applied sciences

MDPI

Article

Mask-Wearing Detection in Complex Environments Based on Improved YOLOv7

Guang Feng [1], Qun Yang [1], Chong Tang [2], Yunhai Liu [2], Xiaoting Wu [1] and Wenyan Wu [3,*]

[1] College of Automation, Guangdong University of Technology, Guangzhou 511400, China; von@gdut.edu.cn (G.F.); 2112104101@mail2.gdut.edu.cn (Q.Y.); 2112104216@mail2.gdut.edu.cn (X.W.)
[2] College of Computer, Guangdong University of Technology, Guangzhou 511400, China; 2112105015@mail2.gdut.edu.cn (C.T.); 2112105295@mail2.gdut.edu.cn (Y.L.)
[3] Center of Campus Network & Modern Educational Technology, Guangdong University of Technology, Guangzhou 510006, China
* Correspondence: wuwy@gdut.edu.cn

Abstract: Wearing masks is an effective protective measure for residents to prevent respiratory infectious diseases when going out. Due to issues such as a small target size, target occlusion leading to information loss, false positives, and missed detections, the effectiveness of face mask-wearing detection needs improvement. To address these issues, an improved YOLOv7 object detection model is proposed. Firstly, the C2f_SCConv module is introduced in the backbone network to replace some ELAN modules for feature extraction, enhancing the detection performance of small targets. Next, the SPPFCSPCA module is proposed to optimize the spatial pyramid pooling structure, accelerating the model convergence speed and improving detection accuracy. Finally, the HAM_Detect decoupled detection head structure is introduced to mitigate missed and false detections caused by target occlusion, further accelerating model convergence and improving detection performance in complex environments. The experimental results show that improved YOLOv7 achieved an mAP of 90.1% on the test set, a 1.4% improvement over the original YOLOv7 model. The detection accuracy of each category improved, effectively providing technical support for mask-wearing detection in complex environments.

Keywords: object detection; YOLOv7; decoupled head; mask-wearing detection; attention mechanism

check for updates

Citation: Feng, G.; Yang, Q.; Tang, C.; Liu, Y.; Wu, X.; Wu, W. Mask-Wearing Detection in Complex Environments Based on Improved YOLOv7. *Appl. Sci.* **2024**, *14*, 3606. https://doi.org/10.3390/app14093606

Academic Editor: Rosario Pecora

Received: 20 February 2024
Revised: 18 March 2024
Accepted: 8 April 2024
Published: 24 April 2024

1. Introduction

With the outbreak of seasonal flu during the winter and spring seasons, wearing masks has become a crucial protective measure for residents during their daily outings [1]. However, in crowded public places such as stations, hospitals, schools, and malls, there are still many pedestrians not wearing masks. Relying solely on manual inspections to encourage mask wearing not only increases labor costs but also raises the risk of virus transmission. Therefore, researching a real-time detection system for whether people are wearing masks in complex environments is essential. This can help reduce the risk of virus transmission, safeguarding the health of individuals and others. Establishing such a system holds significant practical significance in building a healthier and safer society.

As societal demands and deep learning advancements progress, current mainstream object detection algorithms can be broadly categorized into two types: two-stage object detection algorithms based on candidate box regions and single-stage object detection algorithms based on regression analysis. Common two-stage object detection algorithms include Fast R-CNN [2], Faster R-CNN [3], and Mask R-CNN [4], while single-stage object detection algorithms include YOLO, SSD [5], and Retina-Net [6]. Currently, most object detection algorithms are suitable for face mask detection, but the effectiveness of existing algorithms for face mask detection in complex environments is often suboptimal due to challenges such as complex backgrounds, target occlusion, and motion blur. Researchers worldwide have made significant strides in developing algorithms for face mask detection. Sun et al. [7] proposed MDDC-YOLO, an improved mask-wearing detection algorithm based on YOLOv5 specifically designed for dense crowds under surveillance perspectives. It achieved a 6.5 percentage point increase in detection accuracy compared with YOLOv5. However, it did not address the missed and false detections of small targets in complex environments. Li et al. [8] addressed the issues of dense targets, occlusion, and small-scale objects in mask-wearing detection in complex environments such as malls and stations by introducing improved DenseNet into YOLOv5. This achieved a detection accuracy of 97.8% but with slower model detection speeds. Fu et al. [9] focused on problems related to correct mask wearing, different shooting angles, and occlusion. They introduced convolutional attention mechanisms and improved spatial pyramid pooling into the detection head of YOLOv7, achieving an mAP of 93.8% on the test set. However, their approach was limited to detecting mask wearing by single individuals in simple environments. Nonetheless, the aforementioned research efforts still require further improvement in mask-wearing detection accuracy and speed in complex environments, small target detection, and addressing information loss due to target occlusion.

To address the above-mentioned issues and improve the effectiveness of mask-wearing detection in complex environments, the main contributions of this paper are summarized as follows.

We utilize web scraping techniques to collect images of individuals wearing masks in various complex scenarios from the internet. Combined with the publicly available dataset AIZOO, the images are reorganized and annotated using LabelImg software [10]. The dataset is then split into training and testing sets. Then, by employing techniques such as cropping, flipping, and Gaussian blurring for data augmentation, the diversity of the dataset can be increased, thereby enhancing the model's generalization performance.

We introduce an improved real-time mask-wearing detection model based on YOLOv7. In the backbone network, the C2f_SCConv module is introduced to replace some efficient layer aggregation networks (ELANs), reducing the network parameters and increasing the receptive field, thereby facilitating the extraction of richer facial mask features. The spatial pyramid pooling structure is optimized at the neck, introducing the Spatial Pyramid Pooling-Fast, Cross-Stage Partial Channel with Attention (SPPFCSPCA) mechanism module, which integrates an attention mechanism into the spatial pyramid pooling structure to improve the training speed and accelerate model convergence. The detection head incorporates the HAM_Detect decoupling head to mitigate issues such as target occlusion, false positives, and false negatives, optimizing detection performance in complex environments and further accelerating model convergence while enhancing detection accuracy.

The experimental results validate that the improved YOLOv7 model addresses the issues of small target false negatives and false positives in mask-wearing detection in complex environments. It ensures the superiority of training loss and detection accuracy performance, providing effective technical support for improving mask-wearing detection in complex environments.

2. Related Works

2.1. YOLOv7

YOLOv7 [11] is an outstanding single-stage object detector that offers six different versions: YOLOv7-tiny, YOLOv7, YOLOv7-d6, YOLOv7-e6, YOLOv7-e6e, and YOLOv7-w6. Among them, YOLOv7 focuses on achieving a balance between accuracy and speed when performing inference on edge devices. This paper builds upon this version with improvements aimed at enhancing mask-wearing detection performance.

The YOLOv7 network structure comprises three parts: the input stage, backbone, and the neck and head. The input stage is responsible for scaling the input image to meet the requirements of the backbone network. After preprocessing and data augmentation, the processed image is fed into the backbone network for feature extraction. The neck part merges the extracted features to generate features of different sizes: large, medium, and small. These fused features are then passed to the detection head, which outputs a tensor containing the class confidence, object center coordinates, width, and height. YOLOv7 uses the sigmoid function to map the output values between 0 and 1 and applies a non-maximum suppression algorithm for filtering. During inference, YOLOv7 adopts a center-based detection approach where the model predicts the center coordinates, width, height, and object class for each object. Finally, objects are classified and localized based on the prediction results, followed by non-maximum suppression for filtering, resulting in the detection output.

The YOLOv7 network structure, as shown in Figure 1, takes 640 × 640 three-channel RGB training images as input. It adopts the Mosaic data augmentation method proposed in YOLOv4, which involves randomly cropping four input images and then stitching them together to form a single training image. This enriches the dataset and enhances training efficiency. The backbone network consists mainly of efficient layer aggregation network (ELAN) modules and MP (MaxPool) modules. The ELAN modules are used for image feature extraction and channel control, while the MP modules maintain consistency in the number of channels before and after processing. The neck and detection head comprise the Spatial Pyramid Pooling, Cross-Stage Partial Channel (SPPCSPC) modules, extended ELAN (E-ELAN) modules, and MP2 modules. It adopts the same Path Aggregation Feature Pyramid Network (PAFPN) structure as YOLOv5 for aggregating features of multiple sizes. This structure forms large, medium, and small IDetect detection heads and decouples the feature information obtained from the neck. The RepVGG Block (REP) structure is used to adjust the number of channels in the decoupled feature information. Finally, the generated feature maps are predicted and output using 1 × 1 convolutions. Through the collaboration of these modules, the YOLOv7 network achieves excellent results in object detection tasks.

