electronics

MDPI

Article

Mask Detection Method Based on YOLO-GBC Network

Changqing Wang [1,2,3], Bei Zhang [1,2,3], Yuan Cao [1,2,3,*], Maoxuan Sun [1,2,3], Kunyu He [1,2,3], Zhonghao Cao [1,2,3] and Meng Wang [1,2,3]

[1] College of Electronics and Electrical Engineering, Henan Normal University, Xinxiang 453007, China
[2] Henan Key Laboratory of Optoelectronic Sensing Integrated Application, Xinxiang 453007, China
[3] Henan Engineering Laboratory of Additive Intelligent Manufacturing, Xinxiang 453007, China
* Correspondence: xyuan_cao@163.com

Abstract: For the problems of inaccurate recognition and the high missed detection rate of existing mask detection algorithms in actual scenes, a novel mask detection algorithm based on the YOLO-GBC network is proposed. Specifically, in the backbone network part, the global attention mechanism (GAM) is integrated to improve the ability to extract key information through cross-latitude information interaction. The cross-layer cascade method is adopted to improve the feature pyramid structure to achieve effective bidirectional cross-scale connection and weighted feature fusion. The sampling method of content-aware reassembly of features (CARAFE) is integrated into the feature pyramid network to fully retain the semantic information and global features of the feature map. NMS is replaced with Soft-NMS to improve model prediction frame accuracy by confidence decay method. The experimental results show that the average accuracy (mAP) of the YOLO-GBC reached 91.2% in the mask detection data set, which is 2.3% higher than the baseline YOLOv5, and the detection speed reached 64FPS. The accuracy and recall have also been improved to varying degrees, increasing the detection task of correctly wearing masks.

Keywords: mask detection; global attention mechanism; feature fusion network; content-aware reassembly of features

check for updates

Citation: Wang, C.; Zhang, B.; Cao, Y.; Sun, M.; He, K.; Cao, Z.; Wang, M. Mask Detection Method Based on YOLO-GBC Network. *Electronics* **2023**, *12*, 408. https://doi.org/10.3390/electronics12020408

Academic Editor: Gwanggil Jeon

Received: 8 December 2022
Revised: 24 December 2022
Accepted: 27 December 2022
Published: 13 January 2023

Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

The outbreak of COVID-19 has had a significant impact on the world and seriously threatens the safety of human life. COVID-19 is mainly transmitted through droplets and air. Therefore, correct wearing of masks is conducive to mitigating the spread of viruses [1], while the method of manual supervision is inefficient and there is a risk of infection. At this stage, the intelligent detection method of masks is affected by interference factors such as crowd occlusion, small target samples, and too-dense crowds. It is difficult to accurately identify whether people wear the standard. To ensure the effective solution of the above problems, it is necessary to develop a high-precision automatic detection method to detect whether people in public places wear masks correctly, which is also conducive to intelligent prevention and control of the epidemic.

In recent years, with the rapid development of deep learning, computer vision has been widely used in face detection [2,3], medical image analysis [4], traffic sign detection [5], remote sensing detection [6] and other fields. At the same time, there are many classical target detection methods, which can be roughly divided into two categories: The two-stage detection algorithm is represented by R-CNN [7] and Faster-RCNN [8], which are based on feature extraction first into candidate frames, and then classification by a convolutional neural network. These algorithms have high accuracy but slow speed. The one-stage network is represented by SSD [9], YOLO [9–14], and FCOS [15], which complete classification and regression tasks in the process of generating candidate boxes. These algorithms are fast but low in accuracy.

Although the current detection algorithms have good detection performance, they are still the following shortcomings: The feature extraction of the backbone network is insufficient, and it is impossible to learn more target region information. The single feature fusion network structure has poor fusion between different scale feature maps, and easy to lose small target region information. The traditional nearest neighbor sampling method has limitations, and it is unable to capture the rich semantic information required in intensive pre-training tasks, leading to missed detection. The applied non-maximum suppression threshold method is too violent, which leads to inaccurate detection results of the model.

To further improve the accuracy of the mask-wearing detection task, this article proposes the YOLO-GBC algorithm on the basis of YOLOv5. The experimental results show that YOLO-GBC is superior to YOLOv5 in detecting crowded situations and improves the detection accuracy of mask-wearing methods in different scenes.

The proposed model is mainly improved from the following four aspects:

1. Global attention mechanism (GAM) [16] is embedded in the backbone network to strengthen the network to focus on key information.
2. The weighted bidirectional feature pyramid network (BIFPN) [17] is used to replace the path aggregation network, and the feature information is fused by cross-layer cascading to improve the feature fusion ability.
3. Nearest neighborhood sampling is replaced with content-aware reassembly of features (CARAFE) [18] to enhance receptive field and improve computing speed.
4. In the preselection filtering stage, the attenuation method [19] of continuous Gaussian confidence reduction is integrated into the network to improve the accuracy and recall of object detection.
5. The rest of this article is arranged as follows: Section 2 describes the related work, Section 3 describes our improved approaches, Section 4 presents and analyzes the experimental, Section 5 presents qualitative evaluation, Section 6 summarizes the conclusions.

2. Related Work

2.1. Existing Work

At present, some scholars have conducted research in the field of mask detection. Pooja et al. [20] proposed to combine mask detection with the AI field and use both together through an existing monitoring system and innovative neural network algorithm to check whether a person wears a mask. Nagrath et al. [21] proposed the SSDMNV2 model for mask detection. In this model, the author uses an SSD algorithm to detect faces in real time and uses the pre trained MobileNetV2 model to predict whether people wear masks. Loey et al. [22] proposed a hybrid model of facial-mask detection based on depth and classical machine learning. ResNet-50 is used for feature extraction, and traditional machine-learning algorithms are used to classify whether a mask is worn. The experimental results show that the accuracy of the proposed model is better than that of other algorithms, but this experiment is tested on a simulated data set, and the generalization ability is poor. HE et al. [23] proposed a face-mask detection algorithm based on HSV, HOG features and SVM to check whether the face wears masks regularly, which achieves high detection accuracy but slow detection speed and poor robustness. Jiang et al. [24] proposed a different scale face detection algorithm based on Faster R-CNN to solve the problem that it is difficult to identify small-scale faces in the detection process. The algorithm has high detection accuracy, but the model structure is complex, and the detection effect for small objects is not satisfactory. Guo et al. [25] proposed the YOLOv5 CBD model, which combines the coordination attention mechanism and the weighted bidirectional feature pyramid network of complex factors in natural scenes. The detection accuracy has been significantly improved, but it only focuses on the identification of the mask itself. Wang et al. [26] proposed an improved lightweight mask detection method based on YOLOv5 to solve the problem that the existing models cannot satisfy the requirements of high precision and real-time performance.

