

A study of the lightweight algorithm for recognizing masks incorporating attention mechanism and G-GhostNet

Cong Deng^{1, †}, Wenkai Li^{2, †}, Xirui Wang^{3, †}, Bingkun Xin^{4, 5, †}

¹ School of Materials, Beijing Institute of Technology, Beijing, China.

² School of Software, Taiyuan University of Technology, Taiyuan, China.

³ School of Electrical Engineering, University of Liverpool, Liverpool, UK

⁴ School of Electronic Information and Automation, Tianjin University of Science and Technology, Tianjin, China.

[†]These authors contributed equally.

⁵ 20028202@mail.tust.edu.cn

Abstract. As COVID-19 and humanity's protracted battle makes wearing masks a norm, intelligent recognition of mask-wearing and wearing criteria is investigated to lessen the labor of epidemic control and detection personnel. To achieve the purpose of not having to manually identify mask use in a busy, open setting, to be able to provide a real-time warning of the whole mask non-wearing phenomenon in mobile devices, and to address the issue of sluggish efficiency and poor precision of traditional target detection and tracking means in complex scenes, this work presents a lightweight mask identification method that combines G-GhostNet with an attention mechanism. Firstly, YoLov5's Backbone network is replaced with a more lightweight G-GhostNet with fewer convolution kernels to reduce the parameters and computation more significantly; secondly, an attention mechanism is introduced to G-GhostNet's feature mixing process to increase discovered features' weight values. Then, the dichotomous unmasked and masked datasets are prepared to enhance the model accuracy and speed by the loss function. The experiments compare the effects of Squeeze-and-Excitation Networks (SE), Convolutional Block Attention Module (CBAM), Efficient Channel Attention (ECA), and Coordinate Attention (CA), the four attention processes on the enhanced model with the model without the added attention mechanisms. The results show that adding the attention mechanisms all make the model recognize masks better, among which adding CA has the best recognition effect with a 2% improvement in accuracy but 1.3 times increase in Latency time; in contrast, adding SE, ECA balanced speed and accuracy.

Keywords: Lightweight algorithm, Mask identification, Modified G-GhostNet, Attention mechanism, Speed and accuracy

1. Introduction

Due to the impact of the new crown epidemic, wearing masks has become the norm, and there is an urgent need for an intelligent recognition method for mask-wearing, which can reduce the operational difficulty in image recognition technology by using the powerful feature extraction capability of deep learning and convolutional neural networks, to apply computer vision technology for face mask

recognition.

Deploying deep learning models in mobile and performing real-time data processing is a major trend in the development of mobile electronic devices today. Traditional deep models have the problems of high computation and high energy consumption, which are not suitable for electronic devices with low computing power and little energy storage on the mobile side. AlexNet has 60 million parameters, which will occupy 228M memory in float32 devices, and the FLOPS of its convolutional layer occupies 663M memory. The number of VGG parameters is even three times that of AlexNet. This type of problem is commonly solved by compressing the model size for optimization, but the accuracy will be reduced [1]. MobileNet V3 obtains better performance by reducing the floating-point number on top of MobileNet V2 [2]. GhostNet, as a new deep model for mobile, improves MobileNet V3. In the ImageNet classification task, GhostNet achieves a 75.7% Top-1 correct rate with similar computation, which is higher than 75.2% of MobileNet V3 [3]. The current new version of GhostNet proposes cross-layer inexpensive operations that can be used in different network structures, further optimizing the memory required to run the model and improving the running speed on devices such as GPUs [4-5].

Mask wearing recognition needs to identify the area where the mask is worn, which generates a lot of redundant information, and the difficulties are high hardware configuration requirements, high cost, and long running time. However, in the process of feature extraction by mainstream neural networks, rich or even redundant information typically ensures a thorough grasp of the supplied material. Excellent deep neural networks may include redundant feature maps that are crucial components. We cannot avoid these redundancies, but we can cost-effectively process this information.

In 2021, Yinggu Jin, Tao Zhang, Yaning Yang, et al. proposed the MobileNet V2 model based on MobileNet to identify mask-wearing detection, which enabled some accuracy and performance development of recognition techniques [6]. In 2022, Zixun Ye, Hongying Zhang, and Yujun He proposed GhostNet to fuse the spatial attention mechanism FocusNet, using the Yolov3 framework thus improving the performance of the technique. In this paper, we refer to their method and then use the Yolov5 framework (Yolov5 has smaller parameters and higher accuracy than Yolov3) and use the improved GhostNet i.e. G-GhostNet, which fuses spatial attention mechanism and temporal attention mechanism to compare the effect [5,7-10].

In this paper, as the Backbone feature extraction network, a modified version of GhostNet is employed to obtain more feature parameters by applying less computation. First, the initial convolution layer is replaced by a series of linear transformations to obtain similar information feature outputs as conventional methods with less computational effort. Secondly, this paper fuses four attention mechanisms SE, CBAM, ECA, and CA features respectively to enhance the extraction network, focusing on the face-wearing mask part, and classifies the mask detection as a binary classification problem, where 0 means being masked and 1 means not being disguised. Finally, algorithm performance is verified by comparing the extracted features with the real image without a mask.

2. Methodology

In this paper, a modified lightweight mask recognition algorithm is used to identify the presence or absence of a mask, and the goal is achieved through a two-part design of hardware and software, as shown in figure 1. The model is deployed on Jetson TX2 through camera sampling, and pre-processing of the collected images is done to complete face detection and localization, face feature extraction, and face feature matching, to frame the mask detection region, identify whether or not a disguise is worn, and then export any resulting images to output image results.

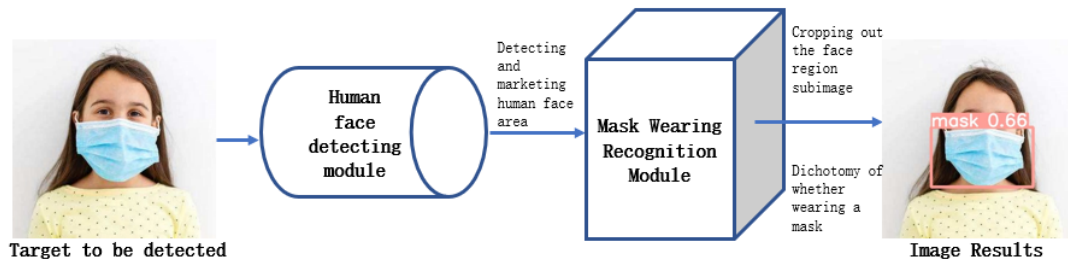


Fig. 1. Flow chart of mask-wearing recognition.

