applied sciences

MDPI

Article

Lightweight Human Ear Recognition Based on Attention Mechanism and Feature Fusion

Yanmin Lei [1], Dong Pan [1,2,*], Zhibin Feng [3] and Junru Qian [4]

[1] Department of Electrical and Information Engineering, Changchun University, Changchun 130022, China; leiym@ccu.edu.cn
[2] Institute of Science and Technology, Changchun Humanities and Sciences College, Changchun 130028, China
[3] Aviation Basic College, Air Force Aviation University, Changchun 130022, China; fzb0431@163.com
[4] Jilin Province Key Laboratory of Measuring Instrument and Technology, Jilin Institute of Metrology, Changchun 130103, China; qjr1107007169@163.com
* Correspondence: 200401079@mails.ccu.edu.cn

Abstract: With the development of deep learning technology, more and more researchers are interested in ear recognition. Human ear recognition is a biometric identification technology based on human ear feature information and it is often used for authentication and intelligent monitoring field, etc. In order to make ear recognition better applied to practical application, real time and accuracy have always been very important and challenging topics. Therefore, focusing on the problem that the mAP@0.5 value of the YOLOv5s-MG method is lower than that of the YOLOv5s method on the EarVN1.0 human ear dataset with low resolution, small target, rotation, brightness change, earrings, glasses and other occlusion, a lightweight ear recognition method is proposed based on an attention mechanism and feature fusion. This method mainly includes the following several steps: First, the CBAM attention mechanism is added to the connection between the backbone network and the neck network of the lightweight human ear recognition method YOLOv5s-MG, and the YOLOv5s-MG-CBAM human ear recognition network is constructed, which can improve the accuracy of the method. Second, the SPPF layer and cross-regional feature fusion are added to construct the YOLOv5s-MG-CBAM-F human ear recognition method, which further improves the accuracy. Three distinctive human ear datasets, namely, CCU-DE, USTB and EarVN1.0, are used to evaluate the proposed method. Through the experimental comparison of seven methods including YOLOv5s-MG-CBAM-F, YOLOv5s-MG-SE-F, YOLOv5s-MG-CA-F, YOLOv5s-MG-ECA-F, YOLOv5s, YOLOv7 and YOLOv5s-MG on the EarVN1.0 human ear dataset, it is found that the human ear recognition rate of YOLOv5s-MG-CBAM-F method is the highest. The mAP@0.5 value of the proposed YOLOv5s-MG-CBAM-F method on the EarVN1.0 ear dataset is 91.9%, which is 6.4% higher than that of the YOLOv5s-MG method and 3.7% higher than that of the YOLOv5s method. The params, GFLOPS, model size and the inference time per image of YOLOv5s-MG-CBAM-F method on the EarVN1.0 human ear dataset are 5.2 M, 8.3 G, 10.9 MB and 16.4 ms, respectively, which are higher than the same parameters of the YOLOv5s-MG method, but less than the same parameters of YOLOv5s method. The quantitative results show that the proposed method can improve the ear recognition rate while satisfying the real-time performance and it is especially suitable for applications where high ear recognition rates are required.

Keywords: YOLOv5s-MG; ear recognition; accuracy; attention; feature fusion

Citation: Lei, Y.; Pan, D.; Feng, Z.; Qian, J. Lightweight Human Ear Recognition Based on Attention Mechanism and Feature Fusion. *Appl. Sci.* **2023**, *13*, 8441. https://doi.org/10.3390/app13148441

Academic Editor: Andrea Prati

Received: 10 June 2023
Revised: 15 July 2023
Accepted: 18 July 2023
Published: 21 July 2023

1. Introduction

Due to the increasing computing power of computers, biometric identification technology has been widely developed and applied. Biometric recognition is based on human features for identification. At present, the commonly used biometric features are mainly face, ear, iris, palmprint, gait, voice and fingerprint. Because the human ear is stable, non-invasive, easy to collect, has low requirements for image acquisition equipment, and

is not easily affected by factors such as glasses, emotions and makeup, it has attracted the attention of many researchers. In 2023, Oyebiyi O.G. et al. used a search criterion of publications not later than 10 years and downloaded 1121 articles from 10 databases: Taylor & Francis, Springer, Science Direct, ACM, Emerald, Sage, Elsevier, Wiley, MIT and IEEE explore [1].

At present, ear recognition methods mainly include traditional methods and deep learning-based methods. The deep learning methods for ear recognition mainly include two-stage Faster R-CNN and single-stage SSD and YOLO series. In 2017, Zhang Y. proposed an efficient and fully automatic 2D ear detection system utilizing multiple scale faster R-CNN on the UND-J2 database [2]. In 2017, Fan, T.Y. et al. [3] proposed a face and ear detection method by using Faster R-CNN. In 2022, Aman Kamboj et al. [4] introduced a new database, NITJEW, and used the modified deep learning models Faster-RCNN and VGG-19 for ear detection and ear recognition tasks, respectively. In 2021, Kamboj A. et al. [5] proposed the CED-Net model for ear recognition and compared it with the Faster-RCNN and SSD methods. Qian J.R. [6] used YOLOv3 for dynamic human ear recognition. In 2021, Quoc H.N. [7] used YOLOv5 to locate multiple tiny ears with a short inference time. The authors of [4,8–13] also used different deep learning network models to achieve ear recognition. In addition to 2D ear recognition, the authors of [14,15] also implemented 3D ear recognition by deep learning. In 2022, Bahadir K. et al. [16] implemented gender recognition based on human ear images and deep learning.

Through the study of the above references, it is found that the method based on deep learning can realize dynamic, static, 2D and 3D ear recognition with various pose changes. However, when CNN is used to solve the problem of ear recognition, there will be a large amount of calculation, parameter quantity and model size.

The real-time performance of a single-stage network is better than that of a two-stage network. In order to further improve the real-time performance of single-stage networks, many scholars use lightweight networks to improve single-stage networks. Common lightweight networks include MobileNet series, Shufflenet series and Ghostnet. In 2023, Lei, Y. et al. [17] proposed the YOLOv5s-MG method which has better real-time performance than YOLOv5s-V3, YOLOv5s-V2 and YOLOv5s-G. However, while improving real-time performance, it may also reduce the accuracy of human ear recognition on some human ear datasets. For a practical application of the ear recognition system, the ear recognition rate is the most important evaluation index. If the human ear recognition rate is low, there will be missed recognition or false recognition, which will affect the application of the system. Because the mAP@0.5 value of the YOLOv5s-MG method is 0.855, the goal of this paper is how to improve it to more than 0.9. on the EarVN1.0 ear dataset in this paper.

The main contribution of this paper is to propose a lightweight ear recognition method based on the attention mechanism and feature fusion named YOLOv5s-MG-CBAM-F. In this method, the CBAM attention mechanism was added to the connection between the backbone network and the neck network of the lightweight human ear recognition method YOLOv5s-MG. Then, the SPPF layer and cross-regional feature fusion in the above method were added to further improve the accuracy of ear recognition. Therefore, the YOLOv5s-MG-CBAM-F method proposed in this paper can not only improve the accuracy of human ear recognition but also ensure real-time performance.

The other parts of this paper are arranged as follows. YOLOv5s-MG, CBAM attention mechanism, SPPF network and PANet network are introduced in Section 2. The improved YOLOv5s-MG-CBAM-F method is proposed in Section 3. Section 4 provides the experiments and results analysis. Section 5 provides the conclusion of the paper.

2. Related Works

2.1. YOLOv5s-MG

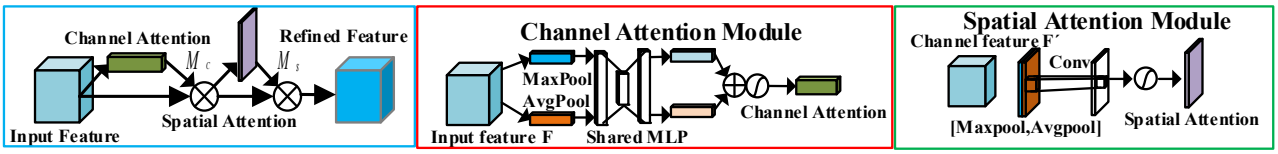
YOLOv5s-MG [17] is a lightweight YOLOv5s human ear recognition method based on MobileNetV3 and the idea of the Ghostnet. In this method, the backbone network of the YOLOv5s was replaced by the MobileNetV3 lightweight network and the C3 module

and Conv module in the YOLOv5s neck network are replaced by the C3Ghost module and GhostConv module, which realizes the lightweight of feature extraction and feature fusion of YOLOv5s simultaneously. Compared with other methods, the YOLOv5s-MG method has good real-time performance, but for EarVN1.0 human ear datasets, the human ear recognition rate has a slight decrease. Therefore, based on the high real-time performance of the YOLOv5s-MG method, this paper further improves the accuracy of human ear recognition, so that the improved human ear recognition method can have relatively good real time and accuracy at the same time.