The above detection methods cannot distinguish between wearing and standard wearing, which makes the detection model unable to achieve a good detection effect. Therefore, to solve these problems, we use the detection method based on YOLO-GBC to detect the wearing condition of the mask, so that it can meet the real-time detection requirements.

2.2. YOLOv5 Network Structure

YOLO was first proposed by Joseph Redmon in 2015. It uses the features learned by the convolutional neural network to detect objects. After continuous improvement and optimization, it has developed into a relatively mature object detection algorithm. As a fifth-generation detection algorithm, YOLOv5 has achieved high speed and accuracy and is easy to deploy on a mobile terminal. Therefore, the YOLOv5 network is selected as the basic framework in this paper, and the structure is shown in Figure 1. The overall structure consists of four parts: input, backbone, neck, and detect. In the backbone part, CBS is used to improve the network feature representation capability, while maintaining the accuracy of feature extraction, reducing memory consumption and parameters. The SPPF [27] pooling module is used to realize the information fusion of local features and global features, and, finally, three effective feature layers are obtained. In the neck part, an FPN [28] and PAN [29] feature pyramid is used to combine feature information of different scales to fuse the features of the effective feature layer to obtain more abundant effective features. In the detection part, the enhanced feature layer output is convolved by the neck network, and, finally, the classification prediction results and regression prediction results are obtained.

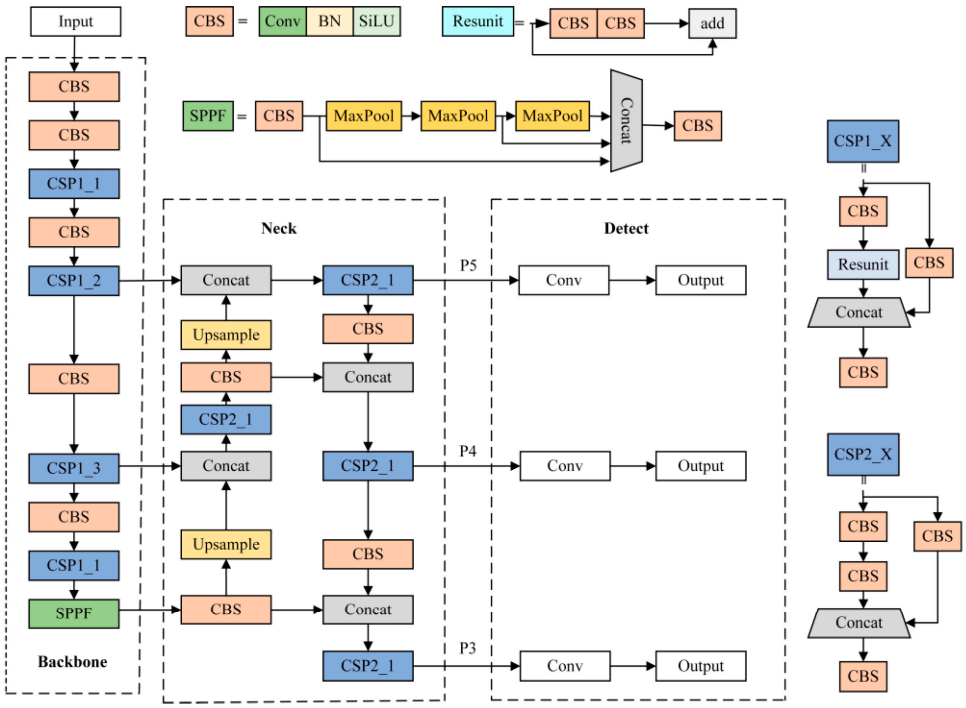


Figure 1. The structure of YOLOv5 network.

2.3. Attention Mechanism

The attention mechanism [30] first appeared in the 1990s. Volodymyr [31] applied it to the visual field in 2014. After Ashish Vaswani [32] proposed the transformer structure in 2017, the attention mechanism was widely used in the network design of NLP and CV-related issues. Attention mechanisms in neural networks can be seen as a resource allocation problem, which allocates computing resources to more important tasks. Generally speaking, the more model parameters there are, the more information is stored in the training process, which will lead to the problem of information overload. Through the introduction of an

attention mechanism, more attention can be focused on key information, reducing attention to other information, or filtering out irrelevant information. This can solve the problem of information overload when computing power is limited and improve the efficiency and accuracy of task processing. For example, SENet [33] achieved significant improvements using a 2D global pool and small computational overhead to calculate channel concerns. ECA [34] proposed a local cross-channel interaction strategy without dimension reduction, which effectively avoids the impact of dimension reduction on channel attention learning. CBAM [35] uses a channel-attention module and spatial-attention module to process the input feature layer, and quickly filter high-value information from a large amount of information to improve the extraction ability of the network. Therefore, the introduction of an attention mechanism into networks has become a mainstream idea in the current deep-learning field, which has a positive effect on improving the detection performance of the network.

2.4. Feature Pyramid Technology

A feature pyramid is a basic component of the recognition system used to detect objects of different scales. Multi-scale target recognition is a challenge for computer vision. As the feature pyramid has different resolutions at different scales, targets of different sizes can have appropriate feature representations at corresponding scales. By fusing multi-scale information, targets of different sizes can be predicted at different scales, which greatly improves the performance of the model. At present, there are many solutions to the problem of scale change in target recognition. For example, FPN can obtain fused multi-scale features, but it is not suitable for very small and large targets. The CNN layer in the hierarchical form of a feature pyramid is used in SSD to generate feature maps of different scales, but the detection performance of small targets is poor. Therefore, this article improves the feature-pyramid structure to make features at different scales have rich semantic information, so as to achieve a better fusion effect.

3. Proposed Method

This section first introduces the overall framework of the proposed YOLO-GBC network, and then explains in detail the improvement strategies of attention mechanism, feature-fusion method and post-processing method.