In this paper, firstly, we use G-GhostNet, a lightweight algorithm for target recognition, to lighten YoloV5, based on which four attention mechanisms SE, CBAM, ECA, CA are constructed. Adding them to G-GhostNet achieves the purpose of reinforcement of mouth features and muzzle features.

In this paper, there are three main improvements to YoloV5: (1) Initial candidate frame improvement, by optimizing the initial candidate frame algorithm, adapting the image scaling input, normalizing the small, medium and large targets, and speeding up the generation of 9 candidate frame anchor points with obvious differences, so that the distance between the target frame and the real frame is returned to the minimum. (2) Switching the Backbone network to G-GhostNet network can solve the problem that the original network has more parameters and is computationally intensive. Using lightweight G-GhostNet convolution can preserve redundancy at a lower cost. (3) Attention mechanism is added to all G-GhostNet convolutions to enhance the expression of features and provide rich information supplements. This model uses Avg IOU as a loss function to evaluate the goodness of candidate box localization parameters, and the optimal target frame is obtained by guiding the regression task with loss in classification (cls loss), loss in localisation (box loss), and loss in object confidence (obj loss).

2.1. Introduction of YoloV5

Yolo series is a deep learning-based regression method, which is a component of a network with a single stage for target detection at fast detection speed and can easily run in real time. YoloV1 is a lightweight design framework based on the network structure of GoogleNet, but replacing the Inception module with a 1×1 reduction layer and a 3×3 convolutional layer [11]. However, the localization accuracy of YoloV1 is insufficient, Joseph Redmon, Ali Farhadi et al. combined the advantages of networks such as VGG16, introduced the Anchor mechanism, and combined image fine-grained features to connect shallow and deep features, which facilitates the detection of small-sized targets [12]. With the introduction of ResNet, YoloV3 can alleviate the gradient explosion and gradient dispersion problems caused by network deepening with the help of residual network ideas and extend the previous network to Darknet-53 [13]. To reduce the gap between YoloV3 and Faster R-CNN, YoloV4 was developed. However, YoloV5 has a smaller depth, smaller feature map width, faster recognition, and lighter framework than YoloV4 [14,15]. Therefore, this paper is based on the YoloV5 algorithm for improvement.

2.2. Attention Mechanism

The attention mechanism is used in neural nets to obtain the weight values corresponding to the features using some network layers to assign higher weights to the feature maps.

2.2.1. The attention mechanism SE emerged to solve the following problem: during convolutional pooling, losses are introduced due to the varying importance of the feature map's channels. As illustrated in Figure 2, a residual block is structured with the SE module added to it. The input feature map is first pooled globally, then, to add more nonlinear processing and to account for the channels' complex correlations, it is processed by a descending fully connected layer, by a layer connected in an upward motion, then, sigmoid layer, finally, by a feature map full multiplication to get the channel weights for each one [7].

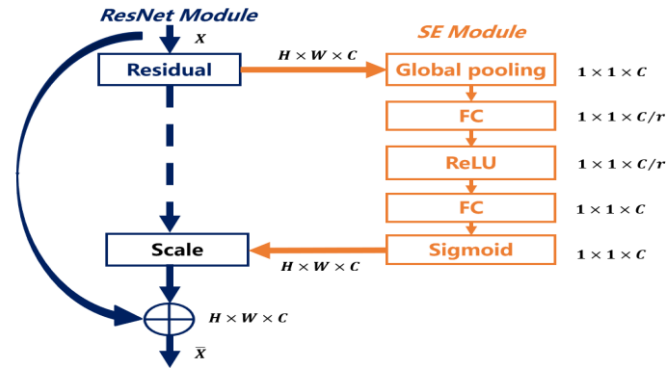


Fig. 2. Structure of residual block with SE module added to it.

2.2.2. *Attention mechanism CBAM incorporates two attention mechanisms*, as in Figure 3, with one more spatial (spatial) attention than SE. The channel attention of CBAM has one more global max pooling than SE [8].

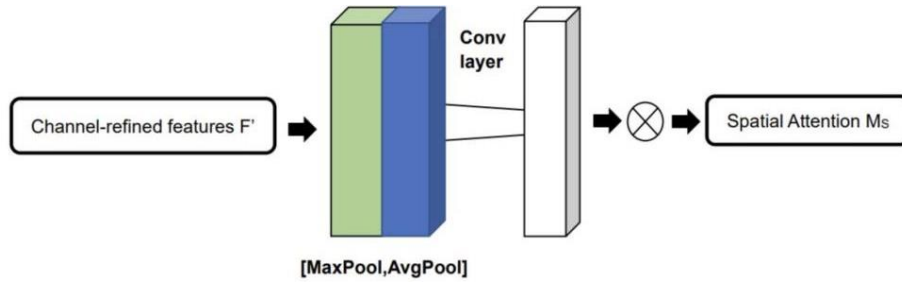


Fig. 3. Channel attention and spatial attention module of CBAM.

2.2.3. *Attention mechanism ECA efficiently implements local cross-channel interaction* with 1-dimensional convolution to extract inter-channel dependencies to avoid SE compression and dimensionality reduction and reduce the adverse effects due to dependency on learning channels[9]. As in Figure 4, the flowchart of ECA attention mechanism.

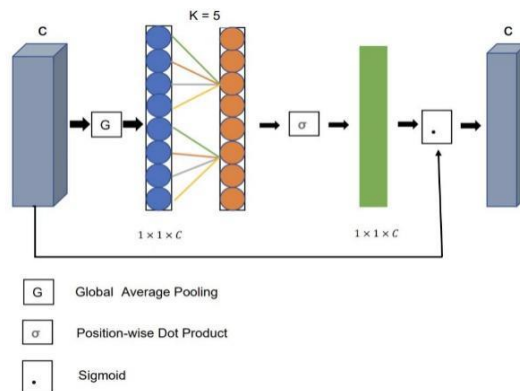


Fig. 4. ECA attention mechanism flowchart.

