recognition errors in mask-wearing for partially occluded faces. This may be due to the complex scene in the WIDER Face dataset and the blurring of occluded faces in the collected crowd images. Therefore, the generalization of the proposed algorithm on WIDER Face test images needs to be further improved to better detect masks blocking faces in dense crowds.

Figure 6. Detection Results of The WIDER Face Dataset

To quantitatively analyze the detection performance of the five algorithms and compare them with the other four algorithms, the detection results on the WIDER Face dataset are shown in Table 1 including mAP50, mAP 0.5:0.95, and F1 indicators.

Table 1 shows that the proposed algorithm performs better at detecting faces in images with masks, achieving 95.9% mAP50, 66.7% mPA0.5:0.95, and 92.8% F1, which were improved by 3.68%, 23.06%, and 4.98%, compared to Hou et al. (2021), respectively. The detection performance of the Hou et al. (2021) method is slightly poorer than the improved YOLOv5s network due to the use of multi-scale mixed pyramid convolutional networks. The detection mPA50 of Tsai and Chi (2022) is only 92.5% due to not considering occlusion factors. Verma et al. (2022) and Akash et al. (2021) used traditional detection algorithms, but their detection performance is poor for small-scale and complex occlusion scenes, with mPA50 lower than 87%. Therefore, the proposed algorithm outperforms other lightweight object detection algorithms in terms of accuracy performance.

Comparison of Different Algorithms on the AIZOO Dataset

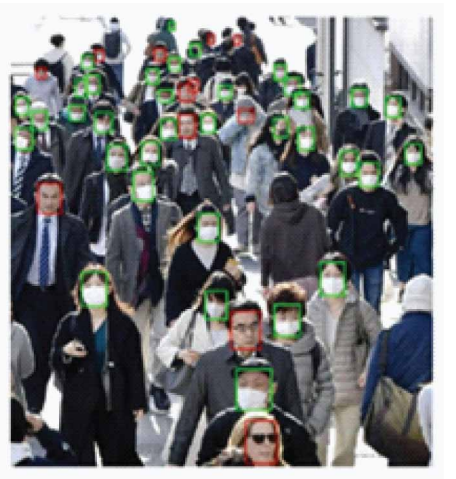
The proposed algorithm was also tested on the AIZOO dataset for image detection. Figure 7 shows some of the image detection results obtained.

Figure 7 shows the proposed algorithm can still detect faces nearby, even if they are slightly covered. However, it sometimes struggles to accurately identify blurry faces that are further away.

Table 1. Performance Comparison of Algorithms Based on WIDER Face Datasets

Algorithm	mPA50/%	mPA0.5:0.95/%	F1
Verma et al. (2022)	84.8	/	/
Akash et al. (2021)	86.1	/	/
Tsai and Chi (2022)	89.3	46.9	85.1
Hou et al. (2021)	92.5	54.2	88.4
Proposed algorithm	95.9	66.7	92.8

Figure 7. Detection Results of the AIZOO Face Dataset



Overall, the performance of the proposed algorithm on the AIZOO dataset is ideal, meeting the requirements of mask-wearing detection in dense populations.

Moreover, the mAP50, mAP0.5:0.95, and F1 indicator values of the simulation results of the five algorithms on the AIZOO dataset are listed in Table 2.

Table 2 shows that the proposed algorithm maintains a significant leading advantage in detection performance. The fusion of CBAM and Focal Loss functions enables the algorithm to pay more attention to important image information, improve detection performance, and reduce computational complexity. The detected mAP50, mPA0.5:0.95, and F1 values are 96.5%, 67.1%, and 94.3%, respectively. Hou et al. (2021) used a multi-scale mixed pyramid convolutional network, while Tsai and Chi (2022) used an improved deep neural network. Both methods achieve good accuracy in facial detection, with mAP50 exceeding 90%. However, their detection performance needs to be improved because neither method considers the occlusion factors. Verma et al. (2022) and Akash et al. (2021) use traditional detection algorithms. However, their detection performance is poor for small-scale and complex occlusion scenes with mPA50 below 90%. Therefore, the proposed algorithm performs better than other object detection algorithms on the AIZOO dataset.