3.1. Overall Network Structure

In view of the problems of large-scale changes, background interference, and serious occlusion in the target region, the YOLOv5 network cannot satisfy the existing detection demands. Therefore, a high-precision mask-wearing detection algorithm model (YOLO-GBC) is established by integrating the feature extraction technology of attention mechanism into the network structure, optimizing the feature fusion network, improving the sampling method, and the method of pre-selection box screening. The network structure is shown in Figure 2.

The network imports the image to be detected from the input port. First, the channel number of the input image is expanded through the backbone network, and the embedded global attention mechanism used to extract key information, so that the network pays more attention to the target to be detected. Then, the bidirectional pyramid structure is used to fuse the high- and low-level semantic information of the feature-map output from the backbone network; the feature map containing a large amount of semantic information is input into the detection network to enhance the feature-extraction capability of the network and reduce the interference of the complex background. The fusion process adopts the CARAFE sampling method, which considers the difficult and easy balance of the samples and accelerates the convergence speed. Finally, in the post-processing stage, the loss function and Soft-NMS are used to complete the object-detection and classification tasks, improve the detection performance of the network, and achieve high-precision mask-wearing detection tasks.

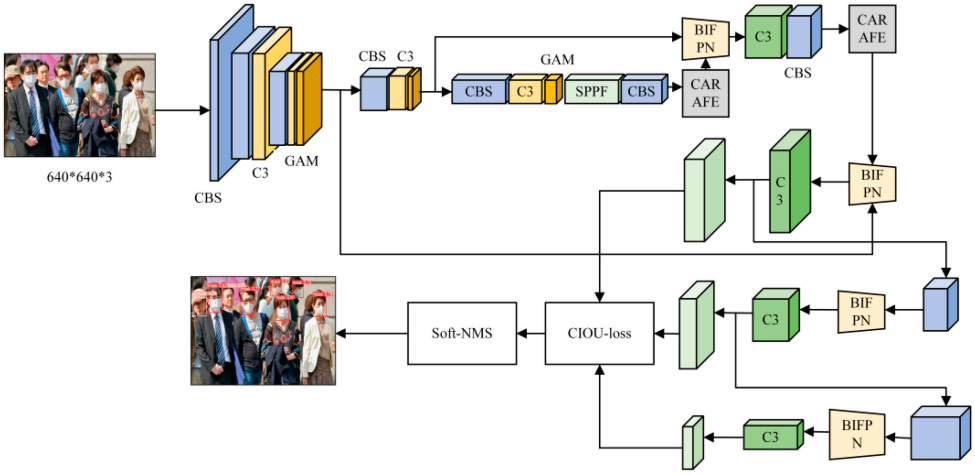


Figure 2. YOLOv5-GBC network structure.

3.2. Global Attention Mechanism

The attention mechanism simulates human visual attention, using the principle of human eye to scan the global image, find the target region that needs attention, inject more attention to that region, highlight useful information and suppress useless information, and improve feature-extraction ability. To solve the problems of loss of important information and insufficient cross-latitude feature interaction in the detection process, we introduced a global attention mechanism (GAM), which captures important features in 3D channels, space width, and space height. The overall structure is shown in Figure 3. In the channel attention module, the three-dimensional arrangement is utilized to maintain three-dimensional information, and multi-layer perceptron (MLP) is used to amplify cross-latitude channel and spatial correlation. In the spatial attention module, two convolutional layers of 7×7 are used to perform feature fusion to enhance the attention of channels to spatial information. By reducing information dispersion and enlarging the ability of global information interaction, the network expressiveness is enhanced, which makes the model more targeted to the target region, and, thus, improves the detection performance of the network.

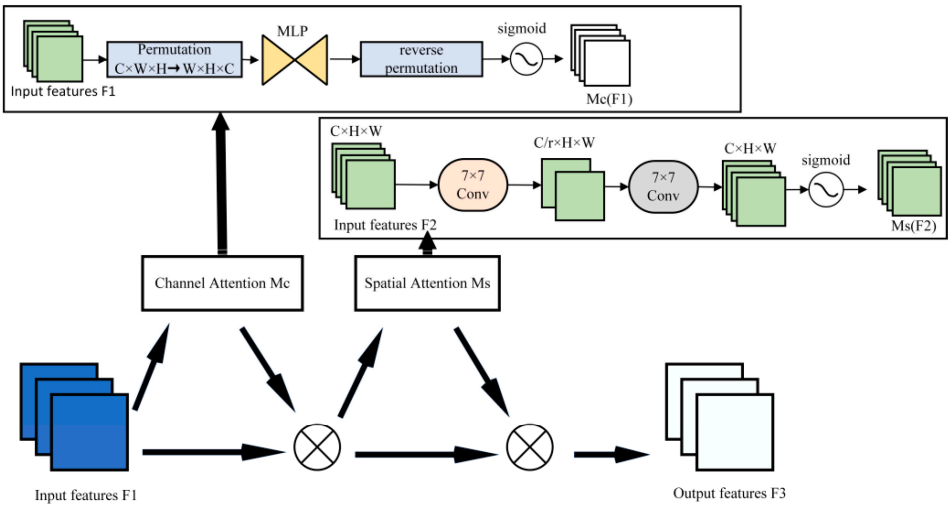


Figure 3. GAM module structure.

Given the input feature mapping $F_1 \in R^{C \times H \times W}$, the expressions for intermediate state F_2 and input state F_3 are as follows:

$$F_2 = M_C(F_1) \otimes F_1 \quad (1)$$

$$F_3 = M_S(F_2) \otimes F_2 \tag{2}$$

where M_C is the channel attention map and M_S is the spatial attention map; $\otimes$ represents the multiplication of pixels.

The attention mechanism is introduced into the backbone network, and the network changes are shown in Figure 4. The backbone network of YOLOv5 is shown in Figure 4a, and the backbone network of YOLO-GBC is shown in Figure 4b. It can be seen from the figure that the GAM module is integrated into the feature fusion of the backbone network. First, the feature map is convolved by the size of 6×6 in the first layer to expand the multiple of the input channel and extract features. Then, the fusion operation is carried out by convolution standardized activation function, C3 module learns residual features, and GAM module extracts weighted features. Finally, several pooled cores of different sizes in the SPPF module are cascaded to combine the characteristic maps of various sensing regions, which improves the receptive field of the network.