2.2. CBAM Attention Mechanism

The attention mechanism assigns different weights to different parts of the image, so that the network pays attention to the important information of the target and ignores other secondary information. At present, the commonly used attention mechanisms mainly include SENet [18], CANet [19], ECANet [20] and CBAM [21].

SENet channel attention strengthens the expression ability of the network by establishing the relationship between the channel features of the feature map, thereby improving the sensitivity of the channel and enabling the network to improve its information utilization. ECANet is based on SENet, abandoning the fully connected layer, but using 1×1 convolution to capture the relationship between channels to avoid the impact of incomplete information caused by dimensionality reduction. CANet is a lightweight attention that obtains location information by changing the pooling method. It not only avoids the problem of information loss, but also considers direction-related location information. CBAM is a lightweight attention method combined with Channel Attention Module (CAM) and Spatial Attention Module (SAM), which uses CAM and SAM in turn for feature maps. CAM enables the network to learn channels containing key information, and SAM enables the network to learn key information on the channel feature map. The structure of CBAMNet is shown in Figure 1a.



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

Figure 1. CBAMNet. (**a**) The structure of CBAMNet; (**b**) CAM; (**c**) SAM.

Firstly, the input feature map will go through the CAM module, as shown in Figure 1b. The feature map in the channel is compressed and the weights are assigned, so that the network can play an important role in analyzing the channel where the key information is located, as shown in Equations (1) and (2).

$$F\prime = M_C(F) \otimes F \tag{1}$$

$$\begin{aligned} M_c(F) &= \sigma(MLP(AvgPool(F)) + MLP(MaxPool(F))) \\ &= \sigma\left(W_1\left(W_0\left(F^c_{avg}\right)\right) + W_1(W_0(F^c_{\max}))\right) \end{aligned} \tag{2}$$

where F is the input feature and $F \in R^{H\times W\times C}$, $M_C(F)$ is a one-dimensional channel attention feature map, σ is a Sigmoid activation function, MLP is a multi-layer perception, W_1 and W_0 are two shared weights, F^c_{avg} and $F^c_{\max}$ represent the average pooling feature map and the maximum pooling feature map, respectively.

Then, it will enter the SAM module, as shown in Figure 1c, and use the results of CAM output as CAM input, use the spatial relationship between features to generate a spatial attention map, keep the spatial dimension unchanged, compress the channel dimension and enable the network to learn the location information of key information.

The SAM module is shown in Equations (3) and (4).

$$F'' = M_S(F\prime) \otimes F\prime \tag{3}$$

$$\begin{aligned} M_s(F,\prime) &= \sigma\left(f^{7\times7}([AvgPool(F'); MaxPool(F\prime)])\right) \\ &= \sigma\left(f^{7\times7}\left(\left[F^s_{avg}; F^s_{\max}\right]\right)\right) \end{aligned} \tag{4}$$

where $M_s(F,\prime)$ is a one-dimensional spatial attention feature map, F^s_{avg} and $F^s_{\max}$ represent the average pooling feature map and the maximum pooling feature map, respectively.

The four attention mechanisms introduced have their own advantages and disadvantages. Among them, the CBAM attention mechanism considers both spatial information and channel information, which can improve the feature extraction ability. Therefore, this paper uses CBAM to improve the ear recognition rate of the YOLOv5s-MG method and the YOLOv5s-MG-CBAM method is constructed.

2.3. SPPF Network

In 2015, He K.M. proposed the SPP module [22], that is, spatial pyramid pooling. The main goal of SPP is to generate a fixed-length vector without considering the size or proportion of the input image. The purpose of SPPF and SPP is the same, but the structure is slightly different. Figure 2 is the SPPF network structure. For any size of the image, the SPPF module is a maximum pooling module with 5×5, 9×9 and 13×13, which is connected in a series to extract the features of the image. The output features of each pooling layer are connected with the input image features to form a fixed-length feature map. Compared with an SPP network, an SPPF network maintains multi-scale information extraction. Therefore, by adding an SPPF network in YOLOv5s-MG method, the accuracy can be effectively improved.

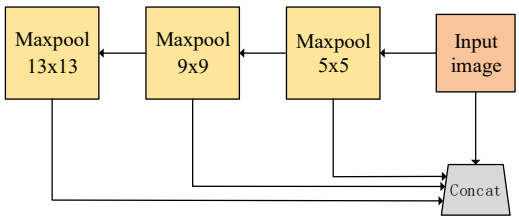


Figure 2. SPPF network.

2.4. PANet Network

In 2017, Lin et al. [23] first proposed the feature fusion network FPN network. FPN uses the hierarchical semantic features of the convolutional network itself to construct a feature pyramid, integrates high-level features and low-level features and determines the target location. It is conducive to the use of the top-level strong semantic features (conducive to classification), and the use of the underlying high-resolution information. In 2018, Shu Liu et al. [22] believed that low-level features are very useful for recognition tasks. The path between high-level features and low-level features is long, and the efficiency of information transmission in the network is not high. Therefore, PANet network is proposed based on FPN network. In order to improve the utilization of low-level information and accelerate the dissemination efficiency of low-level information, a PANet network shortens the information path.

Therefore, by adding a PANet network to the YOLOv5s-MG method, the accuracy can be effectively improved. On the basis of the YOLOv5s-MG-CBAM method, the YOLOv5s-MG-CBAM-F method is proposed through adding an SPPF and PANet network.

3. Proposed Method

3.1. The Overall Framework of the Proposed Method

In this section, a new ear recognition method is proposed. Figure 3 illustrates the overall framework of the method which can be mainly divided into three parts: ear dataset, ear recognition model and ear recognition. The ear recognition model includes the three processes of lightweight, attention and feature fusion.

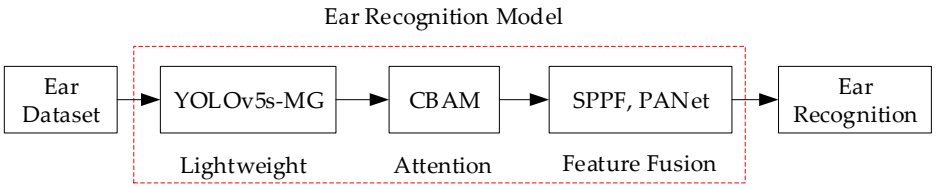


Figure 3. The total structure block diagram of the proposed method.

3.2. The Proposed Method YOLOv5s-MG-CBAM-F

In Section 2.1, a YOLOv5s-MG lightweight ear recognition model is introduced, which reduces the parameter amount and calculation amount of the network model and the model size from two aspects of feature extraction and feature fusion. Section 2.2 introduces four attention mechanisms, each with advantages and disadvantages, but their common purpose is to strengthen the feature extraction ability. In order to further improve the accuracy of human ear recognition on the basis of improving real-time performance, this section will further improve the YOLOv5s-MG method. CBAMNet takes into account both channel information and spatial information, so in this paper, CBAMNet was used to improve the YOLOv5s-MG lightweight network. The improved human ear recognition network YOLOv5s-MG-CBAM-F is shown in Figure 4.

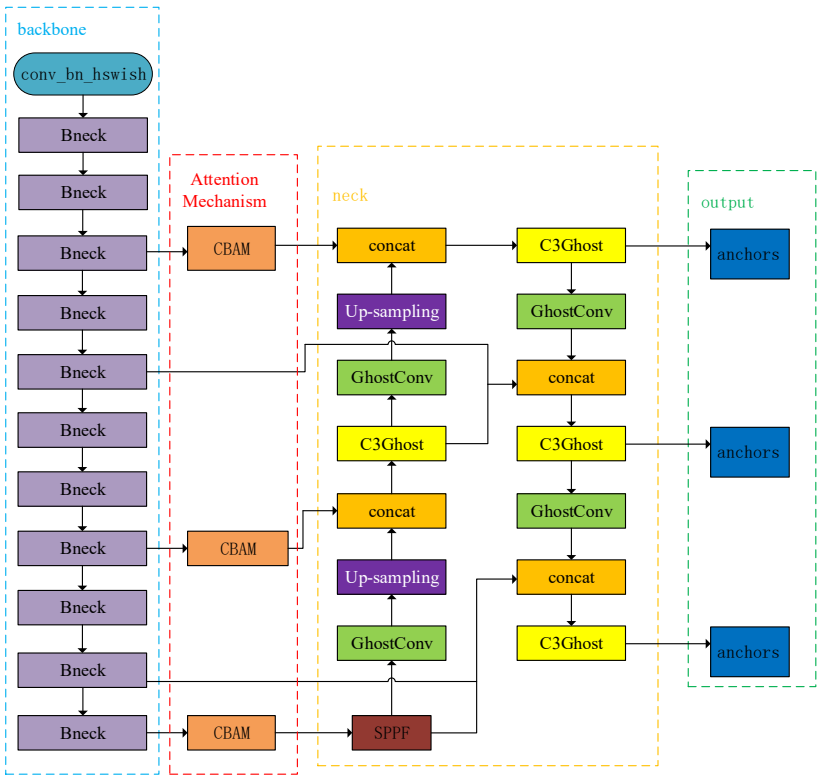


Figure 4. The structure block diagram of the proposed algorithm YOLOv5s-MG-CBAM-F.