2.2.4. The attention mechanism CA decomposes channel attention into 2 encoding processes of aggregated features along different directions, as in Figure 5, capturing long-distance dependencies in one path and retaining precise location data in another, respectively, and pooling[10].

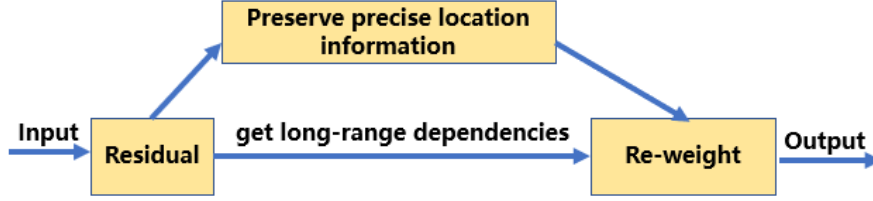


Fig. 5. Structure of the residual block with the CA module added.

2.3. Method Improvement

The model is optimized by improving the algorithm of the initial candidate frame, replacing the convolutional layers of the original model, and adding different attention mechanisms to the G-GhostNet convolution to significantly reduce the model computational parameters and model size with improved accuracy, allowing the model to be deployed lightly into mobile devices. Figure 6 shows the network architecture of this model, which is based on the architecture of YoLov5. G-GhostConv replaces Focus and C3Net in the original network, and G-GhostC3 replaces the C3 module, additionally decreases the Backbone network's layer count from 10 layers of Backbone convolution to 6 layers.

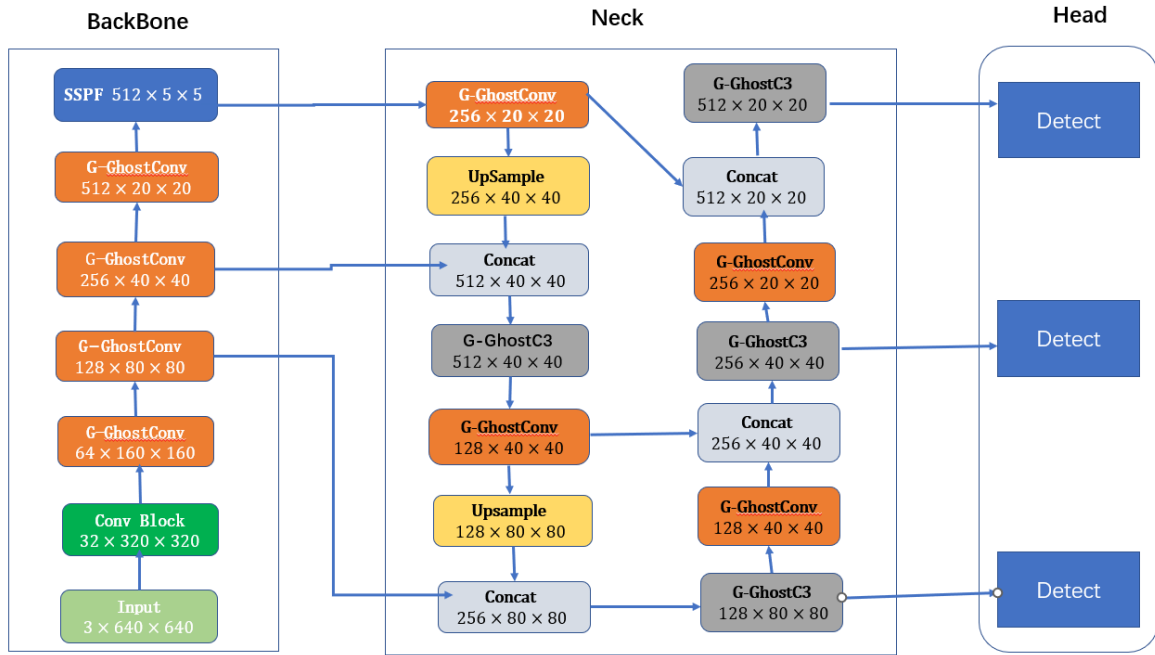


Fig. 6. Network architecture diagram of this model.

2.3.1. Initial candidate frame improvement. Since the target instances of the same category have similar aspect ratios, several bounding boxes with high probability can be prepared in advance in the dataset and then used as a benchmark for prediction. The training of the model will be easier and the predicted bounding box will be more accurate the closer the initial candidate box parameters are to the actual bounding box. The design of the initial candidate frame parameters has a direct impact on the frequency of detecting the target and the accuracy of a frame's intended location. The candidate region box is calculated from our input bounding box, but the training input size is not unique. By normalizing ImageNet so that the input image size is resized to the same 640×640 , the bounding box size also changes. To make the model both efficient and fast in detecting cross-scale targets, and to make the model anchors conform to the normalized size, the model uses 80×80 candidate frames to detect small targets, 40×40 candidate frames to detect medium-sized targets, and 20×20 candidate frames to

detect large targets, and the feature map is divided into three channels of different sizes, with a total of nine candidate frames of different sizes, to speed up positive and negative sample learning, while using the improved KMeans algorithm to make the anchor regress to a suitable candidate region frame.

Although the traditional KMeans will cluster the boxes in the dataset, many datasets are not very different from each other due to the similar size of the boxes and the 9 boxes clustered out, which is not conducive to model training instead. To improve this problem, this paper uses an improved KMeans algorithm so that the randomly selected centroids converge to the global optimal solution, as shown in the figure 7, the initial clustering center is set randomly between 0~640, and the distance between all samples is to make the distance between the selected cluster centers as large as possible, the probability of all non-cluster center sample points being selected as the next cluster center is proportional to the size of the previously recorded distance until nine cluster centers are selected.

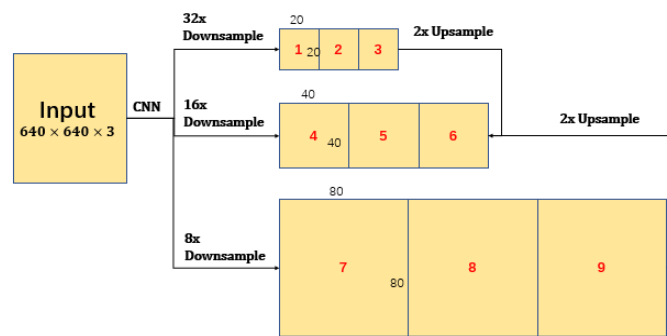


Fig. 7. 9 candidate box sizes of varying sizes.