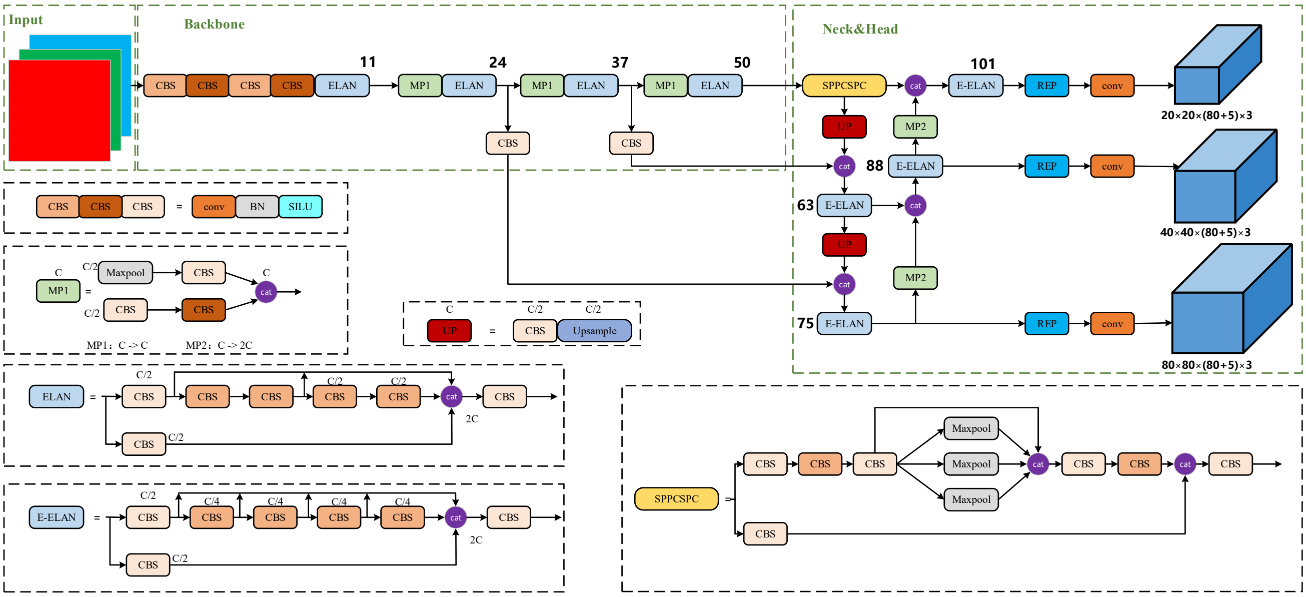


Figure 1. Network structure of YOLOv7.

In recent years, various improvement methods for YOLOv7 have been proposed, leading to a continuous stream of research achievements. Wang et al. [12] introduced the YOLOv7-CPCSDSA detection model for mask-wearing detection. They utilized the CatPConv structure to reduce computational redundancy and memory access, added the SD module to enhance the detection of small targets, and introduced the SA mechanism to enhance the collection of local information. The model achieved an average precision of 88.4%. Praveen et al. [13] proposed an improved YOLOv7 model for apple detection in complex backgrounds. By introducing a specific multi-head attention mechanism to capture complex interactions between regions and features, they improved the accuracy of apple detection, achieving a precision value of 91%. Ding et al. [14] presented an improved mask-wearing detection algorithm based on YOLOv7. They enhanced the perception of small targets by adding attention mechanisms to the backbone network and introducing partial convolution methods. Integration of the DSC structure in the detection head reduced the device requirements and improved computational speed while also enhancing detection performance by reducing the network parameters and computational complexity. Zeng et al. [15] proposed a real-time small object detection algorithm, YOLOv7-UAV, by removing the second downsampling layer and the deepest detection head. They introduced the DpSPPF module to effectively extract feature information of different scales, optimized the K-means algorithm, and used weighted normalization operations. As a result, the average precision increased by 2.89%, while the parameter count decreased by 8.3%.

2.2. Loss Function

In object detection networks, the localization of targets relies on a bounding box regression model, which encompasses three major geometric factors: the overlapping area, center point distance, and aspect ratio. The loss function is used to assess the degree to which the predicted bounding boxes differ from the ground truth boxes. A smaller loss function value typically indicates better model performance. Currently, the main introduced loss functions in object detection networks are compared as shown in Table 1.

Table 1. Comparison of different loss functions.

Name	Loss Function Description
IoU	Intersection over union: the ratio of the intersection area to the union area between the predicted and true bounding boxes. It ranges from 0 to 1, with 1 indicating perfect overlap. However, the IoU does not reflect the distance between non-overlapping boxes or distinguish accuracy when the IoU is the same for multiple predictions.
GIoU	Generalized IoU: an extension of the IoU that introduces the smallest enclosing box for both the predicted and true boxes. It encourages the predicted box to be as close as possible to the true box, especially when they do not overlap. The GIoU reduces to the IoU when the boxes are horizontally aligned.
DIoU	Distance IoU: builds upon the IoU by considering the Euclidean distance between the centers of the predicted and true boxes, along with the overlap area. This helps improve the convergence speed of the loss function.
CIoU	Complete IoU: an extension of the DIoU that also incorporates aspect ratio information to enhance the stability of object box regression. It provides better judgment when the aspect ratios differ significantly.
EIoU	Enhanced IoU: an extension of the CIoU that separates the aspect ratio's influence factor for the predicted and true boxes. It calculates the lengths and widths independently, improving the convergence speed and achieving better localization accuracy.
α IoU	Alpha IoU: introduces a tunable alpha parameter to the IoU for increased flexibility in adapting to different horizontal boundary regression box accuracies. It exhibits better robustness in small datasets and noise.
SIoU	Spatial IoU: an extension of the EIoU that adds the angle loss between the predicted and true box centers. It redefines distance loss, effectively reducing regression freedom. Additionally, it introduces a category information weighting factor to enhance detection model classification accuracy.
WIoU	Weighted IoU: introduces attention-based bounding box losses. WIoU v1 constructs attention-based bounding box losses, while WIoU v2 and v3 further enhance the attention mechanism by introducing gradient gain (focus coefficient) calculation methods.

2.3. Detection Heads

In object detection tasks, the detection head is a specific module used to recognize and locate targets, and its design significantly influences the model's detection accuracy and speed. Currently, detection heads can be mainly categorized into four types: anchor-based, anchor-free, self-attention, and cascade. Among them, anchor-based and anchor-free are the most widely used types. For instance, the YOLO series, SSD series, and most R-CNN series are predominantly designed based on anchors, utilizing predefined anchor boxes to match real target boxes. On the other hand, models such as CornerNet [16], CenterNet [17], and FCOS are designed based on anchor-free principles. The key distinction lies in whether predefined anchor boxes are employed to match real target boxes.

Furthermore, in object detection tasks, target classification and localization regression are highly correlated, but the learning mechanisms for these two tasks are inherently contradictory. Classification tasks, aiming to enhance semantic understanding, require richer global contextual information and coarse features. On the other hand, regression tasks prefer detailed information about the bounding boxes, necessitating fine features. As a result, the coupled detection head for both tasks is widely recognized to significantly impact model convergence, especially for mask targets that are densely distributed and prone to occlusion.

Decoupled detection head (Decouple Head) is a method that decouples the tasks of object classification and localization regression, treating them as independent operations. In recent years, it has emerged as a new direction for improving object detection tasks, yielding significant results. Tian et al. [18] introduced the FCOS model, which employs a Decouple Head structure to separate the two tasks and introduces convolutional layers independently on each branch, allowing each task to make spatial judgments without affecting each other. Wu et al. [19] reinterpreted classification and regression, finding that fc-heads are more suitable for classification tasks, while conv-heads are more suitable for regression tasks. The YOLOX model proposed by Ge et al. [20] utilizes a Decouple Head structure, separating the classification and regression tasks into two independent branches. During prediction, the results are integrated, marking the first introduction of a decoupled detection head in the YOLO series, significantly improving the convergence speed and performance. Xu et al. [21] introduced the PP-YOLOE model, employing an anchor-free detection head design. The detection speed and accuracy of PP-YOLOE are somewhat superior to YOLOX and YOLOv5, making it an advanced industrial target detector with high performance and user-friendly deployment. The YOLOv6 model proposed by the Meituan-Dianping Computer Vision Intelligence team [22], designed for industrial applications, adopts the anchor-free paradigm. It simultaneously reduces a 3 × 3 convolutional layer on both an independent classification and regression branch, reducing computational costs and achieving lower inference latency. The YOLOv8 model, based on optimizing YOLOv5 with Ultralytics [23], replaces the detection head with a Decouple Head structure compared with YOLOv5. It separates the classification and regression tasks and introduces the closure of Mosaic augmentation in the last 10 epochs from YOLOX, effectively enhancing the detection accuracy.

3. Improved YOLOv7

The improved YOLOv7 network structure proposed in this paper is illustrated in Figure 2. It is mainly composed of three parts: the input stage (Input), the backbone network (Backbone), and the neck and detection head (Neck & Head). The modules highlighted in red indicate the improved modules. In the backbone network, the C2f_SCConv module is used to replace certain feature extraction modules, enhancing the detection performance of small targets. In the neck part, the SPPFCSPCA module optimizes the spatial pyramid pooling structure, accelerating the model convergence speed. In the detection head part, the HAM_Detect decoupled head is employed to further accelerate model convergence and improve detection performance in complex environments.