In general, attention can be added to the backbone, neck and output modules of the network in YOLOv5s [24,25]: added to the backbone module mainly to strengthen the

network feature extraction capabilities, added to the neck module to strengthen the feature fusion ability and added to the output module to enhance the output prediction ability of the network.

3.2.1. YOLOv5s-MG-CBAM Ear Recognition Network Based on Attention Mechanism

In this paper, we propose to add a CBAM attention module where the backbone and neck modules are connected, as shown in the red box in Figure 4. It can be seen that the entire network needs to perform three attention operations. The first CBAM is connected to the fourth layer of the backbone network and the second concat layer of the neck network. The second CBAM is connected to the ninth layer of the backbone network and the first concat layer of the neck network. The third CBAM is connected to the twelfth layer of the backbone network and the fourth concat layer of the neck network.

Such a connection makes the attention mechanism analyze shallow features and deep features, making the feature information of the feature fusion network more accurate. At the same time, by weighting the attention of the target in the feature map in different dimensions, it can enhance the network's ability to extract important information from the target in the feature map, thereby improving the detection accuracy.

3.2.2. YOLOv5s-MG-CBAM-F Ear Recognition Network Based on Feature Fusion

Through the study of the improved YOLOv5s-MG-CBAM model, compared with the YOLOv5-MG model, the accuracy of human ear recognition will increase, but there will be some limitations. First, the feature fusion of the neck network is not high for shallow information and deep information utilization; second, the backbone network extraction information is not stable enough. Therefore, in this paper we further improve the YOLOv5s-MG-CBAM model based on feature fusion.

Step 1: Network improvement based on SPPF

In order to further increase the multi-feature extraction ability of the network, the SPPF module is added to the YOLOv5s-MG-CBAM ear recognition model. The specific operation is as follows: after connecting CBAM, the SPPF module is added to the last layer Bneck of the backbone of the YOLOv5s-MG-CBAM model, and finally connected to the last layer GhostConv module on the left side of the neck module. Through the maximum pooling of different sizes of pooling kernels in the SPPF module, the receptive field of the network is improved and the feature extraction ability is increased.

Step 2: Cross-regional feature fusion

As the feature extraction operation of the neural network continues, the underlying features and high-level features in the image will be decomposed layer by layer. The underlying features of the image contain less semantic information, but the feature map has large resolution, sufficient location information and accurate target location. The high-level features of the image have rich semantic information, but the target location is rough. Therefore, the fusion of different levels of features helps the neural network to be more accurate in target detection and recognition.

Although the PANet network introduced in Section 2.4 uses an extended path to enable the underlying information to be transmitted to the higher layer faster and improve the recognition accuracy of large targets, it is not enough for small objects and targets with poor image quality. Therefore, in this section, the PANet network will be improved to solve this problem.

As shown in Figure 4, two channels are added on the basis of the YOLOv5s-MG-CBAM network and PANet network [26]. One path starts from the sixth layer of the backbone network and connects to the third concat layer of the neck network. The other path is starting from the eleventh layer of the backbone network and connecting to the fourth concat layer. The two paths complement the context information of the feature fusion layer. More shallow and deep features are added to the network to supplement the information of small objects and targets with poor image quality.

4. Experimental Results and Analysis

4.1. Human Ear Datasets

In order to train and test the model of the YOLOv5s-MG-CBAM-F method proposed in this paper, three different human ear datasets, namely, CCU-DE, USTB and EarVN1.0, were used. In order to facilitate the comparison of performance between different methods, the selection of three datasets is the same as that in Reference [17]. There are 3274, 7700 and 3201 pictures in the CCU-DE, USTB and EarVN1.0 human ear datasets, respectively. The training set, validation set and test set are divided according to 3:1:1. Because the selected human ear datasets have different characteristics of data size, category, attitude change, resolution, gender, dynamic and static, etc., it can detect the performance of the proposed method very well.

4.2. Experimental Setting

In order to test the real time and accuracy of the YOLOv5s-MG-CBAM-F method proposed in this paper, two groups of experiments are set up. The first group compares YOLOv5s-MG-CBAM with YOLOv5s-MG-SE, YOLOv5s-MG-ECA, YOLOv5s-MG-CA, the original YOLOv5s model, the original YOLOv7 model and YOLOv5s-MG. The second group compares YOLOv5s-MG-CBAM-F with the original YOLOv5s model, the original YOLOv7 model, YOLOv5s-MG-SE-F, YOLOv5s-MG-ECA-F, YOLOv5s-MG-CA-F and YOLOv5s-MG. The connection of YOLOv5s-MG-SE/CA/ECA and YOLOv5s-MG-SE/CA/ECA-F is the same as YOLOv5s-MG-CBAM and YOLOv5s-MG-CBAM-F.

The hardware platforms in the experiment mainly include CPU is Intel(R) Core (TM) i5-10400F and CPU@2.90 GHZ, GPU is NVIDIA GeFORCE RTX 3060, Memory is 16 GB, GPU-accelerated libraries are CUDA11.1.134 and CUDNN8.0.5. Operating system is Windows10 (64 bit). The software is pytorch1.8 and python3.8. The main parameters used in the model training are the following: the initial learning rate is 0.01, batch size is 16, weight attenuation coefficient is 0.0005, momentum coefficient is 0.937, learning rate reduction coefficient is 0.2 and optimizer is SGD with the momentum. The hardware, software and parameters of the experiment in this paper are the same as those in Reference [17].

4.3. Evaluation Indicators

In order to compare the performance between the improved YOLOv5s-MG-CBAM and YOLOv5s-MG-CBAM-F methods and the YOLOv5s-MG [17] method, four evaluation indexes of mAP, model size/MB, GFLOPs/G and params/M are used [17].

mAP is the mean average precision and is a performance index to describe the accuracy of the ear recognition method. In this paper, mAP@0.5 and mAP@0.5:0.95 are used. Model size, GFLOPs and params are performance indexes to describe the real time of the ear recognition method. The larger is mAP and the smaller are model size, GFLOPs and params; and the ear recognition method is better.

4.4. Ear Recognition Experiments of the Improved YOLOv5s-MG-CBAM on Three Datasets

In order to test the accuracy of ear recognition of the YOLOv5s-MG-CBAM method based on the attention mechanism and the YOLOv5s-MG method proposed in this paper, and verify the similarities and differences between the four lightweight networks based on the attention mechanism YOLOv5s-MG-CBAM, YOLOv5s-MG-SE, YOLOv5s-MG-ECA, YOLOv5s-MG-CA and the original YOLOv5s model and YOLOv5s-MG, the six networks with the same parameters on CCU-DE, USTB and EarVN1.0 human ear datasets were trained and tested.

Figure 5 is the training curves of YOLOv5s-MG-CBAM (proposed), YOLOv5s-MG, YOLOv5s-MG-SE, YOLOv5s-MG-CA, YOLOv5s-MG-ECA and YOLOv5s networks by using the training set and validation set in three human ear datasets. The abscissa and the ordinate are the training epoch and mAP@0.5, respectively. It can be seen from Figure 5 when the number of epochs gradually increases, the model tends to converge gradually.

However, the epoch of the model convergence on the CCU-DE, USTB and EarVN1.0 ear dataset is 40, 50 and 500, respectively.