2.3.2. Convolution layer improvement. The overabundance of network parameters is a significant contributor to training challenges for neural networks. To make the model more lightweight and to enhance the network's ability to understand features using redundant ground feature maps, this paper uses a lightweight network model, G-GhostNet, to replace the original convolution of the Backbone in the original model, preserving redundancy more cost-effectively. Compared to a normal convolutional network, as in Figure 8, the final output of the specified size feature map is generally achieved by 2 convolutions and through activation, pooling, etc. The improved convolutional neural network decreases the number of network parameters while increasing network generalization by compressing the scale, and random initialization in the form of weights and weight sharing. In the training of the network model, filters can be trained to be able to detect shapes and edges. As shown in Figure 9, G-GhostNet divides the deep features into Ghost features and complex features, and its integration of the original convolution has two main steps: (1) Ghost features are obtained by a simple linear transformation of the first convolution block; (2) complex features are generated by n convolution blocks, splicing the intermediate features from layer 2 to layer n in the complex feature branch, and then using the mix transform function to obtain the same features as the Ghost features in the same domain. Finally, the two features are simply summed and fused.

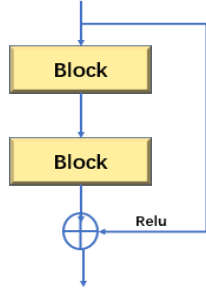


Fig. 8. Original convolutional network.

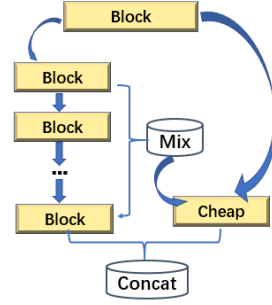


Fig. 9. G-GhostNet convolutional network.

2.3.3. Adding an attention mechanism to convolutional layers. As computational power remains a bottleneck limiting the development of neural networks and models become more complex when there is too much input information, neural networks can focus attention on a portion of the input and by introducing attention, the amount of information processed can be reduced, thus further reducing computational resources. In this paper, a plug-and-play attention mechanism component is introduced into the G-GhostNet convolutional module, using putting on a mask and not putting on a mask as a binary classification task. The experimental model diagram is shown in Figure 10, with the attention mechanism added before the mix transform function. The principle is that the experiment generates features at each Ghost stage by the cheap operation to explore the redundancy between the first and the last module, obtains the complicated features by massive block processing, and adds the attention mechanism to the complicated features plug-ins to enhance the expressiveness of the CHEAP operation and provide a rich information complement to the cheap operation. Then the information is aggregated, Z is globally mean pooled, a fully connected layer transformation is used to Y_n^g homogeneous domain, and information is mixed.

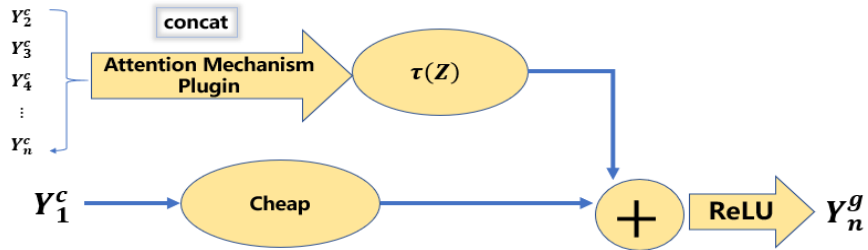


Fig. 10. Improved G-Ghostnet add attention mechanism information mixing flowchart.

Suppose that a stage containing n blocks is defined and its output is denoted $X \in R^n$, the complex features are denoted $Y_n^c \in R^{(1-\lambda)n}$, the ghost features are denoted $Y_n^g \in R^{\lambda n}$ ($0 \leq \lambda \leq 1$), the complex features are enhanced by the attention mechanism, the attention mechanism module has three tensors Q, K, V ; Q is the query vector, and K is the location to focus on, which is the encoder data, and V represents the input of the attention mechanism model. Y_n^c, Y_n^g are generated in the following way, C denotes the cheap operation, which can be a 1×1 or 3×3 convolution. The current stage output features are obtained by combining the above two features.

$$Y_n^c = \text{concat}[L'_n(L'_{n-1}(\dots L'_2(Y_1)))] \text{, } \text{Attention}(Q, K, V)] \quad (1)$$

$$Y_n^g = C(Y_1) \quad (2)$$

$$Y_n = \text{RELU}(\text{conv}_{1 \times 1}(\text{concat}[Y_n^c, Y_n^g])) \quad (3)$$

2.4. Loss Function Selection

In this paper, the initial candidate box's loss function is Avg IOU. The target detection task's loss function is usually divided into two sections: the average intersection ratio (Avg IOU), which is the average distance between the predicted box and the real box, is the most often used measure for assessing the regression loss of the bounding box. The closer the value is to 1, the closer the predicted information is to the desired information. The calculation formula is as follows.

$$Avg\ IOU = \frac{\sum_1^n \frac{|A \cap B|}{|A \cup B|}}{n} \quad (4)$$

In model training, as the number of sample training iterations increases, the appropriate loss function can have a greater impact on the convergence of the model. The loss of this model is mainly composed of the classification loss L_{cls} (determining whether or not the anchor frame and calibration categorization are right), the confidence loss L_{obj} of the target boundary (calculating the confidence of the network), and the locus loss L_{loc} (the error between the prediction frame and the calibration frame).

$$Loss = L_{cls} + L_{obj} + L_{loc} \quad (5)$$

3. Experiment Outcomes and Analysis

3.1. Set of data and experimental details

Algorithms in this paper were implemented using the Pytorch framework. The software training environment was Python-3.7.6, torch-1.12.1, with CUDA acceleration version 11.3.1, NVIDIA GeForce RTX 3060 GPU, 6GB GPU core memory and ubuntu 16.04 operating system.