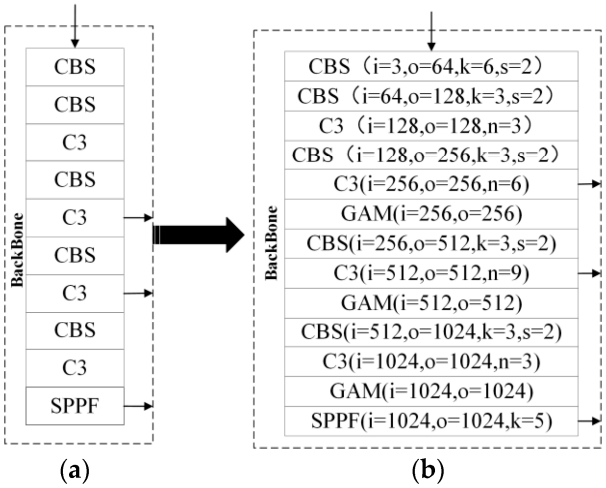


Figure 4. Comparison of backbone network improvements. (**a**) YOLOv5 backbone network structure. (**b**) YOLO-GBC backbone network structure.

3.3. C-FPN Network

YOLOv3 adopts unidirectional fusion network FPN for feature fusion, and its structure is shown in Figure 5a. This structure is easy to lose the underlying target information in the process of transmitting high-level semantic features in a multi-layer network. YOLOv5 adopts a bidirectional network combining FPN and PAN for feature fusion, and its structure is shown in Figure 5b. FPN conveys strong semantic features from the high level to the low level, and PAN conveys strong positioning features from the low level to the high level. The two combine to aggregate different features, effectively enhancing the feature fusion capability of the network. However, from the structure of PAN and FPN, each input of PAN is the feature information processed by FPN, in which is easy to lose the original feature information in feature fusion, affecting the detection effect. Additionally, the FPN architecture employs the nearest-neighbor method for sampling, which only determines the sampling center based on the spatial location between pixels and cannot capture the rich semantic information required for intensive prediction tasks. To solve the problem of insufficient feature fusion and easy loss of target information in Figure 5a,b, the BIFPN structure is proposed for feature fusion. The structure is shown in Figure 5c. BIFPN constructs a two-way channel from top to bottom and from bottom to top. For information from different scales of the backbone network, in the fusion process between different scales, the same feature resolution scale is up sampled, and down sampled, and horizontal connections are added between the original input and output nodes of the same feature to fuse more features without increasing costs. It can reduce the information loss in the transmission process by jumping the connecting line and integrate more features to achieve cross layer and multi-scale information exchange.

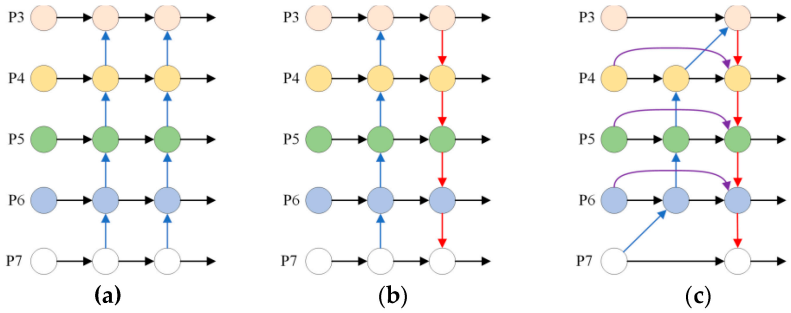


Figure 5. Schematic diagram of feature fusion structure. (**a**) FPN. (**b**) FPN + PAN. (**c**) BIFPN.

To obtain better feature fusion effect, a weighted bidirectional pyramid feature fusion network C-FPN structure is proposed, as shown in Figure 6. This structure enables the original feature information to participate in the training process, so that the information at all stages can be fully integrated, achieving a higher level of weighted feature fusion and bidirectional cross-scale connection, strengthening the specifics in the high-level feature layer, reducing feature information loss during convolution, and increasing the effectiveness of feature fusion. Content-aware reassembly of features (CARAFE) is used to capture feature information in the feature pyramid fusion network, which consists of a kernel prediction module and a content-aware reassembly module. For input size H × W × C, first, the kernel prediction module is utilized to generate the reassembled convolution kernel, then the content-aware reassembly module used to complete the sampling, and finally the shape σ H × σ W × C output characteristic diagram is obtained. The CARAFE sampling method is based on the input content, which can make better use of the surrounding information, aggregate rich semantic information, and obtain a larger perceptual field.

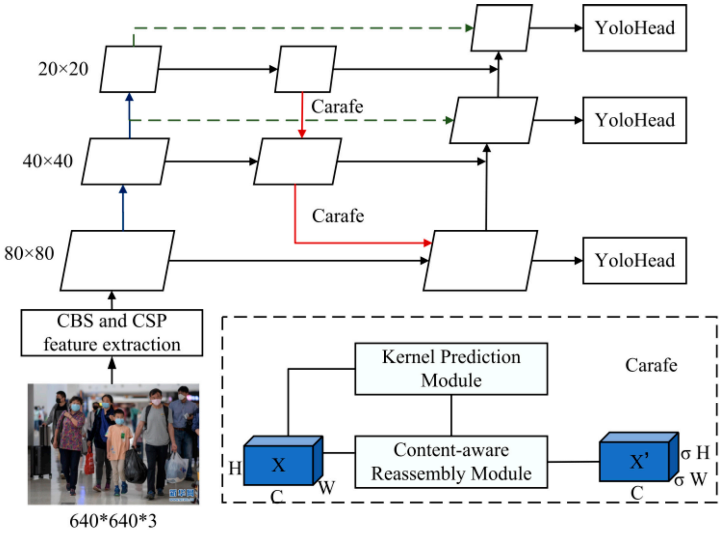


Figure 6. C-FPN structure diagram.

3.4. Soft Non-Maximum Suppression

YOLOv5 uses the NMS method to screen candidate frames. The problem is to forcibly delete the detection box whose *IOU* is greater than the set threshold. If the real target appears in the overlapping zone, the object detection will fail, thus reducing the average detection rate of the algorithm. The expressions are as follows:

$$S_i = \begin{cases} S_i, I_{IOU}(M, b_i) < N \\ 0, I_{IOU}(M, b_i) \geq N \end{cases} \tag{3}$$

where M is the highest scoring predictor box, b_i is the other predictor box except M, S_i is the score value of b_i, and N is the set threshold.

