applied sciences

MDPI

Article

Lightweight YOLOv5s Human Ear Recognition Based on MobileNetV3 and Ghostnet

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: Ear recognition is a biometric identification technology based on human ear feature information, which can not only detect the human ear in the picture but also determine whose human ear it is, so human identity can be verified by human ear recognition. In order to improve the real-time performance of the ear recognition algorithm and make it better for practical applications, a lightweight ear recognition method based on YOLOv5s is proposed. This method mainly includes the following steps: First, the MobileNetV3 lightweight network is used as the backbone network of the YOLOv5s ear recognition network. Second, using the idea of the Ghostnet network, the C3 module and Conv module in the YOLOv5s neck network are replaced by the C3Ghost module and GhostConv module, and then the YOLOv5s-MG ear recognition model is constructed. Third, three distinctive human ear datasets, CCU-DE, USTB, and EarVN1.0, are collected. Finally, the proposed lightweight ear recognition method is evaluated by four evaluation indexes: mAP value, model size, computational complexity (GFLOPs), and parameter quantity (params). Compared with the best results of YOLOv5s, YOLOv5s-V3, YOLOv5s-V2, and YOLOv5s-G methods on the CCU-DE, USTB, and EarVN1.0 three ear datasets, the params, GFLOPS, and model size of the proposed method YOLOv5s-MG are increased by 35.29%, 38.24%, and 35.57% respectively. The FPS of the proposed method, YOLOv5s-MG, is superior to the other four methods. The experimental results show that the proposed method has the performance of larger FPS, smaller model, fewer calculations, and fewer parameters under the condition of ensuring the accuracy of ear recognition, which can greatly improve the real-time performance and is feasible and effective.

Keywords: YOLOv5s; lightweight; MobileNetV3; Ghostnet; ear recognition; real-time

Citation: Lei, Y.; Pan, D.; Feng, Z.; Qian, J. Lightweight YOLOv5s Human Ear Recognition Based on MobileNetV3 and Ghostnet. *Appl. Sci.* **2023**, *13*, 6667. https://doi.org/10.3390/app13116667

Academic Editors: El-Sayed El-Alfy, Motaz Alfarraj and Abdul Jabbar Siddiqui

Received: 5 May 2023
Revised: 27 May 2023
Accepted: 27 May 2023
Published: 30 May 2023

1. Introduction

With the development of internet technology, biometric identification technologies have gradually developed. Biometric recognition is based on human features for identification. Generally speaking, the features that can be used for biometric identification should meet the basic attributes of universality, uniqueness, stability, and collectability [1]. When identifying different types of biometric features, their information retention, area, and anti-counterfeiting should also be considered [2]. At present, the commonly used biometric features are mainly face, ear, iris, palmprint, and fingerprint. As a branch of biometric recognition [3], the human ear has fewer features than the face and is not affected by facial expressions, age, makeup, and other factors [4,5]. Compared with palmprint and fingerprint features, it is not easy to damage and lose information [6–8]. Compared with iris features, it provides rich structural information and texture features [9,10]. Moreover, due to the unique recognition angle of the human ear, the requirements for image acquisition equipment are low, so it is widely studied by scholars.

In recent years, with the rapid development of deep-learning-related technologies, the problem of ear recognition in complex environments has also been solved. Although the accuracy of human ear recognition is improved, the human ear recognition network is more complex, increasing the amount of computation, the number of parameters, and the size of the model. Therefore, on the premise of guaranteeing accuracy, reducing the computational burden of the human ear recognition network, the number of parameters, and model size, improvement in the algorithm of real-time ear recognition is of great importance to practical applications.

At present, there are two main categories of ear recognition methods: traditional methods and methods based on deep learning [11]. The traditional method is to extract the global or local features of the human ear from the human ear image for ear recognition. In 2007, Kumar and Zhang [12] used log-Gabor wavelet technology to extract phase information from human ear grey images for ear recognition. In 2015, scholars, such as Anwar [13], extracted SIFT key points from human ear images, calculated the descriptors of each key point, and used the minimum distance classifier for human ear recognition. In 2007, Nosrati [14] proposed a human ear recognition method based on a two-dimensional wavelet transform, using PCA to complete image classification. In 2018, Omara et al. [15] proposed an improved unconstrained ear recognition method based on local feature fusion. LPQ, HOG, LBP, POEM, BSIF, and Gabor were used to extract the local features of human ear images, and then discriminant correlation analysis (DCA) was used for fusion dimension reduction. Finally, a support vector machine (SVM) was used for classification. In 2008, Xie Z.X. [16] and other scholars proposed multi-pose ear recognition based on local linear embedding and introduced popular learning algorithms to improve the robustness of ear recognition. In 2019, Qian Y.L. [17] and other scholars proposed a fast 3D ear recognition method based on local and global information for 3D ear recognition methods.

The two-stage network is divided into two steps for ear recognition. The first step is to generate a candidate box for the ear target, and the second step is to determine the ear category. The representative network mainly has a Faster R-CNN network. In 2018, Susan et al. [18] proposed an ear detection system based on the two-stage network Faster R-CNN and verified it on a test set composed of UDN, FERET, WVU, and AWE ear datasets, with a detection rate of 98%. Zhang Y. [11] proposed a multi-scale Faster-R-CNN ear detection algorithm to solve the problem of poor robustness of ear recognition in uncontrolled scenarios, and the accuracy of USTB-WebEar and USTB-Helloear has reached 99%.