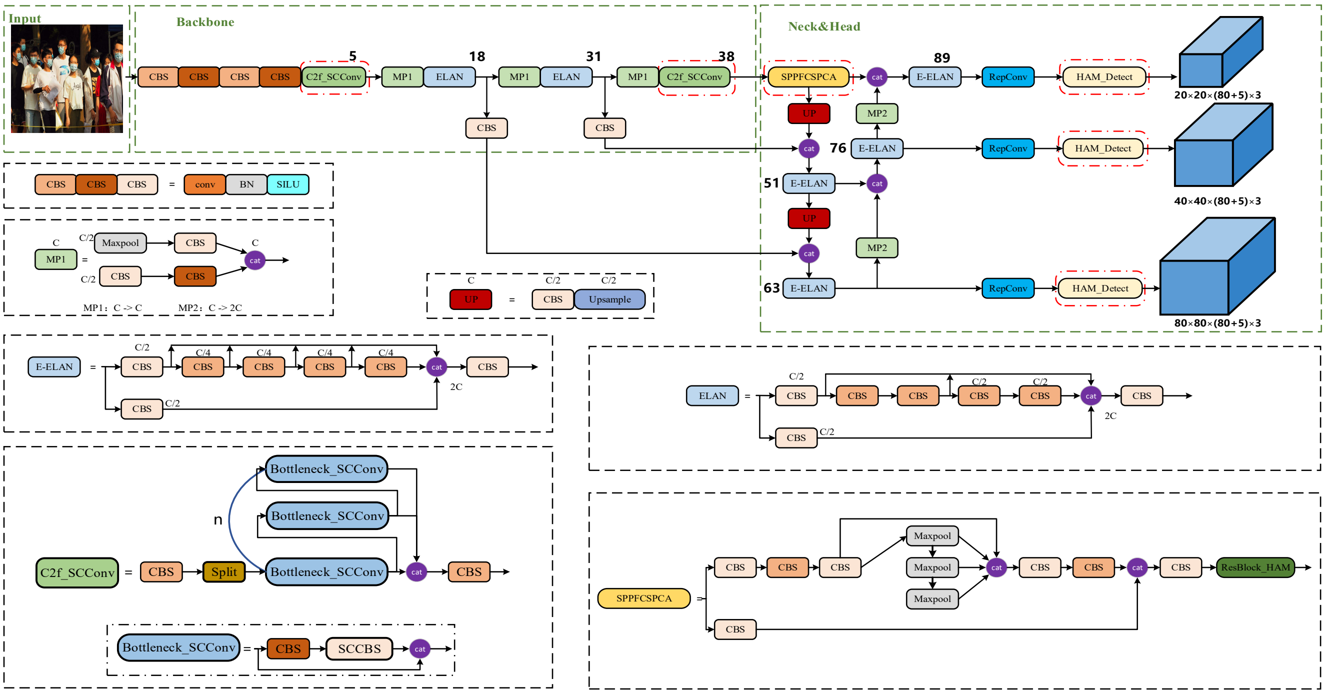


Figure 2. Network structure of improved YOLOv7.

3.1. *C2f_SCConv*

The backbone network primarily introduces the C2f_SCConv module to replace some of the ELEN modules. The C2f_SCConv module combines the C2f structure with Spatial and Channel Reconstruction Convolution (SCConv) [24], The C2f structure originates from YOLOv8, which was designed by integrating the C3 module from YOLOv8 and the ELAN concept from YOLOv7. This ensures it being lightweight while obtaining richer gradient flow information. However, the original C2f structure still lacks a sufficient feature extraction capability for masks in complex environments, as it lacks multidimensional feature information and may exhibit feature redundancy issues when combined with ordinary convolutions. SCConv effectively mitigates feature redundancy, reducing the model parameters and computational costs while enhancing feature representation capabilities and improving network feature extraction performance.

As shown in Figure 3, SCConv consists of two parts: the spatial refined unit (SRU) and the channel refined unit (CRU). For the input feature map, it first undergoes a 1 × 1 convolution to adjust to the appropriate number of channels. Then, it is processed separately by the SRU and CRU modules. Finally, the channel number is restored through another 1 × 1 convolution, followed by a residual operation. The SRU suppresses spatial redundancy through a separate-reconstruct method, while the CRU adopts a strategy of segmentation, transformation, and fusion to reduce channel redundancy.

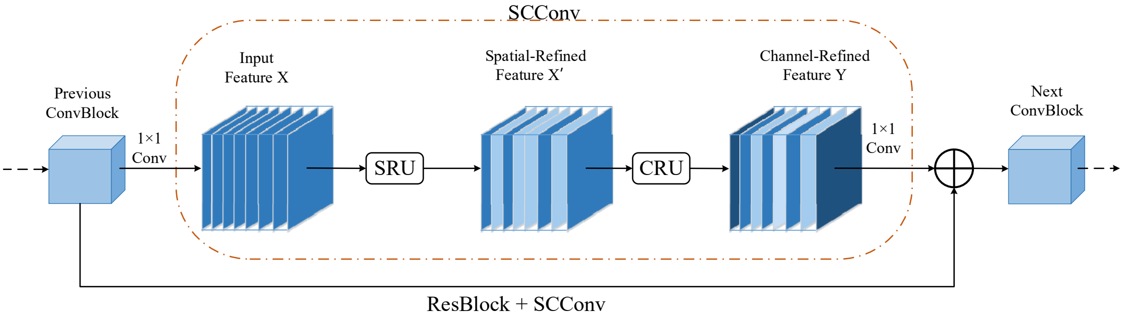


Figure 3. Network structure of SCConv.

As shown in Figure 4, the C2f structure is integrated with the SCConv module. The input features undergo a 1 × 1 convolution to adjust the channel number. Then, instead of using a 1 × 1 convolution, a split operation is employed to split the input features. All bottleneck modules in the original C2f structure are replaced with Bottleneck-SCConv modules, where the value of n is set to one. Compared with the bottleneck module, the Bottleneck-SCConv module enlarges the network's receptive field, enabling the extraction of richer facial mask features.

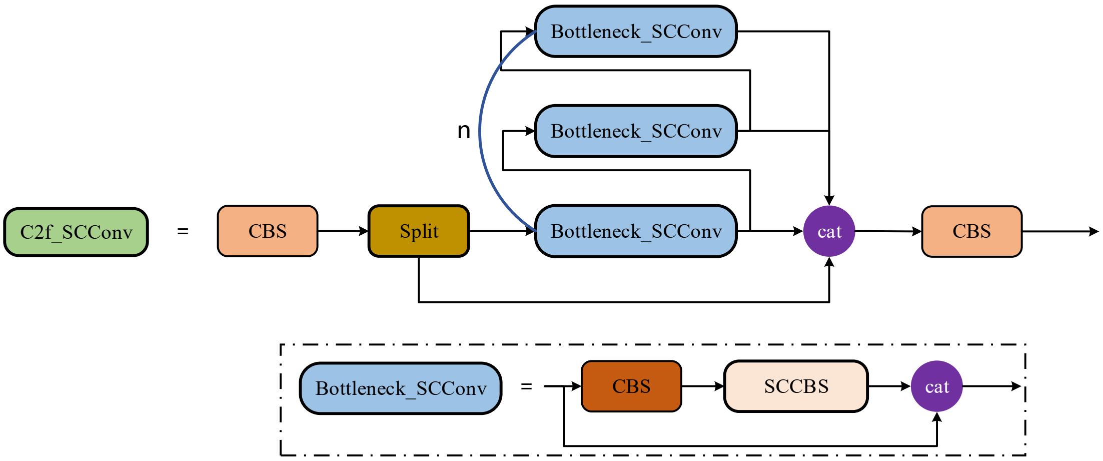


Figure 4. C2f_SCConv module.

3.2. Hybrid Attention Module (HAM)

The attention mechanism scans the entire global image to identify target regions that require focused attention, benefiting the extraction and fusion of features for small targets. However, adding attention can increase the computational cost of the model. The Convolutional Block Attention Module (CBAM), proposed by Woo et al. [25], achieves better performance through a dual attention mechanism in both the spatial and channel dimensions, but it comes with a higher computational cost. Efficient Channel Attention (ECA), proposed by Wang et al. [26], eliminates dimensionality reduction operations and utilizes one-dimensional convolution for cross-channel interaction, resulting in a lower model complexity. However, it uses less channel information, leading to less effective performance in dense scenes.

In order to more effectively utilize attention mechanisms, this paper employs a lightweight and efficient HAM [27] to enhance the model's extraction and fusion of mask features. The HAM consists of channel attention and spatial attention modules, as illustrated in Figure 5. Firstly, the channel attention module generates channel attention maps and refined channel features. Then, the spatial attention module, based on the channel attention module, optimizes the channel features along the channel axis, dividing them into two groups and generating a pair of spatial attention descriptors. Finally, refined features are produced adaptively, emphasizing crucial regions based on these spatial attention descriptors.