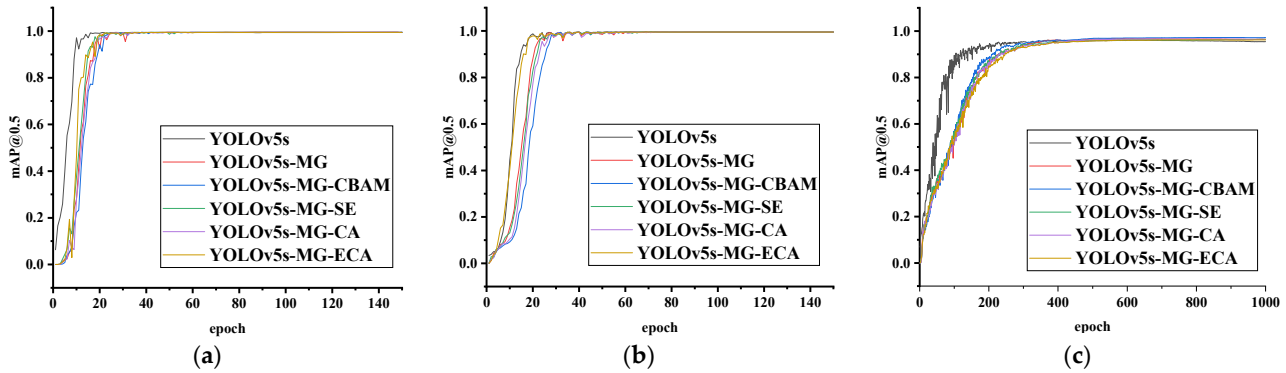


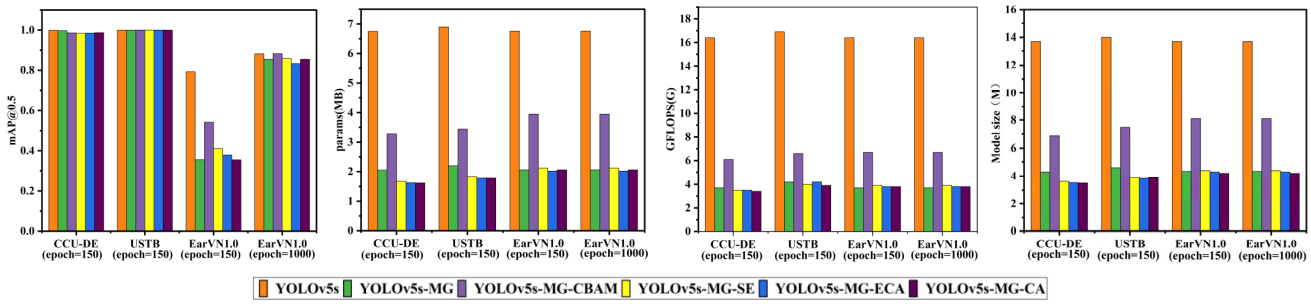
Figure 5. The mAP@0.5 value of YOLOv5s-MG-CBAM (proposed), YOLOv5s-MG, YOLOv5s-MG-SE, YOLOv5s-MG-CA, YOLOv5s-MG-ECA and YOLOv5s models trained on three human ear datasets. (**a**) CCU-DE ear dataset; (**b**) USTB ear dataset; (**c**) EarVN1.0 ear dataset.

As shown in Figure 5, compared with YOLOv5s-MG-SE, YOLOv5s-MG-ECA and YOLOv5s-MG-CA, YOLOv5s-MG-CBAM converges fastest on the EarVN1.0 human ear dataset, and converges slowly on the CCU-DE and USTB human ear datasets. This is mainly because the YOLOv5s-MG-CBAM method is mainly designed for the poor image quality and low resolution of the EarVN1.0 human ear dataset, while the image quality of CCU-DE and USTB human ear datasets is good and overfitting occurs when training with higher resolution.

In this experiment, epoch = 150 and epoch = 1000 were chosen to quantitatively test the difference between the six models. Table 1 and Figure 6 are the experimental results of the six methods on the test sets of three human ear datasets.

Table 1. Quantitative comparison results of the improved YOLOv5s-MG-CBAM and other methods.

Human Ear Dataset	Model	Params (M)	GFLOPS(G)	Model Size (MB)	mAP@0.5
CCU-DE (epoch = 150)	YOLOv5s [17]	6.75 [17]	16.4 [17]	13.7 [17]	0.999 [17]
	YOLOv5s-MG [17]	2.05 [17]	3.7 [17]	4.3 [17]	0.997 [17]
	YOLOv5s-MG-CBAM	3.28	6.1	6.86	0.986
	YOLOv5s-MG-SE	1.68	3.5	3.62	0.984
	YOLOv5s-MG-ECA	1.63	3.5	3.52	0.985
	YOLOv5s-MG-CA	1.62	3.4	3.5	0.987
USTB (epoch = 150)	YOLOv5s [17]	6.9 [17]	16.9 [17]	14 [17]	1 [17]
	YOLOv5s-MG [17]	2.2 [17]	4.2 [17]	4.6 [17]	1 [17]
	YOLOv5s-MG-CBAM	3.44	6.6	7.5	1
	YOLOv5s-MG-SE	1.83	4.0	3.9	1
	YOLOv5s-MG-ECA	1.786	4.2	3.84	1
	YOLOv5s-MG-CA	1.786	3.9	3.9	1
EarVN1.0 (epoch = 150)	YOLOv5s [17]	6.76 [17]	16.4 [17]	13.7 [17]	0.793 [17]
	YOLOv5s-MG [17]	2.06 [17]	3.7 [17]	4.34 [17]	0.356 [17]
	YOLOv5s-MG-CBAM	3.95	6.7	8.1	0.542
	YOLOv5s-MG-SE	2.12	3.9	4.4	0.41
	YOLOv5s-MG-ECA	2.02	3.8	4.3	0.379
	YOLOv5s-MG-CA	2.06	3.8	4.2	0.355
EarVN1.0 (epoch = 1000)	YOLOv5s [17]	6.76 [17]	16.4 [17]	13.7 [17]	0.882 [17]
	YOLOv5s-MG [17]	2.06 [17]	3.7 [17]	4.34 [17]	0.855 [17]
	YOLOv5s-MG-CBAM	3.95	6.7	8.1	0.883
	YOLOv5s-MG-SE	2.12	3.9	4.4	0.859
	YOLOv5s-MG-ECA	2.02	3.8	4.3	0.833
	YOLOv5s-MG-CA	2.06	3.8	4.2	0.855



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

Figure 6. The comparison results of YOLOv5s-MG-CBAM (proposed), YOLOv5s-MG, YOLOv5s-MG-SE, YOLOv5s-MG-CA, YOLOv5s-MG-ECA and YOLOv5s models on three datasets. (**a**) mAP@0.5; (**b**) params; (**c**) GFLOPS; (**d**) the model size.

From Table 1 and Figure 6a, on the USTB human ear dataset, YOLOv5s-MG-CBAM has the highest ear recognition rate and the mAP@0.5 value is 1. CCU-DE follows; mAP@0.5 is above 0.98 and the mAP@0.5 value of YOLOv5s-MG-CBAM was within ±0.2% compared with other methods. Compared with the CCU-DE and USTB ear datasets, the EarVN1.0 dataset has the lowest ear recognition rate. However, on the EarVN1.0 dataset, when epoch = 1000, the mAP@0.5 value of YOLOv5s-MG-CBAM is 0.883. The human ear recognition rate is the highest among the six methods, which is 0.1% higher than that of YOLOv5s, 2.8% higher than that of YOLOv5s-MG, 2.4% higher than that of YOLOv5s-MG-SE, 5.0% higher than that of YOLOv5s-MG-ECA and 2.8% higher than that of YOLOv5s-MG-CA.

It can be seen from Table 1 and Figure 6b–d that the params, GFLOPS and model size of the YOLOv5s-MG-CBAM method are slightly higher than those of YOLOv5s-MG-SE, YOLOv5s-MG-ECA, YOLOv5s-MG-CA and YOLOv5s-MG. This is mainly because CBAM takes into account both channel attention and spatial attention. Theoretically, its parameter quantity and calculation amount are larger than those of SE, ECA and CA, and the experiment also verifies this.

From Table 1 and Figure 6b–d, it can be seen that the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are much smaller than those of YOLOv5s models. On the CCU-DE human ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 48.6%, 37.2% and 50.1% of YOLOv5s, respectively. On the USTB human ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 49.9%, 39.1% and 53.6% of YOLOv5s, respectively. On the EarVN1.0 ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 58.4%, 40.9% and 59.1% of YOLOv5s, respectively.