Ablation Experiment

The effect of the improved YOLOv5s network on the performance of facial mask occlusion detection was verified through ablation experiments. YOLOv5s, YOLOv5s+CBAM, YOLOv5s+Focal Loss, and YOLOv5s+CBAM+Focal Loss were trained four times. YOLOv5s+CBAM+Focal Loss is the

Table 2. Performance Comparison of Algorithms Based on AIZOO Dataset

Algorithm	mPA50/%	mPA0.5:0.95/%	F1/%
Verma et al. (2022)	85.2	/	/
Akash et al. (2021)	86.4	/	/
Tsai and Chi (2022)	90.7	47.2	88.4
Hou et al. (2021)	93.1	56.3	92.1
Proposed algorithm	96.5	67.1	94.3

Table 3. Ablation Results on The WIDER Face Dataset

Indicator value	YOLOv5s	YOLOv5s+CBAM	YOLOv5s+Focal Loss	YOLOv5s+CBAM+Focal Loss
P/%	94.6	95.2	95.6	96.1
R/%	93.1	93.4	92.9	93.7
mPA/%	96.0	96.4	96.6	96.9
F1/%	93.8	94.3	94.2	94.9
Parameters/10^5	17.6	18.1	17.6	18.1
Speed-GPU/ms	1.8	1.5	1.6	1.4

final improved model. The model parameters and test results obtained from training on the WIDER Face and AIZOO datasets are shown in Tables 3 and 4, respectively.

It can be seen from Table 3 and Table 4 that incorporating CBAM into the YOLOv5s model resulted in an increase of 0.95%, 0.54%, and 0.52% in P, R, and mPA compared to the original model, respectively. While the parameter size with the added CBAM only increased by 1.15%, it wasn't enough to significantly improve face detection, especially for occluded faces. This is because the YOLOv5s model is not clear enough to extract facial features and cannot accurately detect all occluded faces, resulting in missed detection. CBAM pays more attention to facial position and features and improves the model's feature extraction ability, which can detect facial positions more quickly and accurately. The detection speed is also naturally improved, with a Speed GPU of only 1.4ms. Similarly, the original model introduced the Focal Loss function to improve the face detection effect by adjusting and balancing positive and negative samples. Taking the WIDER Face dataset as an example, YOLOv5s+Focal Loss improved the accuracy by 1.06% and mPA by 0.63%, compared to the original model. The proposed model combines CBAM and Focal Loss, and its detection performance has been significantly improved. On the WIDER Face and AIZOO datasets, the proposed model achieved P, R, mPA, and F1 values of 96.1%, 93.7%, 96.9%, and 94.9% and 96.5%, 94.1%, 97.2%, and 95.3%, respectively. Moreover, the proposed model maintains a fast detection speed ranging from 1.4ms to 1.6ms.

CONCLUSION

The existing detection methods mainly target application scenarios where pedestrians pass through gates or checkpoints in sequence with a single background and a small number of faces. However, the application results are not satisfactory for complex backgrounds, faces with different sizes, or mutual occlusion. Therefore, this paper proposes an improved mask occlusion face detection algorithm based on YOLOv5. The proposed algorithm incorporates CBAM and Focal Loss functions into the original YOLOv5s network

Table 4. Ablation Results on the AIZOO Dataset

Indicator value	YOLOv5s	YOLOv5s+CBAM	YOLOv5s+Focal Loss	YOLOv5s+CBAM+Focal Loss
P/%	94.8	95.7	95.4	96.5
R/%	93.4	93.9	93.6	94.1
mPA/%	96.2	96.7	96.5	97.2
F1/%	94.1	94.8	94.5	95.3
Parameters/10^5	17.4	17.8	17.4	17.8
Speed-GPU/ms	1.9	1.6	1.6	1.4

to enhance its ability to extract key facial information from masks and accurately detect mask-occluded faces. The experimental simulation results based on the WIDER Face and AIZOO datasets indicate that:

(1) Integrating CBAM into the YOLOv5s network can suppress useless information, reduce computational complexity, and improve detection efficiency. The proposed algorithm has a detection speed of 1.4ms to 1.6ms, which can achieve fast and reliable detection.
(2) The detection performance of the improved YOLOv5s network is superior to other comparative algorithms in terms of accuracy. On the WIDER Face and AIZOO datasets, the proposed model achieved P, R, mPA, and F1 indicators of 96.1%, 93.7%, 96.9%, and 94.5% and 96.5%, 94.1%, 97.2%, and 95.3%, respectively.

Since this article only considers facial recognition under mask occlusion and does not sufficiently consider issues such as changes in light and density, further improvements will be made to the detection algorithm in future research to achieve the goal of facial information recognition under mask occlusion and changes in light and other factors in dense populations. Facial recognition technology can be combined with other biometric systems (such as fingerprint recognition, iris recognition, or voice recognition) to form a multimodal biometric system. Such a system can utilize multiple biological features to improve recognition accuracy and reliability. For example, when facial recognition systems cannot accurately recognize due to occlusion or quality issues, they can be supplemented by other biometric technologies.