This experiment made use of a small-scale picture dataset with 2800 unmasked faces and 2300 masked faces, with annotated information including target type and location information. To guarantee that the training data is as broad as feasible and the test set is universal, the training and test sets are divided at 9:1 to increase model performance and prevent information leakage. Data used for training images are shown in Figure 11, with the unmasked marker 0 and the masked marker 1. Figure 12 depicts the position distribution of the targets to be identified as well as the aspect distribution of the sample pictures in the dataset, which demonstrates that the majority of the objects to be detected are situated in the center of the images, but there are some images where the target objects are placed in other places, which is helpful to the model's generalisation capacity, and the aspect ratios of the training images vary, necessitating image normalisation. In practice, the mask-wearing identification issue involves more tiny targets, and increasing the input size reduces the cropping of huge photos owing to irregular image sizes. The data input size is reduced to 640×640 , which can increase the identification accuracy of tiny targets like masks to some extent.



Fig. 11. Some examples of training data pictures.

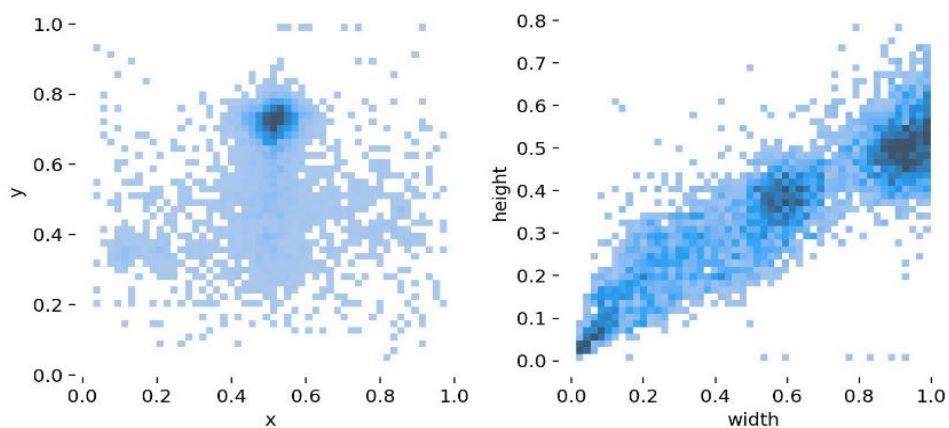


Fig. 12. The distribution of the target's position (left) and the length and breadth distribution of the dataset's sample pictures (right).

The hardware devices for this experiment are the camera and Jetson TX2. as shown in Figure 13, the trained model will detect the target data in real-time through the camera, deploy the target detection model on the Jetson TX2 as shown in Figure 14, and the GPU development platform Xilin FPGA at the embedded end connects to the camera via USB and the network cable connects to the LAN to complete the model inference and output the detection results.

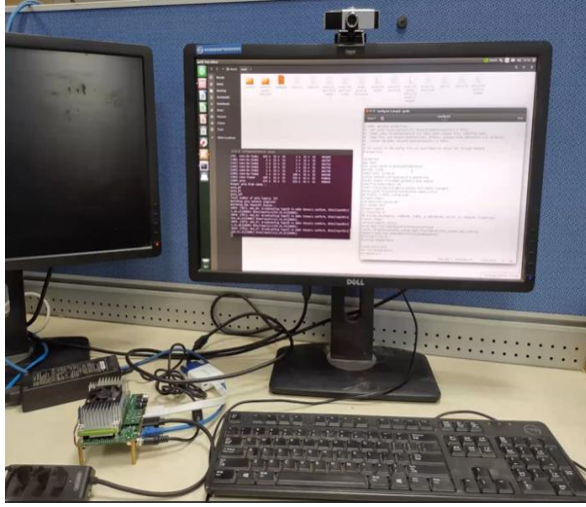


Figure 13. Camera real-time target detection platform.

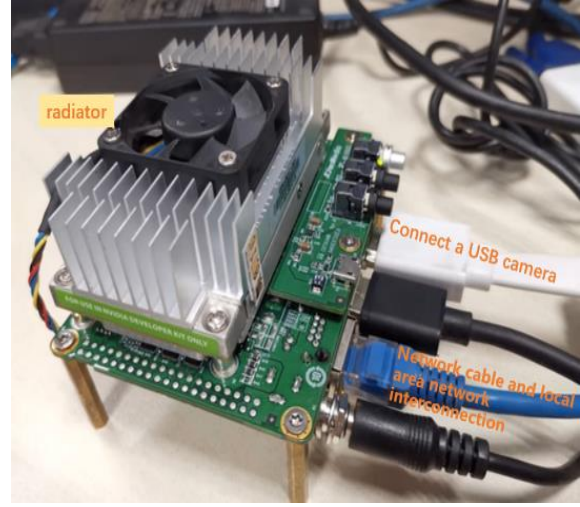


Figure 14. Xilinx FPGA for Jetson TX2.

3.2. Model Evaluation

Model evaluation method, the evaluation index of model optimization adopts the double judgment of Accuracy and Precision.

3.2.1. Accuracy. Represents the ratio of positive samples to all samples judged by the model during detection. Because the accuracy rate does not provide enough information to properly evaluate the model's performance, it is generally used to evaluate the correctness of the entire model.

$$Accuracy = \frac{TP+TN}{TP+FP} \quad (6)$$

3.2.2. Precision. The accuracy rate is the proportion of real positive samples that are anticipated to be positive, and it reflects whether the prediction results are accurate or not. The formula for calculating the accuracy rate is as follows:

$$Precision = \frac{TP}{TP+FP} \quad (7)$$

Among them, P (Positive) denotes a positive example of the anticipated value. N (Negative) means a negative example of the anticipated value. T(True) means the true value is the same as the predicted value. The anticipated value written as F is the inverse of the real value (False). TP denotes that either the real value or the predicted value is a positive sample, and that the real value and the predicted value are the same. TN indicates that either the actual value or the anticipated value is a negative sample, and that the actual and predicted values are the same. FP denotes a difference between the actual and predicted values, where the actual value is either a negative sample or a positive sample. FN indicates that the anticipated value is either a negative sample or the true value is different from the expected value.

3.2.3. MAP@0.5 (standard Mean Precision with IOU=0.5). When the IOU is set to 0.5, the average accuracy of all images in each category is determined, and then all categories are averaged, resulting in mAP@0.5. The larger the value, the higher the model detection accuracy.