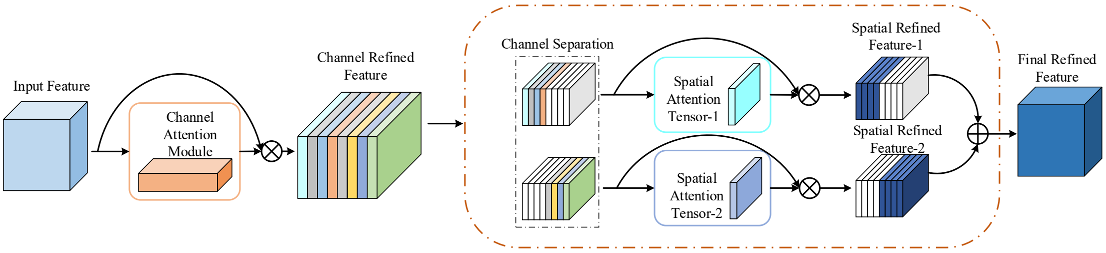

Input Feature
Channel Attention Module
Channel Refined Feature
Channel Separation
Spatial Attention Tensor-1
Spatial Attention Tensor-2
Spatial Refined Feature-1
Spatial Refined Feature-2
Final Refined Feature

Figure 5. Hybrid attention module.

3.3. SPPFCSPCA

The spatial pyramid pooling (SPP) structure can transform feature maps of any size into fixed-size feature vectors, addressing the issues caused by varying input image sizes and simultaneously enhancing the accuracy of object detection. The SPPCSPC module integrates the advantages of spatial pyramid pooling and cross-stage partial networks, resulting in a significant performance improvement compared with SPP, albeit at the cost of increased parameters. This paper proposes improvements to the SPPCSPC structure by incorporating a Hybrid Attention Module-ResBlock_HAM at the output position and transforming the original three parallel Maxpool pooling layers into a sequential arrangement, as shown in Figure 6. Despite the introduction of hybrid attention increasing the parameter count, the use of max pooling can reduce the size of the feature layers. The sequential computation helps mitigate the increase in parameters and not only enable better capture of the target object features but also significantly improve the model convergence speed and, to some extent, mitigate overfitting.

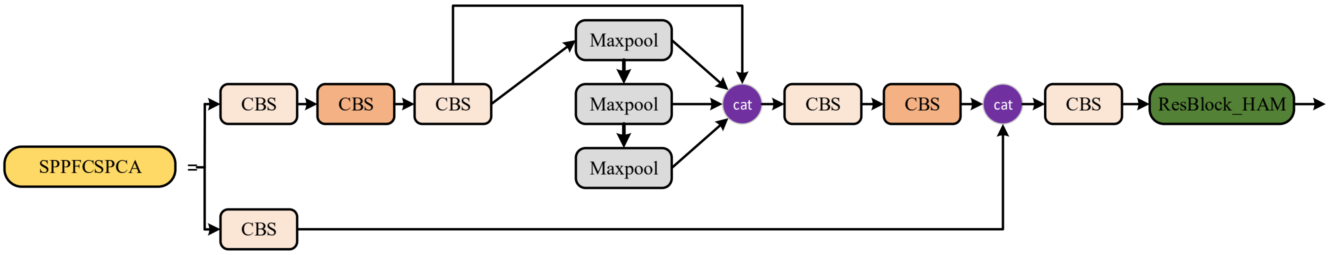


Figure 6. SPPFCSPCA module.

3.4. HAM_Detect

A well-known issue in object detection tasks is the inherent contradiction between classification and regression. This paper, inspired by the study of YOLOX and its related models, leverages the successful application of decoupled heads. The decoupled head structure is incorporated into the YOLOv7 model. However, YOLOX uses decoupled heads to separate and independently learn classification and regression, lacking task-specific learning. To address this, our paper introduces a HAM_Detect module that combines the attention mechanisms and convolution operations. The structure of this head is depicted in Figure 7. The head consists of three subheads responsible for target class prediction, confidence regression, and bounding box regression.

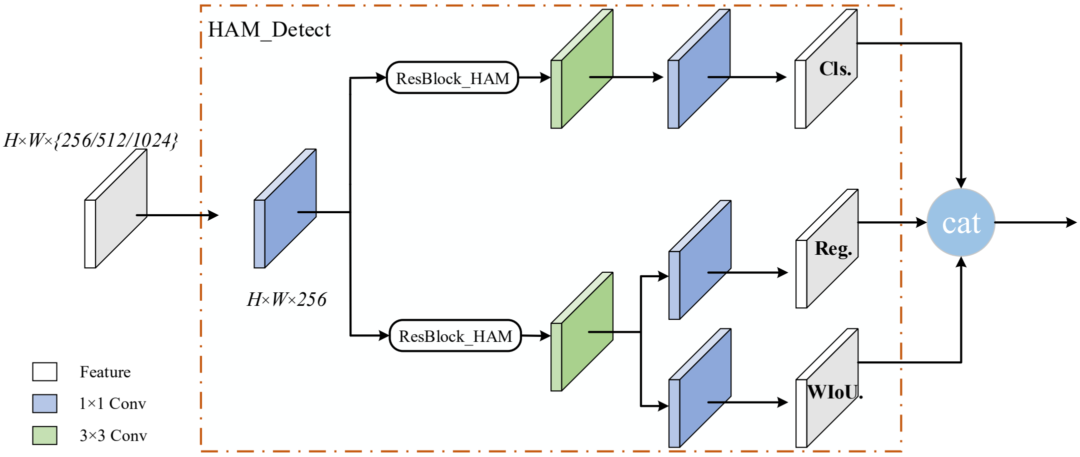


Figure 7. HAM_Detect module.

The HAM_Detect module initially takes the fused feature maps from three different scales into three corresponding scale prediction heads. The fused feature map undergoes a

1 × 1 convolution to adjust the channel count before being fed into both the classification and regression branches. In these two branches, a hybrid attention mechanism is employed to learn the weights and enhance the feature maps. Subsequently, a 3 × 3 convolution extracts high-dimensional features. The classification branch extracts texture information from the input feature map, the confidence regression branch captures background information, and the bounding box regression branch extracts positional information. Finally, a 1 × 1 convolution adjusts the channel count, and the results from each branch are concatenated through the CONCAT operation.

Overall, the HAM_Detect module combines the hybrid attention mechanism with convolution operations. The hybrid attention mechanism dynamically assigns different weights to channels and spatial locations based on different input features, focusing more on crucial features. Meanwhile, convolution operations extract features within local regions, sharing weights across different positions. By integrating the hybrid attention mechanism with convolution operations, the model can better capture the features of target objects, enhance generalization performance, and become more suitable for various real-world scenarios. Since classification and regression tasks require different features, the decoupled head employs distinct branches for learning, adaptively acquiring features based on task requirements. This helps the model more accurately discriminate and locate targets, improving object detection performance in complex scenes.

3.5. Improved Bounding Box Regression Loss Function

The bounding box regression loss function is a crucial component of the object detection loss function, playing a pivotal role in the detection accuracy of object detection models. The YOLOX model employs the IoU [28] as the bounding box regression loss function, which reflects the quality of predictions by considering the intersection over union between the predicted boxes and ground truth boxes. However, calculation methods based on geometric properties like the IoU assume that the annotated ground truth boxes in the dataset are all high-quality samples, reinforcing the fitting between the predicted and ground truth boxes. In reality, annotated datasets often contain a significant number of low-quality samples. Strengthening regression on these low-quality samples may hinder the improvement of detection model performance. Therefore, this paper proposes an improvement to the bounding box regression loss function using the WIoU [29]. It is a dynamic and non-monotonic focusing mechanism that evaluates the quality of all anchor boxes using “outlierness”. Through a gradient gain allocation strategy, it focuses on the anchor boxes of average quality, reducing the adverse gradients generated by extremely high- or low-quality samples. This improvement addresses the deficiencies of the IoU loss function in calculating the regression for low-quality samples, thereby reducing the loss function value and accelerating the model convergence speed.

The IoU schematic diagram for the WIoU bounding box loss function is illustrated in Figure 8. The predicted box is represented by $\overrightarrow{B} = [x\ y\ w\ h]$, while the ground truth box is represented by $\overrightarrow{B}_{gt} = \left[x_{gt}\ y_{gt}\ w_{gt}\ h_{gt}\right]$.