From Table 1 and Figure 6b–d, it can be also seen that the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are much larger than those of YOLOv5s-MG models. Comparing those of the YOLOv5s-MG method, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are increased by 60%, 64.9% and 59.5%, respectively, on the CCU-DE human ear dataset, 56.4%, 57.1% and 63.0%, respectively, on the USTB human ear dataset and 91.7%, 81.1% and 86.6%, respectively, on the EarVN1.0 human ear dataset.

4.5. Ear Recognition Experiments of the Improved YOLOv5s-MG-CBAM-F on Three Datasets

In order to verify the feasibility and effectiveness of the YOLOv5s-MG-CBAM-F based on feature fusion proposed in this paper, and verify the similarities and differences between YOLOv5s-MG-CBAM-F, YOLOv5s-MG-SE-F, YOLOv5s-MG-ECA-F, YOLOv5s-MG-CA-F and the original YOLOv5s model and YOLOv5s-MG, the six networks with the same parameters on CCU-DE, USTB and EarVN1.0 human ear datasets were trained and tested.

The six methods YOLOv5s-MG-CBAM-F (proposed), YOLOv5s-MG, YOLOv5s-MG-SE-F, YOLOv5s-MG-CA-F, YOLOv5s-MG-ECA-F and YOLOv5s are tested by using test sets on three human ear datasets, and the training curves are shown in Figure 7. It can be

seen from Figure 7 that the value of mAP@0.5 gradually tends to be stable and the model converges when epoch increases. The epoch of the model convergence is about 40, 50 and 500 on the CCU-DE, USTB and EarVN1.0 human ear dataset respectively. This is mainly related to the difference pose, resolution, image size and number of each human ear dataset. From Figure 7, it can be found that compared with YOLOv5s-MG-SE-F, YOLOv5s-MG-ECA-F and YOLOv5s-MG-CA-F, YOLOv5s-MG-CBAM-F has the fastest convergence speed on the three human ear datasets.

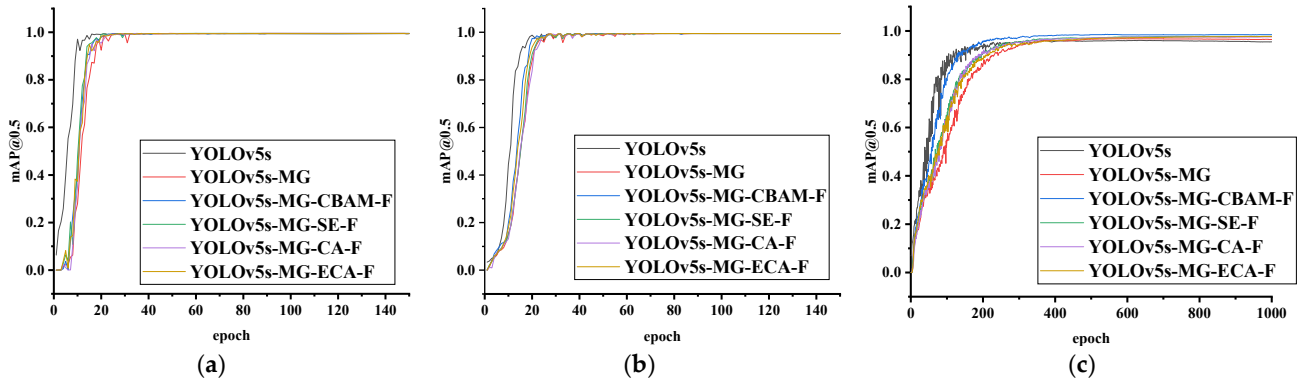
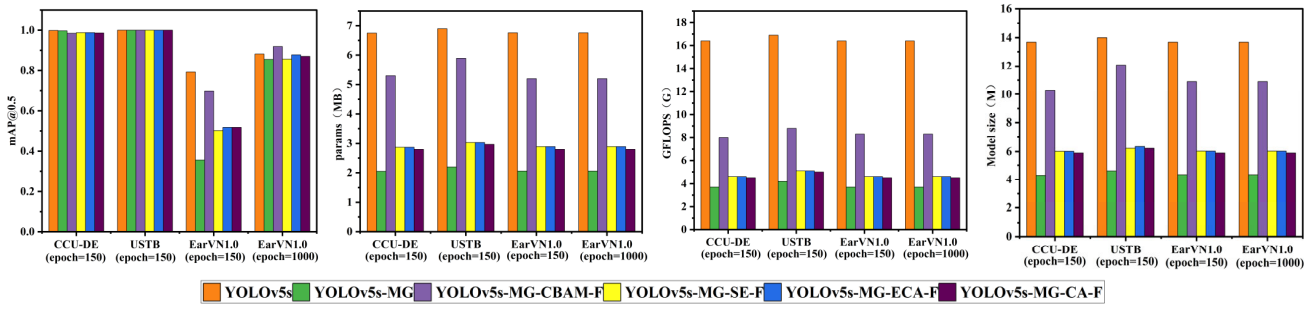


Figure 7. The mAP@0.5 value of YOLOv5s-MG-CBAM-F (proposed), YOLOv5s-MG (proposed), YOLOv5s-MG-SE-F, YOLOv5s-MG-CA-F, YOLOv5s-MG-ECA-F and YOLOv5s models trained on three human ear datasets. (**a**) CCU-DE human ear dataset; (**b**) USTB human ear dataset; (**c**) EarVN1.0 human ear dataset.

In this experiment, epoch = 150 and epoch = 1000 were selected as quantities to verify the difference between the six models. The experimental results on testing a set of the three human ear datasets are shown in Table 2 and Figure 8.

Table 2. Quantitative comparison results of the improved YOLOv5s-MG-CBAM-F and other methods.

Ear Dataset	Model	Params (M)	GFLOPS (G)	Model Size (MB)	mAP@0.5	mAP@0.5:0.95
CCU-DE (epoch = 150)	YOLOv5s [17]	6.75 [17]	16.4 [17]	13.7 [17]	0.999 [17]	0.826
	YOLOv5s-MG [17]	2.05 [17]	3.7 [17]	4.3 [17]	0.997 [17]	0.771
	YOLOv5s-MG-CBAM-F	5.3	8	10.3	0.986	0.781
	YOLOv5s-MG-SE-F	2.87	4.6	6.02	0.988	0.766
	YOLOv5s-MG-ECA-F	2.87	4.6	6.02	0.988	0.777
	YOLOv5s-MG-CA-F	2.8	4.5	5.9	0.986	0.771
USTB (epoch = 150)	YOLOv5s [17]	6.9 [17]	16.9 [17]	14 [17]	1 [17]	0.918
	YOLOv5s-MG [17]	2.2 [17]	4.2 [17]	4.6 [17]	1 [17]	0.903
	YOLOv5s-MG-CBAM-F	5.89	8.8	12.08	1	0.906
	YOLOv5s-MG-SE-F	3.03	5.1	6.23	1	0.903
	YOLOv5s-MG-ECA-F	3.03	5.1	6.35	1	0.904
	YOLOv5s-MG-CA-F	2.97	5	6.23	1	0.906
EarVN1.0 (epoch = 150)	YOLOv5s [17]	6.76 [17]	16.4 [17]	13.7 [17]	0.793 [17]	0.635
	YOLOv5s-MG [17]	2.06 [17]	3.7 [17]	4.34 [17]	0.356 [17]	0.281
	YOLOv5s-MG-CBAM-F	5.2	8.3	10.9	0.698	0.559
	YOLOv5s-MG-SE-F	2.89	4.6	6.03	0.502	0.398
	YOLOv5s-MG-ECA-F	2.89	4.6	6.03	0.518	0.407
	YOLOv5s-MG-CA-F	2.8	4.5	5.9	0.518	0.414
EarVN1.0 (epoch = 1000)	YOLOv5s [17]	6.76 [17]	16.4 [17]	13.7 [17]	0.882 [17]	0.712
	YOLOv5s-MG [17]	2.06 [17]	3.7 [17]	4.34 [17]	0.855 [17]	0.706
	YOLOv5s-MG-CBAM-F	5.2	8.3	10.9	0.919	0.755
	YOLOv5s-MG-SE-F	2.89	4.6	6.03	0.856	0.702
	YOLOv5s-MG-ECA-F	2.89	4.6	6.03	0.878	0.718
	YOLOv5s-MG-CA-F	2.8	4.5	5.9	0.870	0.713



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

Figure 8. The comparison results of YOLOv5s-MG-CBAM-F (proposed), YOLOv5s-MG, YOLOv5s-MG-SE-F, YOLOv5s-MG-CA-F, YOLOv5s-MG-ECA-F and YOLOv5s models on three datasets. (**a**) mAP@0.5; (**b**) the parameter quantity (params); (**c**) the calculation amount (GFLOPS); (**d**) the model size.