3.2.4. MAP@.5:.95 (I O U from 0.5 to 0.95, mean Average Precision with step size 0.05). The average mAP over multiple IOU criteria is shown as mAP@.5:.95 (from 0.5 to 0.95 in 0.05 steps) (0.5, 0.55, 0.6, 0.65, 0.7, 0.75, 0.8, 0.85, 0.9, 0.95). This value is affected by the position of the detection

frame. Generally, the larger the IOU threshold, the higher the requirements for the position of the detection frame. The better the model detection effect, the larger the value.

$$mAP = \frac{\sum_{i=1}^K AP_i}{K} \quad (8)$$

3.3. Experimental Results

The ratio of the projected rectangular box's intersection and union to the actual target before and after the improvement of the initial candidate frame is Avg IOU, as shown in the Table 1, the Avg IOU of the improved KMeans is closer to 1, indicating that the improved KMeans can obtain better candidate frame parameters.

Table 1. The results of improved Avg IOU and the original comparison.

Calculated methods	Avg IOU(%)
Traditional KMeans Method	0.76
Advanced KMeans Method	0.78

The model is assessed using the aforementioned model assessment metrics after 200 training epochs. Table 2 displays the assessment findings. It can be shown that adding the attention mechanism improves the model's identification of masks greatly, about 1-2 %. The recognition accuracy of wearing masks and not wearing masks has been greatly improved. 1% increased to a maximum of 9.6 % (CA), and the recognition accuracy rate of wearing a mask increased from 9.4 % to nearly 9.86% (ECA, SE).

The attention technique allows the G-GhostNet module to provide larger weights to feature channels during deep feature development, allowing the model to more correctly pinpoint the detection target. Other attention mechanisms may turn the feature tensor into a single feature vector channel by two-dimensional global pooling, while CA decomposes the channel attention into two one-dimensional feature codes together with two spatial orientation feature aggregation, which may explain why CA performs best. Long-distance associations may thus be collected along one spatial direction while exact position information is kept along the other, and the produced feature maps can be encoded as orientation-aware and location-sensitive maps, complementing and improving the objects of interest. The reason why CBAM improves the effect is inferior to several other attention mechanisms may be that it uses a large convolution kernel, which leads to the expansion of the target range and the inability to distinguish the target object from the background area well.

Table 2. Test set evaluation results

	Added attention mechanism	Recognition accuracy %	Recognition precision %
Respirator	none	94	94
No Respirator	none	94	92.1
Respirator	SE	94	98.5
No Respirator	SE	94	94.7
Respirator	CBAM	94	98
No Respirator	CBAM	93	93.6
Respirator	ECA	95	98.6
No Respirator	ECA	92	94.8
Respirator	CA	95	98.4
No Respirator	CA	93	96

From Table 3, the results of Latency's test on NVIDIA GeForce RTX 3060 Laptop GPU, 6GB, and 8 batch sizes, compared with the original Yolov5 model, the model mAP@.5 replaced by the G-GhostNet BackBone increased from 91.5% to 93.8%, mAP@.5:.95 increased from 50.7% to 52.4%; the number

of parameters has also dropped significantly, by almost two-thirds, to 6.4M, the number of floating-point operations per second is almost unchanged, and the model memory has dropped in general. At 8.4M, the latency drops by 1.3ms, which may be related to the reduction in the number of convolutions. After adding the attention mechanism, compared with not adding the model, the parameter quantity and model size are unchanged, and the accuracy rate is mostly improved to varying degrees. The mAP@.5:.95 is improved by 0.6%-3.8%, but the time consumption changes. Longer, the delay increases by 1.3ms-9.5ms. Among them, the CA attention mechanism has the highest accuracy, mAP@.5 is 95.6%, mAP@.5:.95 is 56.2%; SE attention mechanism is added After that, the time consumption is the smallest at 8.7ms, and the accuracy ranks third. After adding ECA, the time delay increases by 3.8ms, which is much lower than the delay of 17ms after adding CBAM, and the accuracy after adding ECA ranks second among the four attention mechanisms. It can be seen that SE has the least impact on model delay, the CA attention mechanism has the best recognition effect, and SE and ECA better balance speed and accuracy.

Table 3. G-GhostNet adds attention mechanism performance evaluation.

BackBone Model		Param. (M)	FLOPs(GB)	TrainingLatency(ms)	mAP@.5 (%)	mAP@.5 :.95 (%)	
Yolov5	C3	19.9	16.5	15.4	8.7	91.5	50.7
G-GhostNet	G-GhostNet	6.4	16.4	8.4	7.4	93.8	52.4
SE	G-GhostNet	6.4	16.5	8.4	8.7	92.1	53
CBAM	G-GhostNet	6.4	16.6	8.4	17	95	55.6
ECA	G-GhostNet	6.4	16.5	8.4	11.2	95	55.9
CA	G-GhostNet	6.4	16.7	8.4	16.9	95.6	56.2

From Table 4, in the real scene, comparing the results of detecting masks, after adding the attention mechanism, the recognition of masks and without masks is significantly improved. Among them, the effect of adding CA is the best, the recognition probability of scenes with masks is between 0.40-0.70, and the recognition probability of wearing a mask is between 0.62-0.81. Add SE, ECA effect is second. In the actual inspection and detection of the pure G-GhostNet skeleton model, there was a situation without a mask that was not recognized. The light in this situation was brighter, but after adding the attention mechanism, it could be recognized, indicating that the attention mechanism improved the effective redundancy. The use of redundant features makes the target detection focus on the mouth.

Table 4. Comparison of experimental results (continue).

Mask	No Mask
The result set of pure G-GhostNet skeleton actual.	The result set of pure G-GhostNet skeleton actual.



The result set of adding SE's G-GhostNet model. The result set of adding SE's G-GhostNet model.