The *IoU* is used to measure the degree of overlap between the predicted box and the ground truth box and is defined as follows:

$$L_{IoU} = 1 - IoU = 1 - \frac{W_i H_i}{wh + w_{gt} h_{gt} - W_i H_i} \tag{1}$$

The penalty term for constructing the *WIoU* is defined as the center point distance ratio, expressed by

$$R_{WIoU} = \exp\left(\frac{(x - x_{gt})^2 + (y - y_{gt})^2}{W_g^2 + H_g^2}\right) \tag{2}$$

*WIoUv*1 is defined as follows:

$$L_{WIoUv1} = R_{WIoU} L_{IoU} \tag{3}$$

The outlierness, which describes the quality of the anchor boxes, is defined by

$$\beta = \frac{L_{IoU}}{\overline{L_{IoU}}} \in [0, +\infty] \tag{4}$$

where the term $\overline{L_{IoU}}$ represents the exponentially weighted moving average with momentum m, and it is utilized to construct a non-monotonic focusing coefficient using β:

$$\mathrm{r} = \frac{\beta}{\delta\alpha^{\beta-\delta}} \tag{5}$$

where α, δ represents a manually set parameter, which is experimentally validated to perform well in the YOLOv7 model when its values are set to 1.9 and 3, respectively. Therefore, L_{WIoUv3} is defined as follows:

$$L_{WIoUv3} = rL_{WIoUv1} \tag{6}$$

When the outlierness is relatively small, indicating higher anchor box quality, a smaller gradient gain should be assigned. This is gain is assigned to focus the bounding box regression on regular-quality anchor boxes. Conversely, when the outlierness is larger, indicating lower anchor box quality, a smaller gradient gain should also be assigned. This helps effectively prevent harmful gradients caused by low-quality anchor boxes. As the criteria for dividing anchor box quality dynamically change with $\overline{L_{IoU}}$, *WIoUv*3 can adapt its gradient gain allocation strategy at any moment to best suit the current situation.

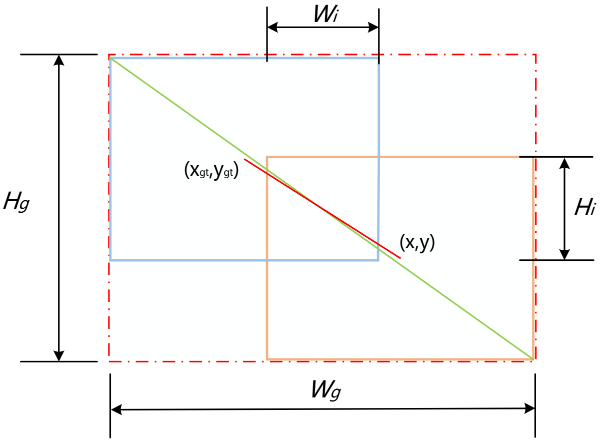


Figure 8. *IoU* schematic diagram.

4. Experiments and Results Analysis

4.1. Experimental Environment

The experiments in this paper were conducted on a cloud server with an Intel(R) Xeon(R) Platinum 8375C CPU with 90 GB of RAM and an RTX4090 GPU with 24 GB of VRAM. The development language used in this paper is the open-source Python machine learning library PyTorch deep learning framework, with Python version 3.10 and PyTorch version 2.1.0. The advantages of the PyTorch framework lie in its support for GPU-accelerated computation and large-scale floating-point operations, facilitating model training. The specific configuration of the experimental environment and the hyperparameter settings are shown in Tables 2 and 3, respectively.

Table 2. Experimental environment configuration.

Configuration Name	Value
Operating system	Linux (Ubuntu)
Programming language	Python
Training framework	Pytorch 2.1.0
Framework environment set-up	CUDA 12.1
CPU/GHz	Intel(R) Xeon(R) Platinum 8375C @ 2.10 GHz
Memory	90 GB
GPU	NVIDIA GeForce RTX4090

Table 3. Experimental hyperparameters.

Parameter	Value
lr0	0.01
lrf	0.2
weight_decay	5×10^{-4}
Batch size	16
Epochs	300

4.2. Dataset

This paper is based on the publicly available AIZOO dataset. Additionally, Python web scraping techniques were employed to gather images of individuals wearing masks in various complex scenarios from the internet, expanding the dataset. After reorganization, the dataset consisted of over 10,000 images with diverse sizes, angles, and lighting conditions. The images were reannotated using LabelImg software to create a dataset specifically for scenarios involving mask wearing in complex scenes, including images from real-life situations such as stations, schools, hospitals, and malls. Some examples are shown in Figure 9. The dataset includes three classes: faces without masks (face), masks (mask), and faces with masks (face_mask). The dataset was split into training and testing sets at a 4:1 ratio.

Figure 9. Partial examples of the dataset.

4.3. Evaluation Metrics

This paper uses precision (P), recall (R), and mean average precision (mAP) as the evaluation metrics to validate and assess the detection performance of the model. The formulas for calculating P and R are as follows:

$$P = \frac{M_{TP}}{M_{TP} + M_{FP}} \times 100\% \tag{7}$$

$$R = \frac{M_{TP}}{M_{TP} + M_{FN}} \times 100\% \tag{8}$$

For the "face" category annotated by the data in this paper, as an example, TP represents the number of face images without masks correctly detected as the "face" category after model training. *FP* is the number of images containing masks or faces with masks incorrectly detected as the "face" category. *FN* is the number of face images without masks incorrectly detected as either having masks or as faces with masks after model training. Precision (*P)* describes the accuracy of the detection model in precisely classifying this category, while recall (*R*) describes the model's ability to avoid missing detection in this category. The average precision (*AP*) is the area enclosed by the precision-recall curve and the positive half of the coordinate axes. *AP* evaluates the detection model for this category for both the precision and recall aspects. The *mAP* is the mean of *AP* for all categories in the detection model, providing an effective evaluation of the model's detection performance across all categories. The formulas for calculating *AP* and the *mAP* are shown in Equations (9) and (10). In this paper, the detection threshold for the intersection over union (IoU) is set to 0.5, making the evaluation metric *mAP*@0.5:

$$AP = \int_0^1 P(R)\mathrm{d}R \tag{9}$$

$$mAP = \frac{\sum\limits_{i=1}^{m} AP_i}{m} \tag{10}$$

4.4. Experimental Design and Result Analysis

4.4.1. Ablation Experiment

To validate the impact of the proposed improvement modules on detection performance, this paper conducts ablation experiments using YOLOv7 as the baseline. The experiments involved the C2f_SCConv module, the SPPFCSPCA spatial pyramid structure, and the attention decoupling head. As shown in Table 4, integrating the C2f_SCConv module into the feature extraction network enhanced the feature extraction performance. Both the C2f_SCConv module and the SPPFCSPCA spatial pyramid structure significantly improved the object detection performance, with the YOLOv7-B model showing a 1% increase in mAP@0.5 compared with YOLOv7. After incorporating the improved attention decoupling head, the model's detection accuracy further improved. The final mAP@0.5 reached 90.1%, representing a 1.4% improvement over YOLOv7. At the same time, there was a slight increase in the model parameters, which also slightly raised the computational performance requirements for the devices.

Table 4. Ablation experiment.

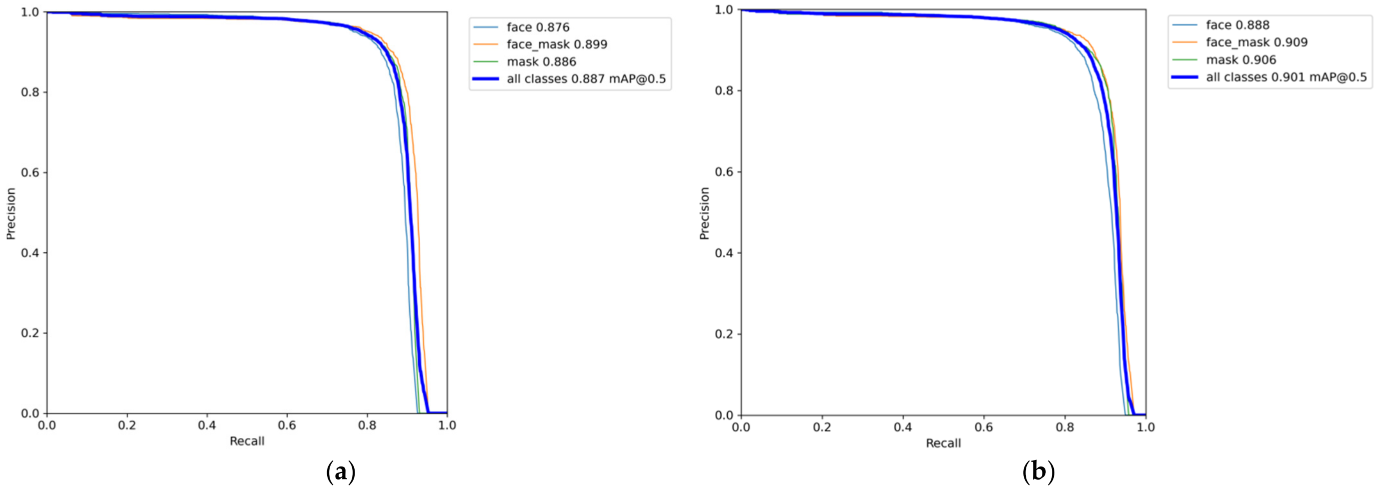
Model	C2f_SCConv	SPPFCSPCA	HAM_Detect	mAP@0.5 (%)	Parameters (M)	GFLOPs
YOLOv7	×	×	×	88.7	35.48	105.1
YOLOv7-A	√	×	×	88.9	38.58	44.2
YOLOv7-B	√	√	×	89.7	38.58	44.2
YOLOv7-C	√	√	√	90.1	54.11	135.6

4.4.2. Contrastive Experiment

(1) mAP Comparison

To demonstrate the improvement effect of the proposed method on the YOLOv7 network more intuitively, a comparison between the improved YOLOv7 network and the original YOLOv7 network in terms of the training results is presented in Figures 10 and 11. Figure 10a,b depicts the precision-recall (P-R) curves of the YOLOv7 network and the proposed method, respectively. These curves include the mean average precision (mAP) values for classes such as face, face_mask, mask, and all classes. The horizontal axis represents the recall, while the vertical axis represents the precision. Figure 11 illustrates the mAP comparison between the two networks. It is evident that both curves exhibit an upward trend with increasing training iterations, and they tend to stabilize at approximately 150 epochs of training. The improved YOLOv7 network outperformed the original

YOLOv7 network in terms of accuracy, showing advantages in mask-wearing detection in complex environments.