As facial recognition technology becomes more widely used and sophisticated, public concerns regarding ethics and privacy are rising. These issues are particularly prominent in sensitive applications such as security monitoring and identity verification. These concerns include the following:

(1) Fairness: Facial recognition technology may lead to unfair bias and discrimination. For example, if the training data is imbalanced or contains biases, the system may be more inclined to recognize faces of certain races or genders. This may lead to unfair treatment, such as mistakenly marking someone as a suspect or refusing their services.
(2) Privacy: Facial recognition technology may infringe upon an individual's privacy rights. Collecting, storing, and using facial data without consent is unethical. Moreover, the leakage of facial data may lead to serious consequences, such as identity theft or harassment.
(3) Transparency and interpretability: The decision-making process of facial recognition systems should be transparent and interpretable. If someone is misidentified or their identity isn't recognized, the user should have the right to know the reason. Without transparency and interpretability, these systems can breed distrust and confusion.

CONFLICT OF INTEREST

We wish to confirm that there are no known conflicts of interest associated with this publication and there has been no significant financial support for this work that could have influenced its outcome.

FUNDING

This work was supported by Planning subject for the 14th five year plan of Shanxi education sciences (No.GH-21105 and No.GH-220335), Educational Reform and Innovation Project of Higher education in Shanxi Province (No.J20221040).

PROCESS DATES

Received: 2/1/2024, Revision: 3/19/2024, Accepted: 3/19/2024

REFERENCES

Abdessamad, A., & Korichi, M. (2021). A deeper Newton descent direction with generalised Hessian matrix for SVMs: An application to face detection. *International Journal of Mathematical Modelling and Numerical Optimisation*, *11*(2), 196–208. doi:10.1504/IJMMNO.2021.114485

Akash, A. A., Akhand, M. A. H., & Siddique, N. (2021). Robust face detection integrating novel skin color matching under variant illumination conditions. *International Journal of Image*. *Graphics and Signal Processing*, *13*(2), 1–15. doi:10.5815/ijigsp.2021.02.01

Arora, M., Garg, S., & A, S. (2021). Face mask detection system using Mobilenetv2. *International Journal of Engineering and Advanced Technology*, *10*(4), 127–129. doi:10.35940/ijeat.D2404.0410421

Chen, B., Ju, X., Xiao, B., Ding, W., Zheng, Y., & de Albuquerque, V. H. C. (2021). Locally GAN-generated face detection based on an improved Xception. *Information Sciences*, *572*(11), 16–28. doi:10.1016/j.ins.2021.05.006

Guo, Q., Wang, Z., Fan, D., & Wu, H. (2022). Multi-face detection and alignment using multiple kernels. *Applied Soft Computing*, *122*(108808), 108808. doi:10.1016/j.asoc.2022.108808

Hou, S., Fang, D., Pan, Y., Li, Y., & Yin, G. (2021). Hybrid pyramid convolutional network for multiscale face detection. *Computational Intelligence and Neuroscience*, *2021*(9963322), 1–15. doi:10.1155/2021/9963322 PMID:34035802

Jia, S., Hu, C., Li, X., & Xu, Z. (2021). Face spoofing detection under super-realistic 3D wax face attacks. *Pattern Recognition Letters*, *145*, 103–109. doi:10.1016/j.patrec.2021.01.021

Kumar, A., Kalia, A., & Kalia, A. (2022). ETL-YOLO v4: A face mask detection algorithm in era of COVID-19 pandemic. *Optik (Stuttgart)*, *259*(169051), 169051. Advance online publication. doi:10.1016/j.ijleo.2022.169051 PMID:35411120

Kumar, A., Kalia, A., Verma, K., Sharma, A., & Kaushal, M. (2021). Scaling up face masks detection with YOLO on a novel dataset. *Optik (Stuttgart)*, *239*(5), 166744. doi:10.1016/j.ijleo.2021.166744

Li, T., & Donta, P. K. (2023). Predicting green supply chain impact with snn-stacking model in digital transformation context. *Journal of Organizational and End User Computing*, *35*(1), 1–19. doi:10.4018/JOEUC.334109

Liu, X., Balestrieri, E., & Melcher, D. (2022). Evidence for a theta-band behavioural oscillation in rapid face detection. *The European Journal of Neuroscience*, *56*(7), 5033–5046. doi:10.1111/ejn.15790 PMID:35943892