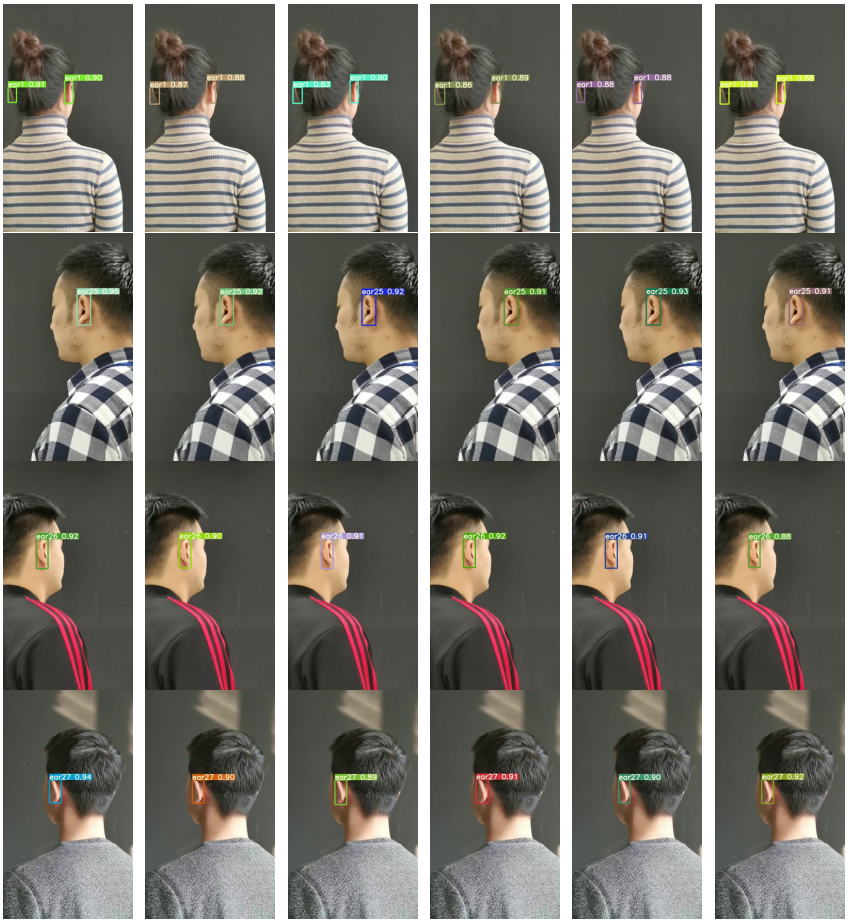
From Table 2 and Figure 8a, on the USTB human ear dataset, YOLOv5s-MG-CBAM-F has the highest ear recognition rate in six methods and the mAP@0.5 value is 1. CCU-DE follows; mAP@0.5 is above 0.98 and the mAP@0.5 value of YOLOv5s-MG-CBAM-F was within ±0.3% compared with other methods. Compared with the CCU-DE and USTB ear datasets, the EarVN1.0 dataset has the lowest ear recognition rate. However, on the EarVN1.0 dataset, when epoch = 1000, the mAP@0.5 value of YOLOv5s-MG-CBAM-F is 0.919. The human ear recognition rate is the highest among the six methods, which is 3.7% higher than YOLOv5s, 6.4% higher than YOLOv5s-MG, 6.3% higher than YOLOv5s-MG-SE-F, 4.1% higher than YOLOv5s-MG-ECA-F and 4.9% higher than YOLOv5s-MG-CA-F. However, the mAP@0.5 value of YOLOv5s-MG-SE-F, YOLOv5s-MG-ECA-F and YOLOv5s-MG-ECA-F is lower than that of YOLOv5s.

From Table 2 and Figure 8b–d, we can see that the parameters, calculation and model size of the YOLOv5s-MG-CBAM-F method are lower than those of YOLOv5s, but higher than other methods. However, YOLOv5s-MG-CBAM-F method still has a lower params, GFLOPS and model size.

On the CCU-DE human ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 78.5%, 48.8% and 75.2% of YOLOv5s, respectively. On the USTB human ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 85.4%, 52.1% and 86.3% of YOLOv5s, respectively. On the EarVN1.0 ear dataset, the params, GFLOPS and model size of the improved YOLOv5s-MG-CBAM model are 76.9%, 50.6% and 79.6% of YOLOv5s, respectively.

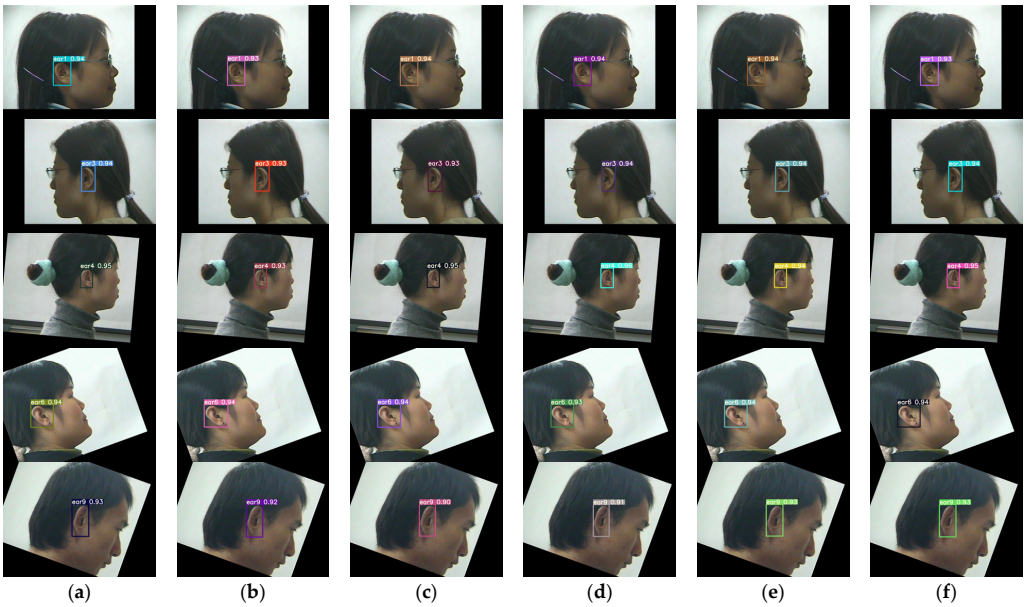
Comparing the experimental results before and after feature fusion, it can be seen from Tables 1 and 2 that on the EarVN1.0 dataset, with the exception that the mAP@0.5 value of YOLOv5s-MG-SE-F is 0.3% lower than that of YOLOv5s-MG-SE, YOLOv5s-MG-CBAM-F, YOLOv5s-MG-ECA-F and YOLOv5s-MG-CA-F were 3.6%, 4.5% and 1.5% higher than YOLOv5s-MG-CBAM, YOLOv5s-MG-ECA and YOLOv5s-MG-CA, respectively. The experiment quantitatively proves that by adding the SPPF network and PANet network in the YOLOv5s-MG-CBAM with an attention mechanism, the feature extraction and feature fusion ability of the network are indeed increased, and the accuracy of human ear recognition is improved. At the same time, the params, GFLOPS and model size are also increased, but they are lower than the YOLOv5s ear recognition network.

The quantitative ear recognition results of the six methods on testing the set of three datasets are shown in Figures 9 and 10 and Table 3.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

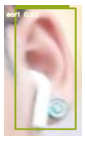
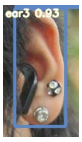
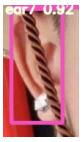
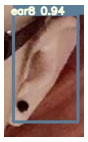
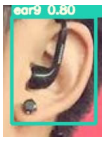
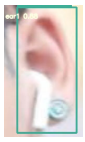
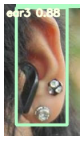
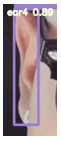
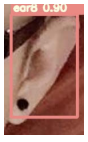
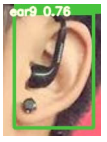
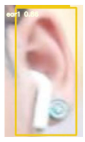
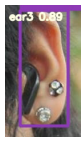
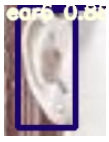
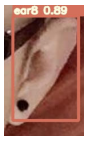
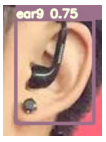
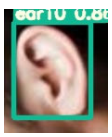
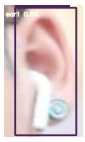
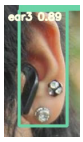
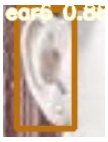
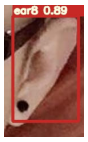
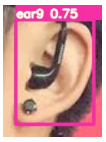
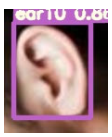
Figure 9. The quantitative ear recognition results of the six methods on CCU-DE datasets. (**a**) YOLOv5s-MG-CBAM-F (proposed); (**b**) YOLOv5s-MG-CA-F; (**c**) YOLOv5s-MG-ECA-F; (**d**) YOLOv5s-MG-SE-F; (**e**) YOLOv5s; (**f**) YOLOv5s-MG.