(**a**) (**b**)

Figure 10. P-R curves. (**a**) YOLOv7. (**b**) Our method.

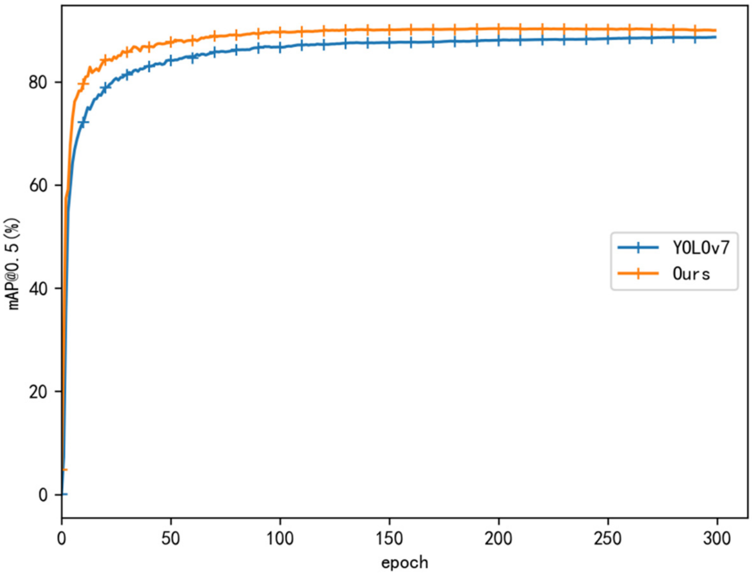


Figure 11. mAP comparison.

(2) Loss Function Convergence Comparison

Figure 12 displays the convergence of the loss function for the original YOLOv7 network and the improved YOLOv7 network proposed in this paper. This comparison provides a visual representation, showing that the loss function values for the improved YOLOv7 network were consistently lower than those of the original network. Moreover, the improved network converged faster, indicating superior performance from the enhanced model.

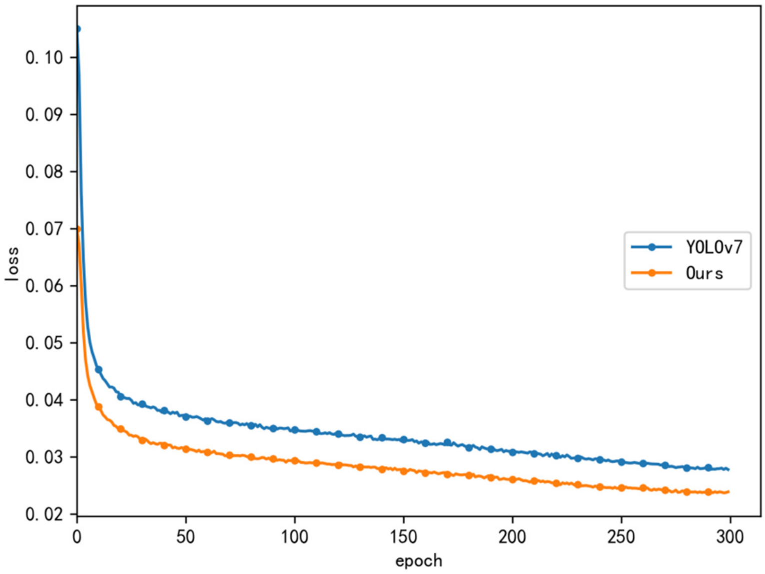


Figure 12. Loss function convergence comparison.

4.4.3. Performance Comparison and Analysis with Mainstream Detection Models

To quantitatively evaluate the performance of the improved model, this paper compares the improved model with Faster R-CNN, YOLOv5s, YOLOX, YOLOv7-tiny [30], the original YOLOv7 model, and the latest YOLOv8s model based on detection metrics. The comparative experimental results are shown in Figure 13 and Table 5. The mAP value of the improved model was higher than those of the other models, with a 6.5% improvement over Faster R-CNN, a 5.2% improvement over YOLOv7-tiny, and 1.7%, 1.6%, 1.4%, and 1.8% improvements over YOLOv5s, YOLOX, YOLOv7, and YOLOv8s, respectively. Compared with the original YOLOv7 model, the mAP values for each category were improved. Specifically, the mAP for faces without masks increased by 1.5%, that for faces with masks increased by 0.9%, and that for the mask category increased by 1.8%. Therefore, based on the comparison of mAP values, the detection performance of the improved model surpassed that of other mainstream detection models.

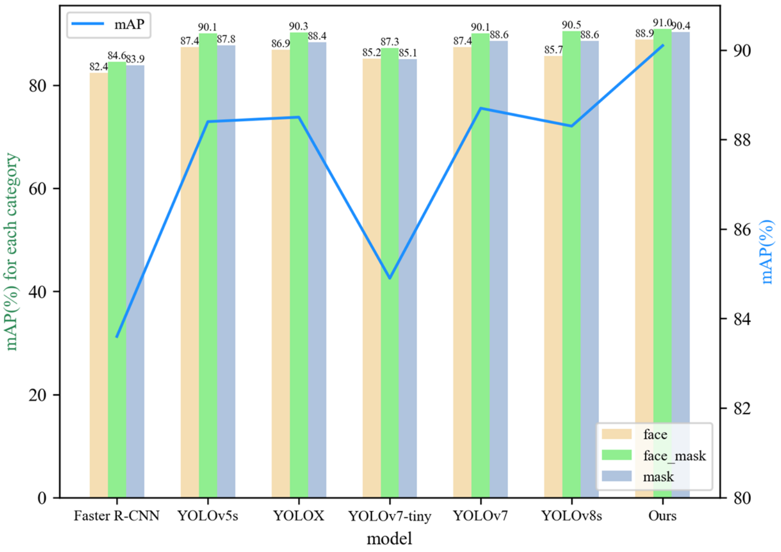


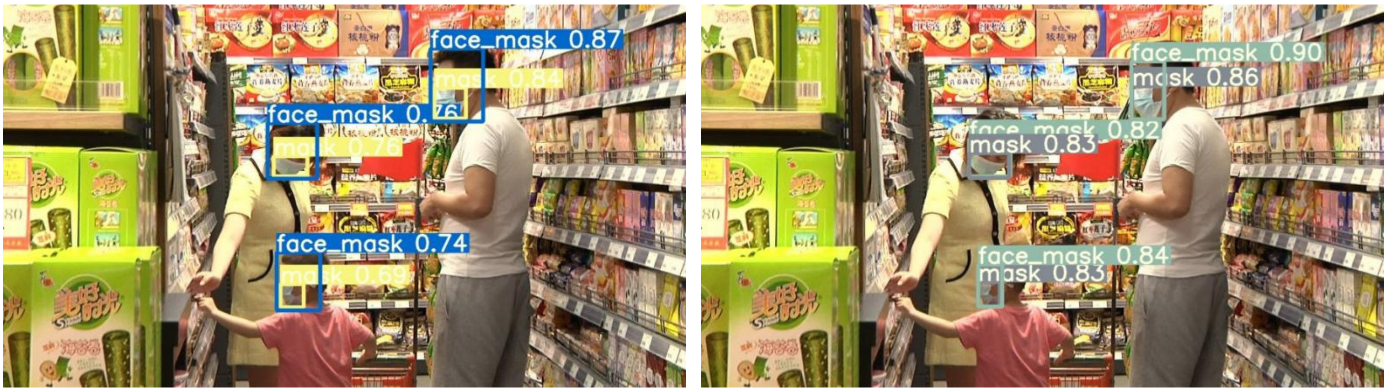
Figure 13. Comparison of mAP values for mainstream detection algorithms.

Table 5. Performance comparison and analysis with mainstream detection models.