Melkani, G., & Maggu, S.Gaurav Melkani and Dr. Sunil Maggu. (2021). Image-based face detection and recognition. *International Journal for Modern Trends in Science and Technology*, *6*(12), 466–470. doi:10.46501/IJMTST061290

Menaka, K., & Yogameena, B. (2021). Face detection in blurred surveillance videos for crime investigation. *Journal of Physics: Conference Series*, *1917*(1), 012024. doi:10.1088/1742-6596/1917/1/012024

Moujahid, A., Dornaika, F., Arganda-Carreras, I., & Reta, J. (2021). Efficient and compact face descriptor for driver drowsiness detection. *Expert Systems with Applications*, *168*(12), 114334. doi:10.1016/j.eswa.2020.114334

Oualla, M., Ounachad, K., & Sadiq, A. (2021). Building face detection with face divine proportions. [IJOE]. *International Journal of Online and Biomedical Engineering*, *17*(04), 63. doi:10.3991/ijoe.v17i04.19149

Pan, Z., Wang, Y., & Zhang, S. (2022). Joint face detection and facial landmark localization using graph match and pseudo label. *Signal Processing Image Communication*, *102*(6), 116587. doi:10.1016/j.image.2021.116587

Su, X., Gao, M., Ren, J., Li, Y., Dong, M., & Liu, X. (2022). Face mask detection and classification via deep transfer learning. *Multimedia Tools and Applications*, *81*(3), 4475–4494. doi:10.1007/s11042-021-11772-5 PMID:34903950

Thome, I., Hohmann, D. M., Zimmermann, K. M., Smith, M. L., Kessler, R., & Jansen, A. (2021). "I spy with my little eye, something that is a face…": A brain network for illusory face detection. *Cerebral Cortex (New York, N.Y.)*, *32*(1), 137–157. doi:10.1093/cercor/bhab199 PMID:34322712

Tsai, T., & Chi, P. (2022). A single-stage face detection and face recognition deep neural network based on feature pyramid and triplet loss. *IET Image Processing*, *16*(8), 2148–2156. Advance online publication. doi:10.1049/ipr2.12479

Verma, A., Baljon, M., Mishra, S., Kaur, I., Saini, R., Saxena, S., & Kumar Sharma, S. (2022). Secure rotation invariant face detection system for authentication. *Computers, Materials & Continua*, *70*(1), 1955–1974. doi:10.32604/cmc.2022.020084

Vinodini, R., & Karnan, M. (2022). Face detection and recognition system based on hybrid statistical, machine learning and nature-based computing. *International Journal of Biometrics*, *14*(1), 3. doi:10.1504/IJBM.2022.119543

Vukovic, I., Cisar, P., Kuk, K., Bandjur, M., & Popovic, B. (2021). Influence of image enhancement techniques on effectiveness of unconstrained face detection and identification. *Elektronika ir Elektrotechnika*, *27*(5), 49–58. doi:10.5755/j02.eie.29081

Wang, W., Huang, X., & Luo, S. (2022). Measuring enterprise mutual information based on the helix model. *Journal of Organizational and End User Computing*, *34*(7), 1–17. doi:10.4018/JOEUC.297117

Wang, Y., Ghani, D. A., & Zhou, B. (2023). A two-stage emotion generation model combining CGAN and pix2pix. *Journal of Organizational and End User Computing*, *35*(1), 1–21. doi:10.4018/JOEUC.330647

Yan, H., Liu, Y., Wang, X., Li, M., & Li, H. (2021). A face detection method based on skin color features and Adaboost algorithm. *Journal of Physics: Conference Series*, *1748*(4), 042015. doi:10.1088/1742-6596/1748/4/042015

Zheng, G., & Xu, Y. (2021). Efficient face detection and tracking in video sequences based on deep learning. *Information Sciences*, *568*, 265–285. doi:10.1016/j.ins.2021.03.027

Zhou, L., Zhao, H., & Leng, J. (2022). MTCNet: Multi-task collaboration network for rotation-invariance face detection. *Pattern Recognition*, *124*, 108425. doi:10.1016/j.patcog.2021.108425

Yue Qi, Associate Professor, Master of Computer Science, Graduated from Taiyuan University of Technology in 2010. Worked in Taiyuan Open University. His research interests include object detection and deep learning.

Yiqin Wang, Master of Computer Science, Associate Professor. Graduated from Tianjin Normal University in 2007. Worked in Jinzhong University. Her research interests include deep learning and image processing.

Yunyun Dong, Lecturer, Doctor of Computer Science,Graduated from Taiyuan University of Technology in 2021. Worked in Taiyuan University of Technology. Her research interests include intelligent information processing and computer simulation.