(**a**) (**b**) (**c**) (**d**) (**e**) (**f**)

Figure 10. The quantitative ear recognition results of the six methods on USTB datasets. (**a**) YOLOv5s-MG-CBAM-F (proposed); (**b**) YOLOv5s-MG-CA-F; (**c**) YOLOv5s-MG-ECA-F; (**d**) YOLOv5s-MG-SE-F; (**e**) YOLOv5s; (**f**) YOLOv5s-MG.

Table 3. Quantitative ear recognition on EarVN1.0 dataset comparison results of the improved YOLOv5s-MG-CBAM-F and other methods.

Method	Ear1	Ear2	Ear3	Ear4	Ear5	Ear6	Ear7	Ear8	Ear9	Ear10
YOLOv5s-MG-CBAM-F										
YOLOv5s-MG-CA-F										
YOLOv5s-MG-ECA-F										
YOLOv5s-MG-SE-F										
YOLOv5s										
YOLOv5s-MG										

From Figure 9, we can see that all the six methods can identify the four types of human ears on the CCU-DE dataset and the range of recognition results is 87–95%. The four improved YOLOv5s-MG methods based on the attention mechanism have similar recognition effects, while the ear recognition results of the YOLOv5s-MG method are slightly lower, which is consistent with the experimental results and theoretical analysis in Table 2. From Figure 9, we can also see that the size of the picture is large, and the human ear is a small target relative to the entire picture. Therefore, the method proposed in this paper is also suitable for small target recognition. The ear recognition results of the side face and the back face are given in the figure, and the ear recognition effect of the front face is better.

From Figure 10, we can see that all the six methods can identify the five types of human ears on the USTB dataset and the range of recognition results is 90–96%. The recognition effects of the six methods on the USTB ear dataset are almost the same. However, the ear recognition effect of the six methods on the USTB dataset is better than that of the CCU-DE ear dataset. For an actual human ear recognition application system, the human ear is often in a dynamic environment with pose changes such as translation, rotation, illumination and contrast changes, or occlusion. Therefore, only when all possible situations in practice are taken into account in the experiment, can the performance of the proposed method be better verified for application. From Figure 10, we can also see that the USTB human ear dataset is obtained by flipping, rotating a certain angle, illumination and contrast on

the original image. Therefore, the method proposed in this paper is also suitable for ear recognition in some dynamic environment changes.

Table 3 is the quantitative ear recognition results of the six methods on EarVN1.0 datasets. From Figure 6, we can see that on 10 types of human ears, the range of recognition results of the proposed YOLOv5s-MG-CBAM-F and YOLOv5s-MG method are 80–94% and 71–92%. This shows that the use of the attention mechanism and feature fusion in the YOLOv5s-MG method can indeed improve the accuracy of ear recognition. Among the six methods, the ear recognition accuracy of the YOLOv5s-MG method is the lowest. Take ear1 as an example, the ear recognition results of the six methods are 93%, 89%, 88%, 88%, 89% and 71%. The recognition result of the YOLOv5s-MG-CBAM-F method is the highest and the recognition result of the YOLOv5s-MG method is the lowest. From Figure 6, we can also see that the image size, brightness and clarity of the EarVN1.0 dataset are different. At the same time, most of the ears have one or more occlusions such as headphones, glasses and earrings. Therefore, the YOLOv5s-MG-CBAM-F method proposed in this paper is suitable for ear recognition with occlusion. At the same time, compared with the CCU-DE and USTB datasets, Table 3 also explains the reason for the low ear recognition rate on the EarVN1.0 dataset. However, from Table 3, we can see that the YOLOv5s-MG-CBAM-F method can meet the human ear recognition in practical applications.

4.6. The Computational Complexity Analysis

In order to test the real time of ear recognition, all experiments were performed under the same conditions and the inference time per image is shown in Table 4. From Table 4, we can see that for the same image, whether on CPU or GPU, the YOLOv5s-MG method has the fastest ear recognition speed, while the YOLOv7 method has the slowest recognition speed. The inference time per image of YOLOv5s-MG-CA-F, YOLOv5s-MG-ECA-F and YOLOv5s-MG-SE-F method is similar.

Table 4. The inference time per image of the six methods on the testing set of three datasets. The unit is milliseconds (ms).

Human Ear Dataset	Device	Model					
		YOLOv5s-MG	YOLOv5s-MG-CBAM-F	YOLOv5s-MG-CA-F	YOLOv5s-MG-SE-F	YOLOv5s-MG-ECA-F	YOLOv7
CCU-DE (epoch = 150)	CPU	69	114	81.2	82.5	82.6	1000
	GPU	12.4	15.9	15	14.5	15.1	711.9
USTB (epoch = 150)	CPU	117.6	170.2	126.9	128.5	128.5	769.2
	GPU	12.8	16.2	15.4	15.1	14.9	975.4
EarVN1.0 (epoch = 150)	CPU	89.3	137.8	96.9	94.5	101.1	909.1
	GPU	12.8	16	14.8	14.3	14.3	803.5
EarVN1.0 (epoch = 1000)	CPU	87.7	137.6	117.4	103.4	113.4	769.2
	GPU	12.6	16.4	15.2	14.8	14.7	807

From Table 4, it can be seen that on the CCU-DE, USTB and EarVN1.0 ear dataset, the inference time per image of the improved YOLOv5s-MG-CBAM-F model increased by 3.5 ms, 3.4 ms and 3.8 ms compared with that of YOLOv5s-MG, respectively, by using GPU.

4.7. Selection Strategy of Ear Recognition Method

In this paper, based on YOLOv5s-MG, two improved ear recognition methods, YOLOv5s-MG-CBAM and YOLOv5s-MG-CBAM-F, were proposed from the perspective of improving accuracy based on the lightweight ear recognition method YOLOv5s-MG. The comparison results of the four methods are shown in Figure 11.

It can be seen from Figure 11 that all the four methods can realize human ear recognition. Among the two improved methods, for the three distinctive human ear datasets, the YOLOv5s-MG-CBAM-F method has the best accuracy.

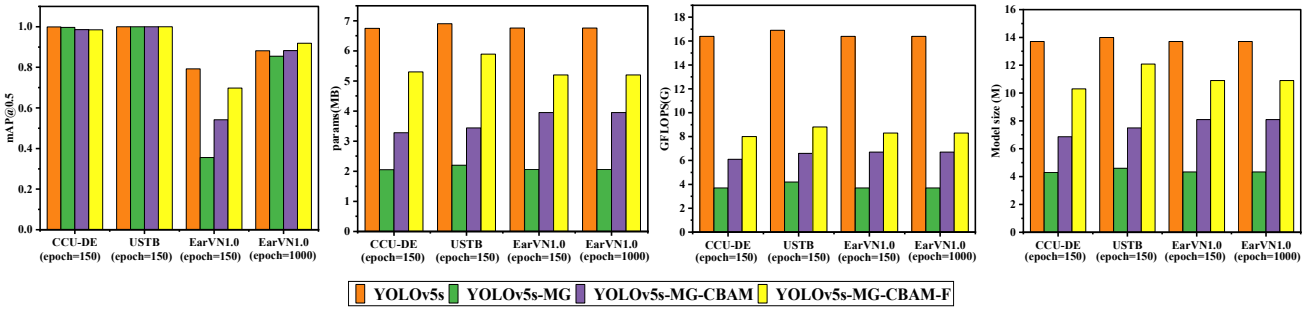


Figure 11. The comparison results of YOLOv5s-MG-CBAM-F (proposed), YOLOv5s-MG-CBAM (proposed), YOLOv5s-MG and YOLOv5s models on three datasets. (**a**) mAP@0.5; (**b**) the parameter quantity (params); (**c**) the calculation amount (GFLOPS); (**d**) The model size.