To solve the problem of incorrect deletion of candidate frames, this article proposes a method of gradually reducing confidence instead of directly deleting candidate frames whose *IOU* is greater than the threshold in NMS. The expressions are as follows:

$$S_i = \begin{cases} S_i, I_{IOU}(M, b_i) < N \\ S_i e^{-\frac{IOU(M,b_i)}{\sigma}}, I_{IOU}(M, b_i) \geq N \end{cases} \quad (4)$$

where $e^{-\frac{IOU(M,b_i)}{\sigma}}$ is the Gaussian penalty function and σ is the set hyper parameter. According to Equation (4), when the overlapping region of the maximum score box and the prediction box is greater than the threshold, the Soft-NMS uses the weighted function attenuation method to make the score S_i of the prediction box smaller, which is different from the traditional NMS method that is directly set to 0, and finally screens out the candidate boxes that are greater than the confidence threshold N. The Soft-NMS uses the weighting function to attenuate the fraction of the prognosis frame overlapped with M, effectively averting the missed detection caused by forced deletion of the prediction frame and improving the detection accuracy.

The comparison of NMS and Soft NMS processes is shown in Figure 7. The left half of the figure represents the screening process of NMS candidate frames. First, the candidate boxes predicted by the network according to the score are arranged, and the prediction box M with the highest score as the target taken. Then, the *IOU* value between target M and other prediction boxes b_i is calculated. Finally, the size relationship between *IOU* and the set threshold N is judged. If the *IOU* is greater than the set threshold, the prediction box is deleted. If the *IOU* is less than the set threshold, the cycle filtering process is continued until all prediction boxes are deleted and the results output. The Soft-NMS dotted box on the right represents the screening process of the Soft-NMS prediction frame. If the *IOU* is greater than the set threshold, the fraction of the predicted frame is attenuated. If the score of the predicted frame is still higher than the confidence threshold O after the penalty, the predicted frame is retained. In real-life scenarios, Soft-NMS is used to handle the false deletion of candidate frames, thus greatly reducing the false detection rate of masks.

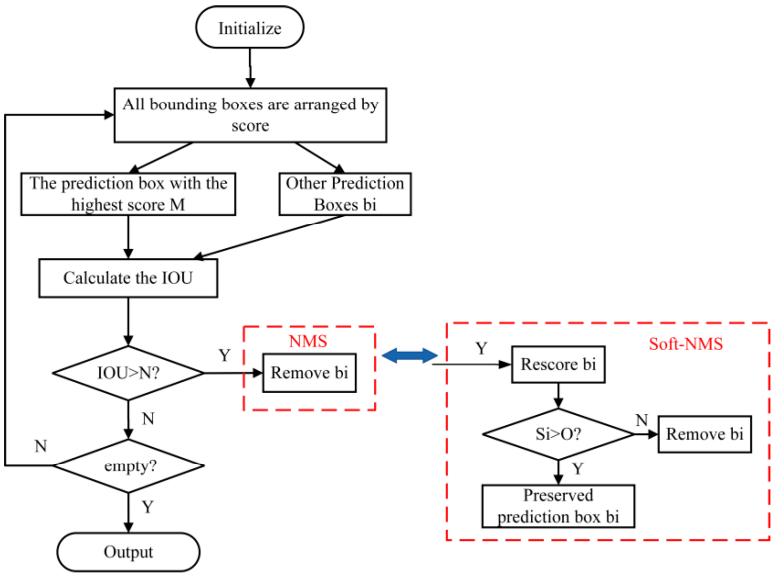


Figure 7. Comparison of NMS and Soft-NMS processes.

4. Experiments and Results

4.1. Experimental Environment and Settings

The experimental platform is Win10 operating system, CPU is Intel i7-12700, GPU is NVIDIA GeForce RTX 3090Ti, and Python 3.6 framework is used. In the model training, the epoch is set to 300, the batch size is 64, the learning rate is 0.01, the weight decay is

0.005, the momentum is 0.9, and the optimizer Adam is selected to calculate the adaptive learning rate.

4.2. Dataset

The data set in this article stems from three ways: flying pulp, web crawling, and actual collection. A total of 11317 pictures were in the data set, including 7786 training sets and 3531 test sets. To realize the detection of mask wearing in complex environments and effectively prevent the fitting phenomenon caused by too singular data set and too large a proportion of targets, the designed data set includes places with dense traffic in different scenarios and the training set and verification set are allocated according to 7:3. The data set contains mask, nomask, and poor label files. Classification standards are shown in Figure 8. As shown in Figure 8a, the face where the nose and mouth are covered is marked as a mask, the face without a mask as shown in Figure 8b is marked as nomask, and the nose and mouth are not covered as shown in Figure 8c,d is marked as poor. Use Labelimg implement to mark data to form a self-made data set.



(**a**) (**b**) (**c**) (**d**)

Figure 8. Dataset classification criteria. (**a**) Mask classification criteria. (**b**) Nomask classification criteria. (**c**) Poor classification criteria. (**d**) Poor classification criteria.

The learning of CNN networks requires massive amounts of data as support, and the process of acquiring data requires a lot of time and vigor. Therefore, some data enhancement methods need to be adopted to expand the image and improve the generalization ability of the model. In this article, we choose the Mosaic data expansion technique [36], in which four images are randomly cropped, scaled, arranged, and then stitched into a new image. By this means, a batch of pictures of difficult scenes is added to the data set, which improves the generalization ability of the model in complex scenes.

4.3. Evaluation Metrics

The experiment introduces four indicators of precision, recall rate, mAP, and FPS as the quantitative standards for judging the detection effect of the model, and the detection results were analyzed and compared. As the frame rate per second, FPS represents the number of images that can be processed per second to evaluate the speed of object detection. The evaluation indicators are calculated as follows:

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

where true positives (*TP*) denote the number of correctly identified target regions. False positives (*FP*) denote the number of improperly identified target regions. False negatives (*FN*) denote the number of unidentified target regions. *Precision* predicts the proportion of correct targets among all targets. *Recall* predicts the correct targets proportion among all positive targets.

$$\mathrm{m}AP = \frac{\sum AP(c)}{c} = \frac{\sum_{c} \int_0^1 P(R) d(R)}{c} \tag{7}$$

where *mAP* represents the mean value of the average precision under all categories to comprehensively measure the multi-category detection effect. *C* is the number of target categories.