Model	Face (%)	Face_Mask (%)	Mask (%)	mAP@0.5 (%)
Faster R-CNN	82.4	84.6	83.9	83.6
YOLOv5s	87.4	90.1	87.8	88.4
YOLOX	86.9	90.3	88.4	88.5
YOLOv7-tiny	85.2	87.3	82.1	84.9
YOLOv7	87.4	90.1	88.6	88.7
YOLOv8s	85.7	90.5	88.6	88.3
Ours	88.9	91.0	90.4	90.1

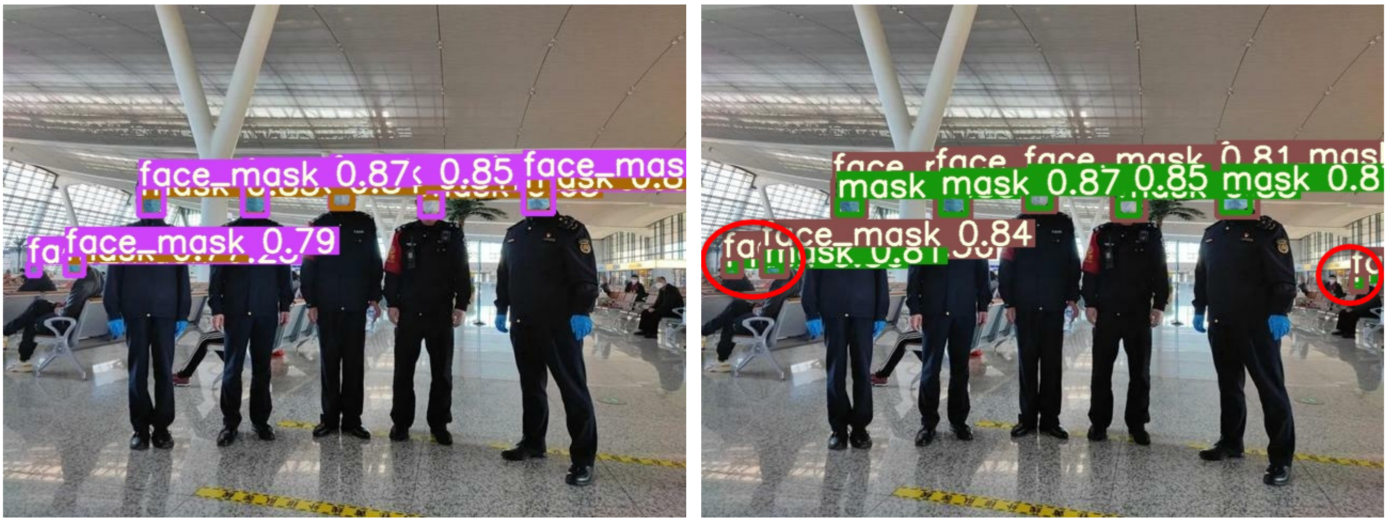
4.4.4. Visualization

To validate the visual effectiveness of the proposed improved model in various mask-wearing detection scenarios, images from the test set with complex backgrounds, small target sizes, partial scene occlusion, and dim lighting conditions were selected for detection. Figures 14–17 illustrate the visual comparison of the detection results between YOLOv7 and the proposed model in the same scenario.



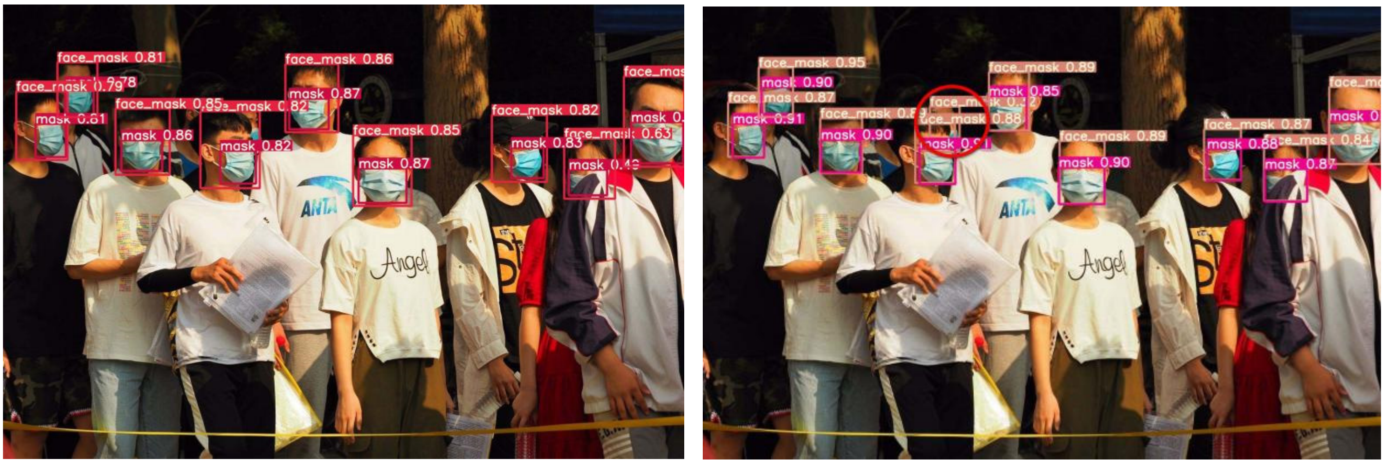
(**a**) (**b**)

Figure 14. Comparison of detection results in complex backgrounds. (**a**) YOLOv7. (**b**) Our method.



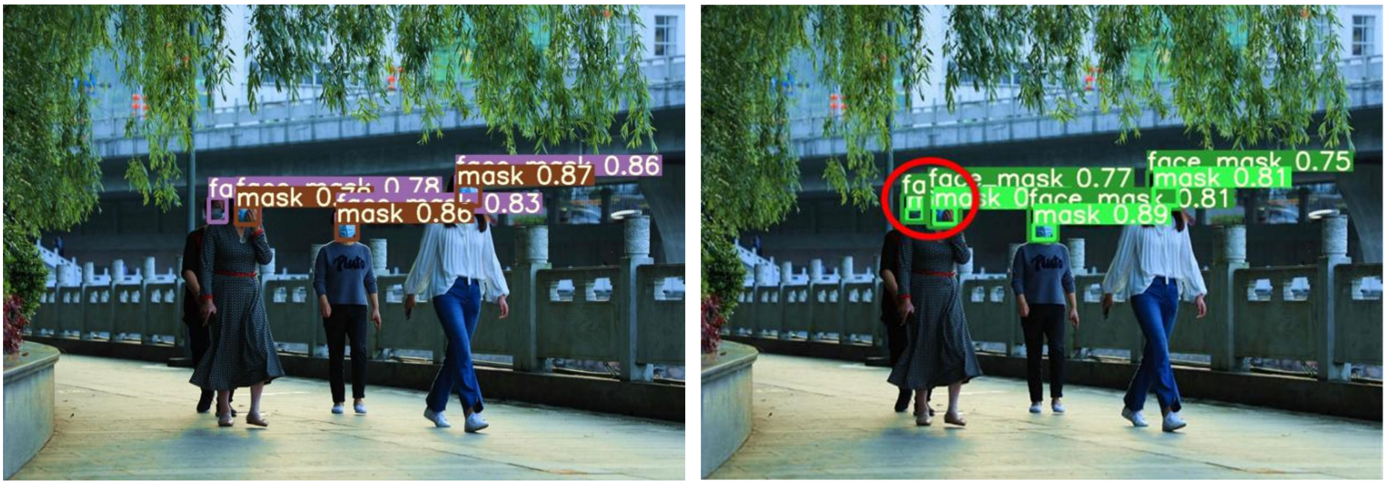
(**a**) (**b**)

Figure 15. Comparison of small target detection. (**a**) YOLOv7. (**b**) Our method.



(**a**) (**b**)

Figure 16. Comparison of detection results for partially occluded faces. (**a**) YOLOv7. (**b**) Our method.



(**a**) YOLOv7 (**b**) Our method

Figure 17. Comparison of detection results in dim lighting conditions. (**a**) YOLOv7. (**b**) Our method.

Figure 14a,b depicts a comparison of the detection results in complex environmental backgrounds. Compared with the YOLOv7 model, the proposed model in this paper demonstrated higher overall accuracy in detecting mask wearing in complex backgrounds.

Figure 15a,b depicts a comparison of the detection results for small-sized targets. YOLOv7 failed to detect faces wearing masks in the distance, and there were issues with incomplete detection of masked targets. However, the model proposed in this paper improved the detection of faces wearing masks in the distance and ensured more complete detection results, as indicated by the red circle in Figure 15b.

Figure 16a,b depicts a comparison of the detection results for partially occluded faces. YOLOv7 failed to detect faces wearing masks when partially occluded, while the model proposed in this paper could correctly detect faces wearing masks even when partially occluded, as indicated by the red circle in Figure 16b.