It can be seen from Figure 11a that the same method has different ear recognition accuracy on different ear datasets, which is mainly related to three factors: the human ear dataset, feature extraction ability of the model and convergence of model training. The human ear dataset with high resolution, a small amount of data and simple posture change has a high ear recognition rate, that is, the mAP@0.5 value is large on the USTB human ear dataset, medium on the CCU-DE human ear dataset and small on the EarVN1.0 human ear dataset. Because the YOLOv5s-MG-CBAM-F method has the strongest feature extraction ability, the mAP@0.5 value is the largest on the EarVN1.0 human ear dataset. On the EarVN1.0 dataset, the mAP@0.5 value is low when the model does not converge at epoch = 150, and is high when the model converges at epoch = 1000.

It can be seen from Figure 11b–d that the real-time performance of the improved methods YOLOv5s-MG-CBAM-F is worse than the method YOLOv5s-MG before improvement. However, because the real-time performance of the YOLOv5s-MG method is particularly good, the improved YOLOv5s-MG-CBAM-F method can still meet the real-time requirements.

According to the theoretical analysis of Section 3 and the experimental results and analysis of Section 4.4, Section 4.5, Section 4.6, Section 4.7, the selection strategy of the human ear recognition method is formulated, that is, according to different human ear datasets, the principle of accuracy priority is adopted when the accuracy and real-time performance meet the requirements at the same time. For the three human ear datasets in this paper, the YOLOv5s-MG method is selected on the CCU-DE and USTB human ear datasets; on the EarVN1.0 dataset, the YOLOv5s-MG-CBAM-F method is selected.

5. Conclusions

In this paper, two improved ear recognition methods are proposed: YOLOv5s-MG-CBAM and YOLOv5s-MG-CBAM-F. Experiments were carried out on three distinctive human ear datasets of CCU-DE, USTB and EarVN1.0, and four performance indicators of mAP@0.5, params, GFLOPS and model size were used to evaluate the improved method. Quantitative experimental results show that the two methods proposed in this paper can realize ear recognition and meet the requirements of real time and accuracy.

Compared with YOLOv5s-MG, the mAP@0.5 value of the proposed YOLOv5s-MG-CBAM-F method on the EarVN1.0 ear dataset increased by 6.4% and the inference time per image of the improved YOLOv5s-MG-CBAM-F model increased by 3.8 ms compared with that of YOLOv5s-MG by using GPU.

The quantitative results show that compared with YOLOv5s-MG, the YOLOv5s-MG-CBAM-F method has the best ear recognition accuracy, especially for images with poor resolution and rich posture in the EarVN1.0 ear dataset. The method proposed in this paper can meet the performance of human ear recognition accuracy and real-time performance, and has a good prospect for the application of human ear recognition devices in the field

of identity recognition. The next step will apply the method proposed in this paper to the actual ear recognition equipment.

Author Contributions: Conceptualization, Y.L. and D.P.; methodology, Y.L. and D.P.; software, D.P.; validation, J.Q. and Z.F.; investigation, Z.F.; data curation, D.P., J.Q. and Z.F.; writing—original draft preparation, D.P.; writing—review and editing, Y.L.; project administration, Y.L.; funding acquisition, Y.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported by the funds of Education Department of Jilin Province (number: 2022LY502L16).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

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

References

1. Oyebiyi, O.G.; Abayomi-Alli, A.; Arogundade, O.T.; Qazi, A.; Imoize, A.L.; Awotunde, J.B. A Systematic Literature Review on Human Ear Biometrics: Approaches, Algorithms, and Trend in the LastDecade. *Information* **2023**, *14*, 192. [CrossRef]
2. Zhang, Y.; Mu, Z. Ear Detection under Uncontrolled Conditions with Multiple Scale Faster Region-Based Convolutional Neural Networks. *Symmetry* **2017**, *9*, 53. [CrossRef]
3. Fan, T.Y.; Mu, Z.C.; Yang, R.Y. Multi-modality recognition of human face and ear based on deep learning. In Proceedings of the 2017 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Ningbo, China, 9–12 July 2017.
4. Kamboj, J.; Rani, R.; Nigam, A. A comprehensive survey and deep learning-based approach for human recognition using ear biometric. *Vis. Comput.* **2022**, *38*, 2383–2416. [CrossRef] [PubMed]
5. Kamboj, A.; Rani, R.; Nigam, A.; Jha, R.R. CED-Net: Context-aware ear detection network for unconstrained images. *Pattern Anal. Appl.* **2021**, *24*, 779–800. [CrossRef]
6. Qian, J.R. Research on Dynamic Human Ear Recognition Method Based on deep Learning. Master's Thesis, Chang Chun University, Changchun, China, 2020.
7. Quoc, H.N.; Hoang, V.T. Human ear-side detection based on YOLOv5 detector and deep neural networks. In Proceedings of the 2021 International Conference on Decision Aid Sciences and Application (DASA), Sakheer, Bahrain, 7–8 December 2021.
8. Tian, Y.; Wang, S.; Li, W. Human ear recognition based on deep convolutional neural network. In Proceedings of the 2018 Chinese Control and Decision Conference (CCDC), Shenyang, China, 9–11 June 2018.
9. Mehraj, H.; Mir, A.H. Human Recognition using Ear based Deep Learning Features. In Proceedings of the 2020 International Conference on Emerging Smart Computing and Informatics (ESCI), Pune, India, 12–14 March 2020.
10. Dvoršak, G.; Dwivedi, A.; Štruc, V.; Peer, P.; Emeršič, Ž. Kinship Verification from Ear Images: An Explorative Study with Deep Learning Models. In Proceedings of the 2022 International Workshop on Biometrics and Forensics (IWBF), Salzburg, Austria, 20–21 April 2022.
11. El-Naggar, S.; Bourlai, T. Image Quality Assessment for Effective Ear Recognition. *IEEE Access* **2022**, *10*, 98153–98164. [CrossRef]
12. El-Naggar, S.; Bourlai, T. Exploring Deep Learning Ear Recognition in Thermal Images. *IEEE Trans. Biom. Behav. Identity Sci.* **2023**, *5*, 64–75. [CrossRef]
13. Yuan, L.; Mao, J.; Zheng, H. Ear Detection based on CenterNet. In Proceedings of the 2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT), Weihai, China, 14–16 October 2020.
14. Mursalin, M.; Islam, S.M.S. Deep Learning for 3D Ear Detection: A Complete Pipeline from Data Generation to Segmentation. *IEEE Access* **2021**, *9*, 164976–164985. [CrossRef]
15. Mursalin, M.; Ahmed, M.; Haskell-Dowland, P. Biometric Security: A Novel Ear Recognition Approach Using a 3D Morphable Ear Model. *Sensors* **2022**, *22*, 8988. [CrossRef] [PubMed]
16. Bahadir, K.; Fatih, Y.; Emin, B. A hybrid approach based on deep learning for gender recognition using human ear images. *J. Fac. Eng. Archit. Gazi Univ.* **2022**, *37*, 1579–1594.
17. Lei, Y.; Pan, D.; Feng, Z.; Qian, J. Lightweight YOLOv5s Human Ear Recognition Based on MobileNetV3 and Ghostnet. *Appl. Sci.* **2023**, *13*, 6667. [CrossRef]
18. Jie, H.; Li, S.; Gang, S. Squeeze-and-Excitation Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2019**, *42*, 2011–2023.
19. Hou, Q.; Zhou, D.; Feng, J. Coordinate Attention for Efficient Mobile Network Design. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021; pp. 13708–13717.
20. Wang, Q.; Wu, B.; Zhu, P. ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–9 June 2020.
21. Woo, S.; Park, J.; Lee, J.Y. CBAM: Convolutional block attention module. In Proceedings of the European Conference on Computer Vision (ECCV), Glasgow, UK, 23–28 August 2020.

22. Liu, S.; Qi, L.; Qin, H. Path Aggregation Network for Instance Segmentation. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 8759–8768.
23. Lin, T.Y.; Dollar, P.; Girshick, R. Feature Pyramid Networks for Object Detection. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 936–944.
24. Yang, K.; Fan, X.J.; Bo, W.H.; Liu, J.; Wang, J.L. Plant disease and pest detection based on vision attention enhancement. *J. Nanjing For. Univ. (Nat. Sci. Ed.)* **2023**, *47*, 11–18.
25. Zhu, R.X.; Yang, F.X. Improved YOLOv5 small object detection algorithm in moving scenes. *Comput. Eng. Appl.* **2023**, *59*, 196–203.
26. Yang, J.H.; Li, H.; Du, Y.Y. A Lightweight Object Detection Algorithm Based on Improved YOLOv5s. *Electron. Opt. Control* **2023**, *30*, 24–30.