4.4. Training Results and Analysis

To observe the changes in network training loss and average accuracy more clearly. YOLOv5 and YOLO-GBC networks were trained for 300 rounds using the same data set. Whether the training trend is correct is determined through the loss function and the real-time solution of mAP to ensure that the model tends to converge normally with the increase of iteration times. If the training process curve goes abnormally, we can choose to interrupt the training to avoid wasting time. Figure 9 shows the loss curve and mAP curve of the baseline network and the proposed model. The change in loss curve is shown in Figure 9a. With the increase in epoch, the loss value continues to decrease. After 100 rounds, the loss value tends to be stable, and the model training converges normally with good stability. The change in mAP is shown in Figure 9b. The YOLO-GBC curve grew rapidly in the first 50 rounds and reached a high point in the 50th round. After the 50th round, the curve fluctuates slightly and remains stable. Compared with the basic model, the mAP curve of the proposed network is always above the YOLOv5 curve, indicating that the accuracy of the proposed model is higher than that of the original model, which proves the validity of the improved strategy.

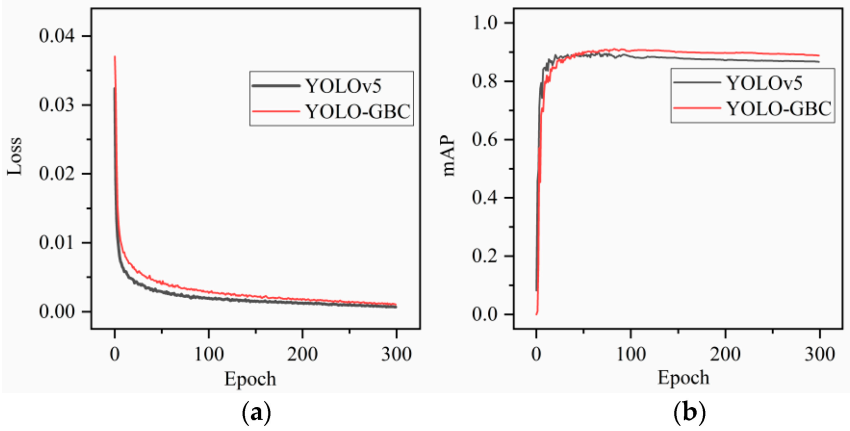


Figure 9. Network training result. (**a**) loss curve; (**b**) mAP curve.

4.5. Comparison Experiment

To comprehensively evaluate the YOLO-GBC algorithm proposed in this article, five classical network models were compared and analyzed on the same data set. Precision, recall, mAP, and FPS were used as evaluation indicators to accurately evaluate various detection methods. The experimental results are shown in Table 1.

Table 1. Performance comparison results of different algorithms.

Model	Map/%	Precision/%	Recall/%	FPS/f/s
SSD(VGG)	72.6	74.4	71.6	62
Faster-RCNN	77.2	78.4	75.9	54
YOLOv4	86.4	87.9	84.5	72
YOLOv5	86.4	89.8	85.2	80
YOLO-GBC	91.2	91.1	86.4	64

As shown in Table 1, the YOLOv5 model has higher detection accuracy and recall compared to the two-stage Fast-RCNN and the one-stage SSD, YOLOv3, and YOLOv4. In comparison to SSD, Faster-RCNN, and YOLOv4, the YOLOv5 model increased by 16.3%, 11.7%, and 2.5% in mAP and 13.6%, 9.3%, and 0.7% in recall. The YOLOv5 model has the

fastest detection speed, with FPS reaching 80. Compared with the YOLOv5 model, the proposed YOLO-GBC model has improved the accuracy and recall rate by 2.3% and 1.2%, respectively. The FPS is slightly lower than the baseline network, but still meets the needs of real-time detection. By comprehensive comparison, YOLO-GBC is superior to the other five classical object detection algorithms.

4.6. Ablation Experiment

To verify the effectiveness of each module, the improved modules were introduced into the network one by one by combining the improvement strategy based on the baseline network YOLOv5. The experimental results were observed after training to verify the necessity of corresponding improved modules. Table 2 shows the results of the experiment.

Table 2. Results of ablation experiments.

Experiment	GAM	C-FPN	Soft-NMS	mAP/%	Precision/%	Recall/%	FPS/f/s
1				88.9	89.8	85.2	80
2	√			90.2↑	90.6↑	86.6↑	72
3		√		89.9↑	90.5↑	86.5↑	80
4			√	90.6↑	91.8↑	84.7↓	88
5	√	√		90.5↑	89.2↓	87.2↑	67
6	√		√	90.9↑	90.4↑	86.2↑	76
7		√	√	90.7↑	90.9↑	86.1↑	72
8	√	√	√	91.2↑	90.5↑	86.4↑	64

Table 2 effectively analyzes all the improvement strategies. Experiments 2–8 show that each improvement strategy contributes to the model to varying degrees. Experiment 1 uses the baseline network YOLOv5 for network training, and the experimental data is used as the control group. The mAP is 88.9%, and the FPS is 80. Experiment 2 incorporates the GAM in the backbone network, which improves the mAP value by 1.3% compared with experiment 1, indicating that the introduction of the attention module makes the network feature extraction more effective. Experiment 3 replaces the bidirectional pyramid fusion network in the current algorithm with a weighted feature pyramid fusion network containing a lightweight sampling operator, which boosted the mAP by 1%, indicating that the improved feature fusion method makes network training more in-depth. In experiment 4, the NMS in the prediction frame screening stage was replaced by Soft-NMS, and the mAP was increased by 1.6%, indicating that the frame score attenuation method with reduced continuous Gaussian confidence is more accurate for prediction frame filtering. Experiment 5 added C-FPN based on experiment 2, and the mAP increased by 0.3%, indicating that the cross-regional feature fusion method has a better fusion effect. Experiment 6 introduced Soft-NMS based on experiment 2, and mAP increased by 0.7%, demonstrating that Soft-NMS has a weighted effect on improving model detection performance. Experiment 7 added Soft-NMS based on experiment 3, and the mAP increased by 1.8%, indicating that the introduction of Soft-NMS can improve the detection accuracy of the model. Experiment 8 introduces all the improved strategies. Although a small amount of FPS was sacrificed, the mAP and recall rates were 2.3% and 1.2% higher than experiment 1, respectively. The YOLO-GBC network is superior to the original model in terms of performance indicators, which proves the advantages of the proposed network in the detection work, thus satisfying the real-time detection requirements of masks in complex scenarios.