Figure 17a,b illustrates a comparison of the detection results in dim lighting conditions. YOLOv7 exhibited omissions in detecting faces wearing masks under dim conditions, detecting only one category. In contrast, the model proposed in this paper could correctly detect faces wearing masks under dim conditions, as indicated by the red circle in Figure 17b.

5. Conclusions

This paper proposes an improved mask-wearing detection model based on YOLOv7, aiming to address the issues of poor detection performance in complex environments, such as complex backgrounds, small target sizes, information loss due to target occlusion, false positives, and false negatives. We introduced the C2f_SCConv module as a feature extraction module, which enlarges the network's receptive field, allowing for the extraction of richer facial mask features and improving the detection performance of small targets. Additionally, we proposed the SPPFCSPCA module, which integrates a hybrid attention mechanism and optimizes the spatial pyramid pooling structure, resulting in faster model convergence. Finally, we introduced the HAM_Detect decoupled head, which incorporates a hybrid attention mechanism and optimizes the loss function, further accelerating model convergence, mitigating issues caused by target occlusion, false positives, and false negatives, and improving the detection performance of the model in complex environments.

Building upon these improvements, our model trained on the enhanced dataset from the AIZOO public dataset exhibited excellent performance. Significant improvements were observed in the model's bounding box loss function compared with the baseline YOLOv7 model. The detection accuracy reached 90.1%, showing notable improvement compared with Faster R-CNN, YOLOv5s, YOLOX, YOLOv7, YOLOv7-tiny, and YOLOv8s. In practical detection scenarios, our proposed model performed well in complex environments, such as those with complex backgrounds, small target sizes, partial scene occlusion, and dim lighting conditions. In the future, we will focus on improving the model's lightweight direction, aiming to reduce the model parameters and computational complexity and optimize the model complexity while maintaining good detection accuracy. This will facilitate deployment of the mask-wearing detection model on mobile platforms with limited computational resources.

Author Contributions: Conceptualization, G.F. and W.W.; Methodology, Q.Y. and C.T.; Software, Q.Y. and Y.L.; Validation, Q.Y.; Formal analysis, Q.Y., C.T., Y.L. and X.W.; Investigation, Q.Y. and C.T.; Resources, G.F., Q.Y. and W.W.; Data curation, Q.Y. and X.W.; Writing—original draft, Q.Y.; Writing—review & editing, G.F. and W.W.; Visualization, X.W.; Project administration, G.F. All authors have read and agreed to the published version of the manuscript.

Funding: This research and the APC was funded by Guangdong Provincial Philosophy and Social Science Planning Project, Project Name: Personalized Guidance Research Based on Cross modal High Order Reasoning Model, grant number GD23YJY08.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available on request from the corresponding author. The data are not publicly available due to contain portrait privacy information.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Bai, J. Scientific Wearing of Masks to Protect Public Health. *People's Daily*, 14 April 2023. [CrossRef]
2. Girshick, R. Fast R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 1440–1448. [CrossRef]
3. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39*, 1137–1149. [CrossRef]
4. He, K.; Gkioxari, G.; Dollár, P.; Girshick, R. Mask R-CNN. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 2961–2969. [CrossRef]
5. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single shot multibox detector. In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; Springer: Cham, Switzerland, 2016; pp. 21–37. [CrossRef]
6. Lin, T.Y.; Goyal, P.; Girshick, R.; He, K.; Dollár, P. Focal loss for dense object detection. In Proceedings of the 2017 IEEE International Conference on Computer Vision (ICCV), Venice, Italy, 22–29 October 2017; IEEE: Venice, Italy, 2017; pp. 2999–3007. [CrossRef]

7. Sun, L.; Zhang, R.; Liu, Y.; Rao, T. Mask Wearing Detection Algorithm for Dense Crowds from the Surveillance Perspective. *Comput. Eng.* **2023**, *49*, 313–320. [CrossRef]
8. Li, M.; Xiao, Q.; Han, Z. Face Mask Wearing Detection Based on Improved YOLOv5. *Comput. Eng. Des.* **2023**, *44*, 2811–2821. [CrossRef]
9. Fu, H.; Gao, J.; Che, L. Mask Wearing Detection Based on Improved YOLOv7. *Liq. Cryst. Disp.* **2023**, *38*, 1139–1147.
10. Yakovlev, A.; Lisovychenko, O. An approach for image annotation automatization for artificial intelligence models learning. *Adapt. Syst. Autom. Manag.* **2020**, *1*, 32–40.
11. Wang, C.-Y.; Bochkovskiy, A.; Liao, H.-Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for realtime object detectors. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Vancouver, BC, Canada, 18–22 June 2023. [CrossRef]
12. Wang, J.; Wang, J.; Zhang, X.; Yu, N. A Mask-Wearing Detection Model in Complex Scenarios Based on YOLOv7-CPCSDSA. *Electronics* **2023**, *12*, 3128. [CrossRef]
13. Praveen, K.S.; Naveen, K.K. Drone-based apple detection: Finding the depth of apples using YOLOv7 architecture with multi-head attention mechanism. *Smart Agric. Technol.* **2023**, *5*, 100311. [CrossRef]
14. Ding, Z.; Guo, J.; Liu, J.; Zhu, H. A mask-wearing detection algorithm based on improved YOLOv7. In Proceedings of the 2023 6th International Conference on Signal Processing and Machine Learning, Tianjin, China, 14–16 July 2023. [CrossRef]
15. Zeng, Y.; Zhang, T.; He, W.; Zhang, Z. YOLOv7-UAV: An Unmanned Aerial Vehicle Image Object Detection Algorithm Based on Improved YOLOv7. *Electronics* **2023**, *12*, 3141. [CrossRef]
16. Law, H.; Deng, J. CornerNet: Detecting objects as paired keypoints. In Proceedings of the European Conference on Computer Vision (ECCV), Munich, Germany, 8–14 September 2018; pp. 734–750. [CrossRef]
17. Duan, K.; Bai, S.; Xie, L.; Qi, H.; Huang, Q.; Tian, Q. CenterNet: Keypoint triplets for object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 6569–6578. [CrossRef]
18. Tian, Z.; Shen, C.; Chen, H.; He, T. FCOS: Fully convolutional one-stage object detection. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Republic of Korea, 27 October–2 November 2019; pp. 9627–9636. [CrossRef]
19. Wu, Y.; Chen, Y.; Yuan, L.; Liu, Z.; Wang, L.; Li, H.; Fu, Y. Rethinking classification and localization for object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10186–10195. [CrossRef]
20. Ge, Z.; Liu, S.; Wang, F.; Li, Z.; Sun, J. YOLOX: Exceeding YOLO seriesin 2021. In Proceedings of the Computer Vision and Pattern Recognition, Nashville, TN, USA, 20–25 June 2021.
21. Xu, S.; Wang, X.; Lv, W.; Chang, Q.; Cui, C.; Deng, K.; Wang, G.; Dang, Q.; Wei, S.; Du, Y.; et al. PP-YOLOE: An evolved version of YOLO. *arXiv* **2022**, arXiv:2203.16250. [CrossRef]
22. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6:a single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.
23. Jocher, G.; Chaurasia, A.; Qiu, J. YOLO by Ultralytics. Available online: https://github.com/ultralytics/ultralytics (accessed on 7 April 2024).
24. Li, J.; Wen, Y.; He, L. SCConv: Spatial and Channel Reconstruction Convolution for Feature Redundancy. In Proceedings of the 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Vancouver, BC, Canada, 18–22 June 2023; pp. 6153–6162. [CrossRef]
25. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional block attention module. In Proceedings of the 2018 European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; LNCS 11211. Springer: Cham, Switzerland, 2018; pp. 3–19.
26. Wang, Q.; Wu, B.; Zhu, P.; Li, P.; Zuo, W.; Hu, Q. ECA-net: Efficient channel attention for deep convolutional neural networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; IEEE: Piscataway, NJ, USA, 2020; pp. 11531–11539. [CrossRef]
27. Li, G.; Fang, Q.; Zha, L.; Gao, X.; Zheng, N. HAM: Hybrid attention module in deep convolutional neural networks for image classification. *Pattern Recognit. J. Pattern Recognit. Soc.* **2022**, *129*, 108785. [CrossRef]
28. Yu, J.; Jiang, Y.; Wang, Z.; Cao, Z.; Huang, T. UnitBox: An Advanced Object Detection Network. In Proceedings of the 24th ACM International Conference on Multimedia, Amsterdam, The Netherlands, 15–19 October 2016. [CrossRef]
29. Tong, Z.; Chen, Y.; Xu, Z.; Yu, R. Wise-IoU: Bounding Box Regression Loss with Dynamic Focusing Mechanism. *arXiv* **2023**, arXiv:2301.10051.
30. Yu, B.; Li, M. Face Mask Recognition Based on Improved YOLOv7-Tiny. In Proceedings of the 2023 4th International Seminar on Artificial Intelligence, Networking and Information Technology (AINIT), Nanjing, China, 16–18 June 2023; pp. 329–333. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.