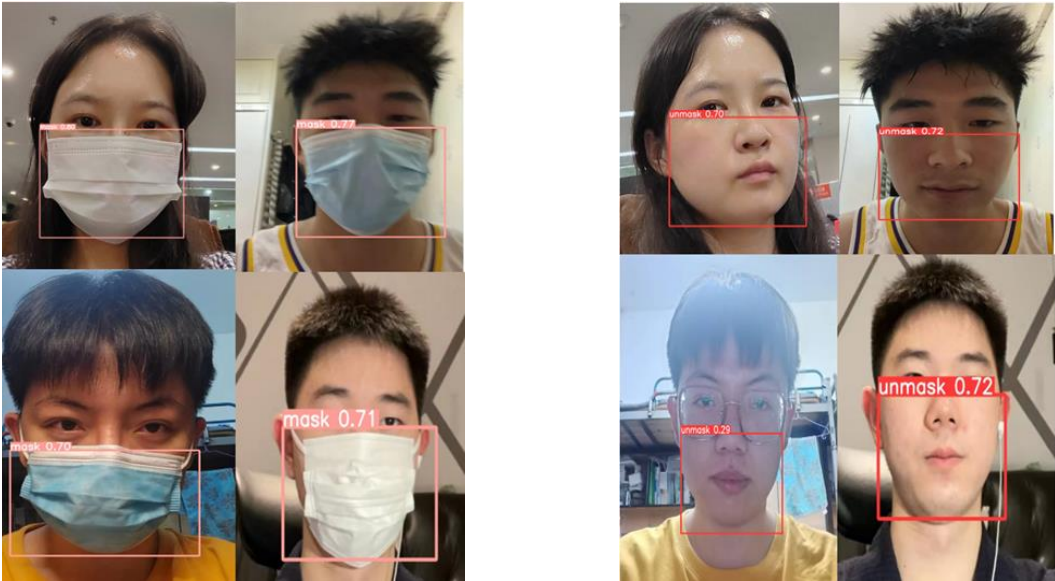


Table 4. (continued).


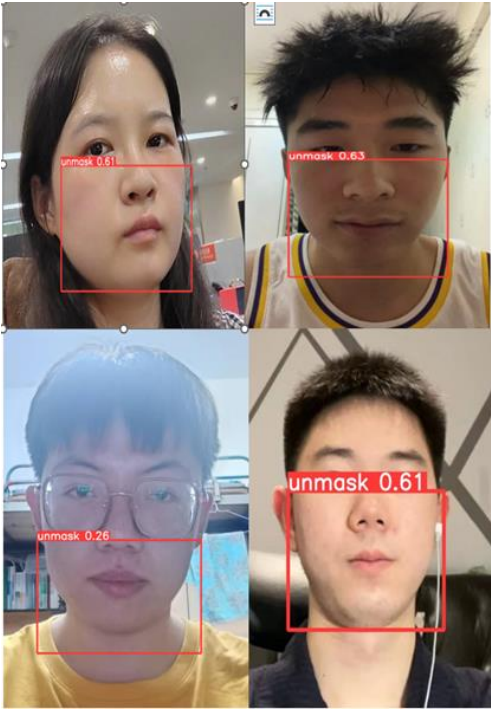

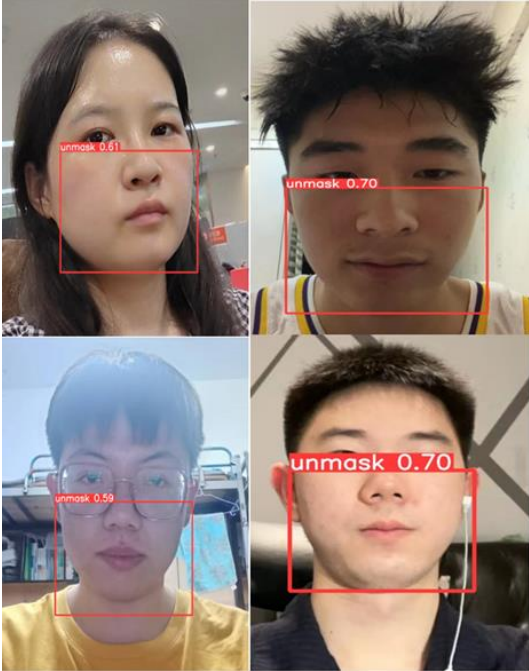
The result set of adding CBAM's G-GhostNet model.		The result set of adding CBAM's G-GhostNet model.	
			
The result set of adding ECA's G-GhostNet model.		The result set of adding ECA's G-GhostNet model.	
			

Table 4. (continued).



The result set of adding CA's G-GhostNet model.				The result set of adding CA's G-GhostNet model.			
							

Figure 15 depicts the evolution of the model's loss function value throughout five situations with 200 training iterations, indicating that the loss function value is greater when just the G - GhostNet module is replaced, and the error between the predicted and real frames is greater; however, when the attention mechanism is included, the loss function value falls dramatically. However, the loss function changes the greatest after including the SE attention mechanism, which might be because the SE attention mechanism only examines the channel information unilaterally and ignores the location information, resulting in a high fluctuation of the IOU's positioning frame.

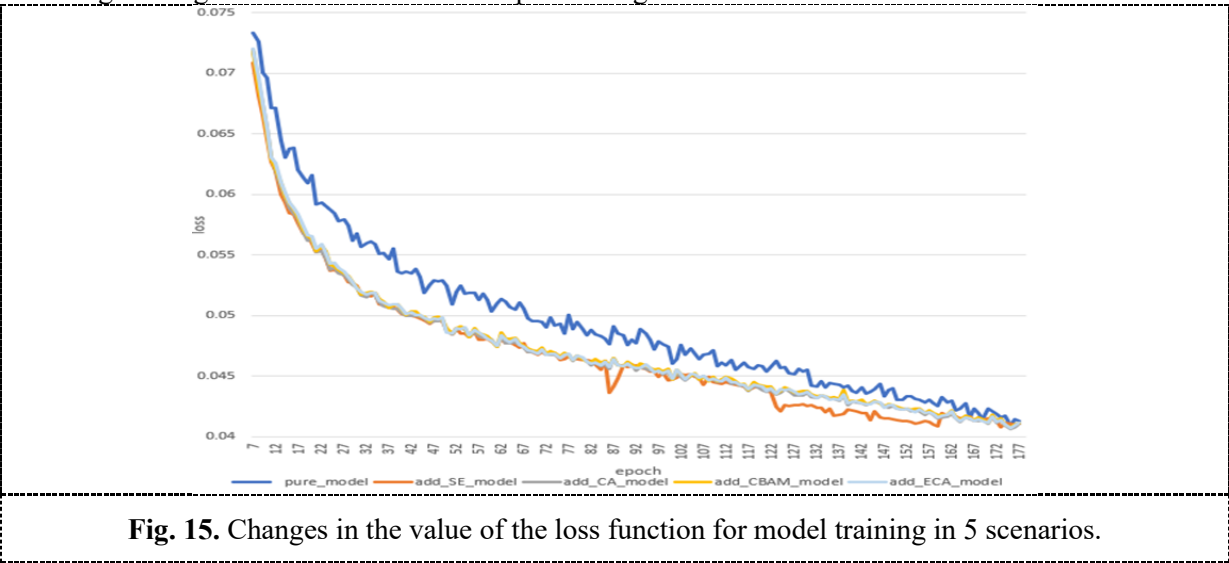


Figure 16 shows the changes in the accuracy of model training in 5 scenarios with 200 epoch training, indicating that the accuracy of the model is not as high as adding the attention mechanism when only the G-GhostNet module is replaced, which further confirms our previous view.