5. Qualitative Evaluation

5.1. Focus Comparison

To prove the detection performance of the proposed YOLO-GBC network more intuitively, this article uses Grad-CAM technology to compare the detection results of different models, so as to observe the specific differences between YOLOv5 and YOLO-GBC methods. The visual comparison results are shown in Figure 10, where the highlighted region

reflects the region of interest of the model. It can be seen from the first column that YOLOv5 network not only focuses on the target region, but also other regions beyond the core region of the face. The focus on the target region is not concentrated, but in the visualization results of the proposed YOLO-GBC model, the highlighted region is concentrated on the face, and the focus region is relatively obvious. It can be seen from the second and third columns that the YOLOv5 model is severely disturbed by the background environment, and the focus region is chaotic, but the anti-interference ability is significantly enhanced, and the focus region is more accurate in the proposed network. It can be seen from the fourth column that YOLOv5 network has obviously insufficient attention to edge targets, but in the proposed network, the brightness of edge detection objects is significantly improved. Therefore, the YOLO-GBC model is more concerned about the target region through the verification of the Grad CAM visualization technology.

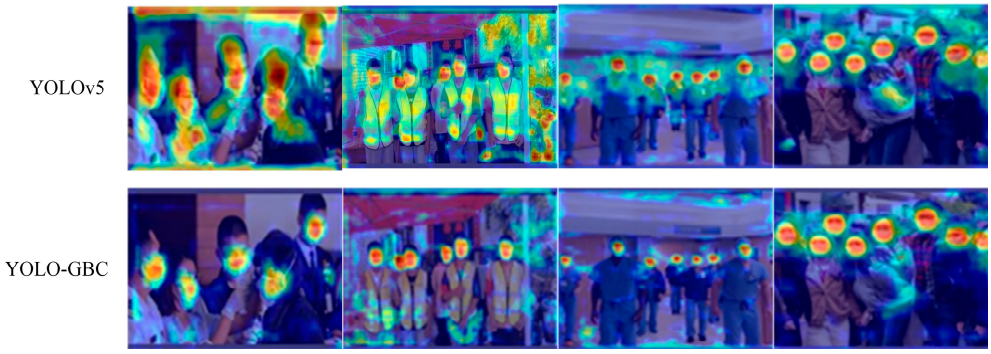


Figure 10. Comparison of the heat map of concern.

5.2. Results Comparison and Display

Figure 11 shows the recognition effects of YOLOv5 and YOLO-GBC in real scenes. The yellow box region indicates the missed target. Figure 11a column shows the case of a crowded scene, and the proposed algorithm can effectively identify the undetected occluded targets. Figure 11b column shows the case of face occlusion, and the proposed algorithm can detect the undetected region. The Figure 11c column shows the case of small target size, and the proposed algorithm is capable of precisely detecting tiny targets at the edge. Figure 11d column shows the case of severe occlusion and small target, the proposed algorithm is still capable of precisely detecting the target region. Therefore, YOLO-GBC can effectively solve the problem of missed detection and false detection of YOLOv5 detection algorithm in the case of dense crowd, severely occluded targets, and small edge targets. At the same time, YOLO-GBC has good detection and recognition effects, improves the detection performance of the target in the actual scene, and further proves the superiority of the algorithm.

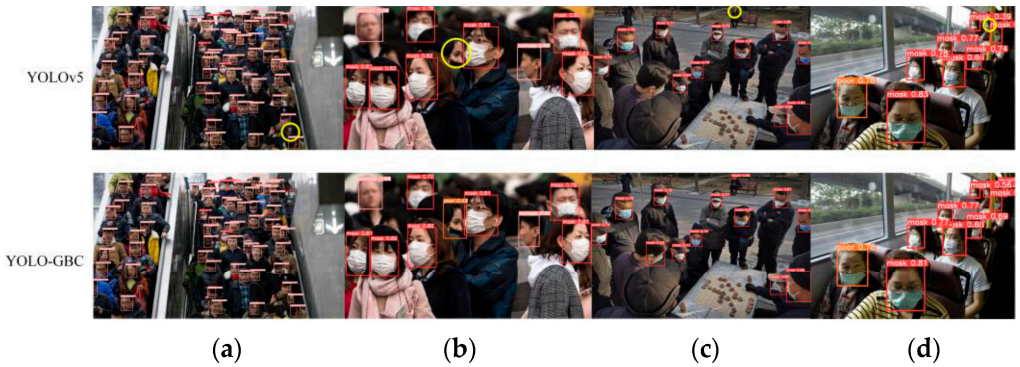


Figure 11. Comparison of detection effect. (**a**) crowded scene. (**b**) face is severely blocked. (**c**) small target. (**d**) serious occlusion and small target.

6. Conclusions

In this article, we propose a standard mask-wearing identification method based on the YOLO-GBC network to address the issues of inaccurate mask detection algorithms and high missed detection rates in real scenes. The collaborative GAM module designs the backbone network, screens high-quality information, and improves the feature-extraction ability of the model. The C-FPN multi-scale feature fusion network is used to enhance the sensing range and increase cross-channel semantic-information extraction. The Soft-NMS method is adopted to reduce missed detection and further enhance the recognition of hidden targets. The experimental results indicate that YOLO-GBC can effectively solve the problem of false detection caused by dense populations, small target scale, and local region blockage. The validity of the YOLO-GBC method is verified by comparing with the classical detection algorithms. At the same time, the algorithm's detection accuracy and recall rate were massively enhanced, which is conducive to promoting the intelligent real-time detection of mask-wearing in actual scenes and is crucial for the prevention and control of epidemics.

In the future, we will make every effort to build a complete mask detection data set to address the overfitting problem caused by the severe sample imbalance of the current public data set. Meanwhile, based on the existing work, we will continue to optimize and improve the network structure for better detection performance and faster detection speed. In addition, a hardware platform is built to combine the algorithm with hardware devices to improve the practicality.