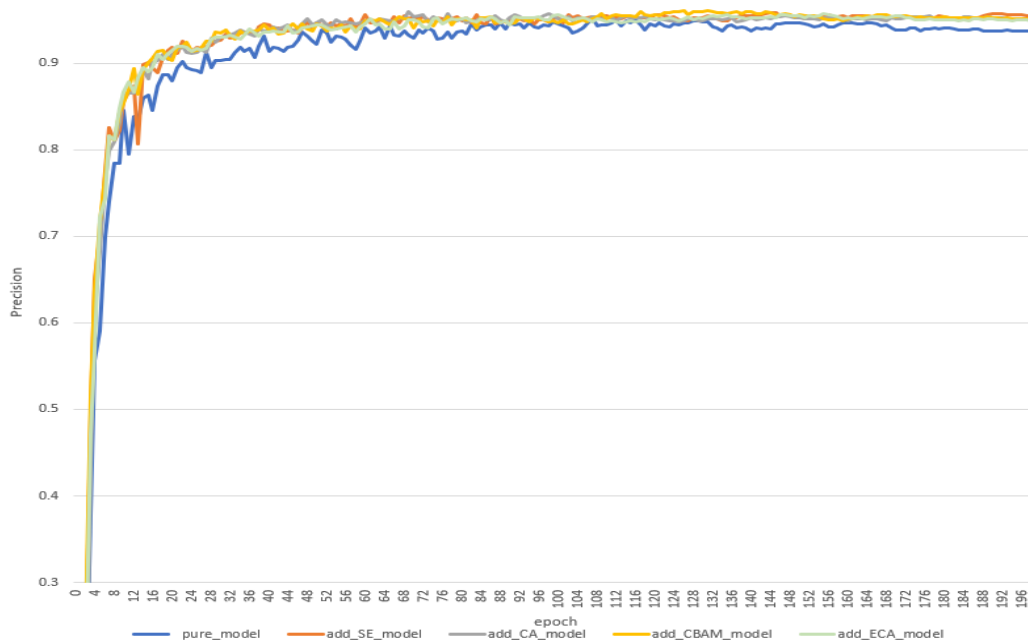


Fig. 16. Changes in the accuracy of model training under 5 scenarios.

4. Conclusion

This paper shows the problem of automatically identifying whether to wear a mask on the mobile terminal when it is required to wear a mask. Based on the improved deep learning algorithm of G - GhostNet, different attention mechanisms SE, CBAM, ECA, CA are added to discuss the effect of model recognition. It has been discovered that including an attention mechanism can increase the model's ability to detect masks, but it will increase the delay. Adding CA to the model offers the best recognition effect, whereas SE has the least influence on model delay. The experimental findings indicate that the Ghost Net algorithm optimized for mask recognition can be implemented using G-GhostNet adding an attention mechanism, but requires a trade-off between accuracy and time consumption as needed. CA attention mechanism works best for recognition, SE, ECA balance speed and accuracy.

References

- [1] Ren Lei , Liu Guoqing , Wang Lihua , et al . A new and efficient mobile deep learning image classification system. 2021 *Int. Th. Tec.* , **(5)**:58-63.
- [2] Wang Lihui, Yang Xianzhao, Liu Huikang, et al . Pedestrian detection and tracking algorithm based on GhostNet and attention mechanism. 2022, *Dat. Col. Proc.*, **37(01)**: 108-121.
- [3] Han K ,Wang Y ,Tian Q, et al. Ghostnet: More features from cheap operations. 2020, *Conf. Com. Vis. Pat. Recog.*.1580-1589.
- [4] Kai Han ,Yunhe Wang, Qi Tian ,et al .GhostNets on Heterogeneous Devices via Cheap Operations. 2022 *Int. J. Com. Vis.* **130**:1050–1069.
- [5] Ye Xun , Zhang Hongying , He Yujun .A lightweight mask-wearing face recognition algorithm fused with attention mechanism.2022, *Com. Eng. App.*:1-10.
- [6] Jin Yinggu , Zhang Tao, Yang Yaning, et al. Research on mask wearing recognition based on MobileNet V2 .2021, *J. Dalian Univ. Nat.*, **23(05)**: 404-409+431.
- [7] X. Hao, C. Shan, Y. Xu, et al. An Attention-based Neural Network Approach for Single Channel Speech Enhancement.2019, *Int. Conf. Aco., Spe. Sig. Proc.* 6895-6899.

- [8] Woo S, Park J, Lee J Y, et al. Cbam: Convolutional block attention module.2018, *Euro. Conf. Com. Vis.* 3-19.
- [9] Liu T, Luo R, Xu L, et al. Spatial Channel Attention for Deep Convolutional Neural Networks., 2022, *Math.* **10(10)**: 1750.
- [10] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. 2021 *Conf. Com. Vis. Pat. Recog.* 13713-13722.
- [11] Han X, Chang J, Wang K. You only look once: Unified, real-time object detection. *Procedia*, 2021, *Comp. Sci.*, **183**: 61-72.
- [12] Redmon J, Farhadi A. YOLO9000: better, faster, stronger. 2017 *Conf. Com. Vis. Pat. Recog.* 7263-7271.
- [13] Lin T Y, Dollár P, Girshick R, et al. Feature pyramid networks for object detection. 2017, *Conf. Com. Vis. Pat. Recog.* 2117-2125.
- [14] Yuan S, Wang Y, Liang T, et al. Real - time recognition and warning of mask wearing based on improved YOLOv5 R6.1., 2022 *Int. J. Com. Vis.***45**.
- [15] Wang Z, Sun W, Zhu Q, et al. Face Mask-Wearing Detection Model Based on Loss Function and Attention Mechanism. 2022, *Com. Int. Neur.***82**.