Author Contributions: Conceptualization, C.W. and B.Z.; methodology, C.W.; software, B.Z.; validation, B.Z.; formal analysis, C.W.; investigation, B.Z.; resources, C.W.; data curation, B.Z.; writing—original draft preparation, B.Z.; writing—review and editing, C.W., Y.C., M.S., K.H., Z.C. and M.W.; supervision, Y.C. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the National Natural Science Foundation of China, grant number 52177004.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

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

References

1. Liang, M.; Gao, L.; Cheng, C.; Zhou, Q.; Uy, J.P.; Heiner, K.; Sun, C. Efficacy of Face Mask in Preventing Respiratory Virus Transmission: A Systematic Review and Meta-Analysis. *Travel Med. Infect. Dis.* **2020**, *36*, 101751. [CrossRef] [PubMed]
2. Kortli, Y.; Jridi, M.; Al Falou, A.; Atri, M. Face Recognition Systems: A Survey. *Sensors* **2020**, *20*, 342. [CrossRef]
3. Almabdy, S.; Elrefaei, L. Deep Convolutional Neural Network-Based Approaches for Face Recognition. *Appl. Sci.* **2019**, *9*, 4397. [CrossRef]
4. Wu, M.; Awasthi, N.; Rad, N.M.; Pluim, J.P.; Lopata, R.G. Advanced Ultrasound and Photoacoustic Imaging in Cardiology. *Sensors* **2021**, *21*, 7947. [CrossRef] [PubMed]
5. Arcos-Garcia, A.; Alvarez-Garcia, J.A.; Soria-Morillo, L.M. Evaluation of Deep Neural Networks for Traffic Sign Detection Systems. *Neurocomputing* **2018**, *316*, 332–344. [CrossRef]
6. Deng, Z.; Sun, H.; Zhou, S.; Zhao, J.; Lei, L.; Zou, H. Multi-Scale Object Detection in Remote Sensing Imagery with Convolutional Neural Networks. *ISPRS J. Photogramm.* **2018**, *145*, 3–22. [CrossRef]
7. Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 580–587.
8. 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]
9. Liu, W.; Anguelov, D.; Erhan, D.; Szegedy, C.; Reed, S.; Fu, C.Y.; Berg, A.C. SSD: Single Shot MultiBox Detector. In *European Conference on Computer Vision*; Springer: Cham, Switzerland; Amsterdam, The Netherlands, 2016; pp. 21–37.
10. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You Only Look Once: Unified, Real-Time Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
11. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7263–7271.

12. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 1–4.
13. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
14. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable Bag-of-Freebies Sets New State-of-the-Art for Real-Time Object Detectors. *arXiv* **2022**, arXiv:2207.02696.
15. 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, Korea, 27 October–2 November 2019; pp. 9627–9636.
16. Liu, Y.; Shao, Z.; Hoffmann, N. Global Attention Mechanism: Retain Information to Enhance Channel-Spatial Interactions. *arXiv* **2021**, arXiv:2112.05561.
17. Tan, M.; Pang, R.; Le, Q.V. EfficientDet: Scalable and Efficient Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 10781–10790.
18. Wang, J.; Chen, K.; Xu, R.; Liu, Z.; Loy, C.C.; Lin, D. Content-Aware ReAssembly of FEatures. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27 October–2 November 2019; pp. 3007–3016.
19. Bodla, N.; Singh, B.; Chellappa, R.; Davis, L.S. Soft-NMS—Improving Object Detection with One Line of Code. In Proceedings of the IEEE International Conference on Computer Vision, Honolulu, HI, USA, 21–26 July 2017; pp. 5561–5569.
20. Pooja, S.; Preeti, S. Face Mask Detection Using AI. In *Predictive and Preventive Measures for COVID-19 Pandemic*; Springer: Singapore, 2021; pp. 293–305.
21. Nagrath, P.; Jain, R.; Madan, A.; Arora, R.; Kataria, P.; Hemanth, J. SSDMNV2: A Real Time DNN-Based Face Mask Detection System Using Single Shot Multibox Detector and MobileNetV2. *Sustain. Cities Soc.* **2021**, *66*, 102692. [CrossRef] [PubMed]
22. Loey, M.; Manogaran, G.; Taha, M.H.N.; Khalifa, N.E.M. A Hybrid Deep Transfer Learning Model with Machine Learning Methods for Face Mask Detection in the Era of the COVID-19 Pandemic. *Measurement* **2021**, *167*, 108288. [CrossRef] [PubMed]
23. He, Y.; Wang, Z.; Guo, S.; Yao, S.; Hu, X. Face mask detection algorithm based on HSV+ HOG features and SVM. *J. Meas. Sci. Instrum.* **2022**, *13*, 267–275.
24. Jiang, H.; Learned-Miller, E. Face Detection with the Faster R-CNN. In Proceedings of the 2017 12th IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017; pp. 650–657.
25. Guo, S.; Li, L.; Guo, T.; Cao, Y.; Li, Y. Research on Mask-Wearing Detection Algorithm Based on Improved YOLOv5. *Sensors* **2022**, *22*, 4933. [CrossRef] [PubMed]
26. Wang, Z.; Sun, W.; Zhu, Q.; Shi, P. Face Mask-Wearing Detection Model Based on Loss Function and Attention Mechanism. *Comput. Intel. Neurosc.* **2022**, *2022*, 2452291. [CrossRef]
27. He, K.; Zhang, X.; Ren, S.; Sun, J. Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2015**, *37*, 1904–1916. [CrossRef]
28. Lin, T.Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature Pyramid Networks for Object Detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
29. Liu, S.; Qi, L.; Qin, H.; Shi, J.; Jia, J. Path Aggregation Network for Instance Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
30. Niu, Z.; Zhong, G.; Yu, H. A Review on the Attention Mechanism of Deep Learning. *Neurocomputing* **2021**, *452*, 48–62. [CrossRef]
31. Mnih, V.; Heess, N.; Graves, A. Recurrent Models of Visual Attention. *Adv. Neural Inf. Process. Syst.* **2014**, *27*, 1–9.
32. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Polosukhin, I. Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* **2017**, *30*, 1–15.
33. Hu, J.; Shen, L.; Sun, G. Squeeze-and-Excitation Networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 7132–7141.
34. Liu, T.; Luo, R.; Xu, L.; Feng, D.; Cao, L.; Liu, S.; Guo, J. Spatial Channel Attention for Deep Convolutional Neural Networks. *Mathematics* **2022**, *10*, 1750. [CrossRef]
35. Woo, S.; Park, J.; Lee, J.Y.; Kweon, I.S. CBAM: Convolutional Block Attention Module. In Proceedings of the European Conference on Computer Vision, Munich, Germany, 8–14 September 2018; pp. 3–19.
36. Zeng, G.; Yu, W.; Wang, R.; Lin, A. Research on Mosaic Image Data Enhancement for Overlapping Ship Targets. *arXiv* **2021**, arXiv:2105.05090.

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.