The single-stage network uses the regression network to obtain the ear target classification and detection box. The representative networks are the SSD and YOLO series. In 2021, Lei Y.M. [19] proposed an ear recognition system based on a single-stage network SSD and lightweight network MobileNetv1, and the ear recognition rate reached 98% on the USTB dataset. Qian J.R. [20] proposed a dynamic human ear recognition method based on YOLOv3, and the recognition accuracy in the CCU-DE human ear database has exceeded 90%.

Through the literature research, we found that the above methods can perform ear recognition, but compared with traditional ear recognition algorithms, the ear recognition algorithm based on deep learning has certain advantages [21]. In deep learning, the single-stage network is faster than the two-stage network. Combining with the auxiliary network can improve the recognition accuracy, but it will also increase the amount of computation, the number of parameters, and the size of the model, resulting in unsatisfactory real-time performance of the network.

A key factor in the practical application of human ear recognition is to improve the real-time performance of the method while ensuring recognition accuracy. Therefore, in this paper based on the single-stage deep neural network YOLOv5s, we propose a human ear recognition method based on the lightweight network MobileNetV3 and Ghostnet. This method can simultaneously make the feature extraction of the backbone network and the feature fusion of the neck network of YOLOv5s lightweight. We use four performance indicators (mAP, model size, GFLOPs, and params) to test the YOLOv5s-MG method

proposed in this paper on three datasets (CCU-DE, USTB, and EarVN1.0) and compare it with other lightweight methods. The simulation results show that the method has very low computational complexity, parameter quantity, and model size while ensuring accuracy and improving the real-time performance of the method.

The main contribution of this paper is to propose a lightweight ear recognition method named YOLOv5s-MG. In this method, we replaced the backbone network of YOLOv5s with MobileNetV3 lightweight network so that feature extraction is also lightweight. Simultaneously, by using the idea of the Ghostnet network, we replaced the C3 module and Conv module of the neck network of YOLOv5s with the C3Ghost module and GhostConv module so that feature fusion is lightweight. Therefore, the method proposed in this paper can not only maximize real-time performance but also ensure the accuracy of human ear recognition.

The rest of the paper is arranged as follows. YOLOv5s and three kinds of lightweight networks are introduced in Section 2. The improved YOLOv5s-MG method is proposed in Section 3. Experiments and results analysis are presented in Section 4. The paper is summarized in Section 5.

2. Related Works

2.1. YOLOv5s

YOLO series is a single-stage deep learning method, mainly including YOLOv1, YOLOv2, YOLOv3, YOLOv4, YOLOv5, YOLOv6, YOLOv7, and YOLOv8. The YOLOv1 [22] algorithm divides the image into several grids for regression tasks and uses a grid containing the center point of the object for classification tasks. Its network architecture is composed of 24 convolutional layers and 2 fully connected layers. On the basis of YOLOv1, YOLOv2 [23] uses batch normalization to eliminate gradient disappearance and gradient explosion problems, replaces the full connection layer with an anchor frame, and introduces the pass-through layer to retain more feature information. Compared with YOLOv1, YOLOv2 consists of 19 convolutional layers and 5 maximum pooling layers, which improves speed and accuracy. YOLOv3 [24] uses the residual network to form the Darknet53 network as the backbone feature extraction network according to the idea of the FPN feature pyramid and uses three different sizes of detection heads to detect samples of different sizes. YOLOv4 [25] has made a comprehensive improvement to the network structure, using Mosaic data enhancement to enhance the input data. The SPPNet network and PANet network are used to construct the neck layer. The activation function is replaced with the Mish function. The CIOU_Loss loss function is used to calculate the regression loss of the target box. Compared with YOLOv3, the target recognition ability of YOLOv4 is greatly improved. On the basis of YOLOv4, YOLOv5 [26] has added some new improvement ideas. YOLOv5 adds adaptive image filling and adaptive anchor frame calculation to the input part to process the data, which increases the diversity of the data and improves the accuracy. CSPDarkNet53 network is mainly used in the backbone network, which introduces the Focus structure and CSP structure to improve the speed without losing training accuracy. The output end uses the GIOU_Loss loss function for regression, and the weighted NMS operation filters multiple target anchor frames to improve the accuracy of target recognition. YOLOv6 [27] is designed by the Meituan team based on the idea of RepVGG and uses the new SIoU Loss as the position loss function, which outperforms other algorithms in speed and accuracy. The main improvement of YOLOv7 [28] is the introduction of an efficient layer aggregation network into the backbone network; the neck network is mainly composed the of SPPCSPC module, MP module, ELAN-M module, and reparameterized module. While the accuracy of YOLOv7 is improved, the network structure has become more complex. The main improvements of YOLOv8 [29] are as follows: the backbone uses the concept of Cross Stage Partial (CSP) to split the feature map into two parts; the C2f module is used in YOLOv5 instead C3 module; and the neck uses multi-scale feature fusion of images, which can further improve the accuracy and the lightweight feature.

Based on the comprehensive consideration of the accuracy and real-time performance of all models of the YOLO series, YOLOv5 is a widely used model and is selected as the basic network for improvement. The network structure diagram is shown in Figure 1.

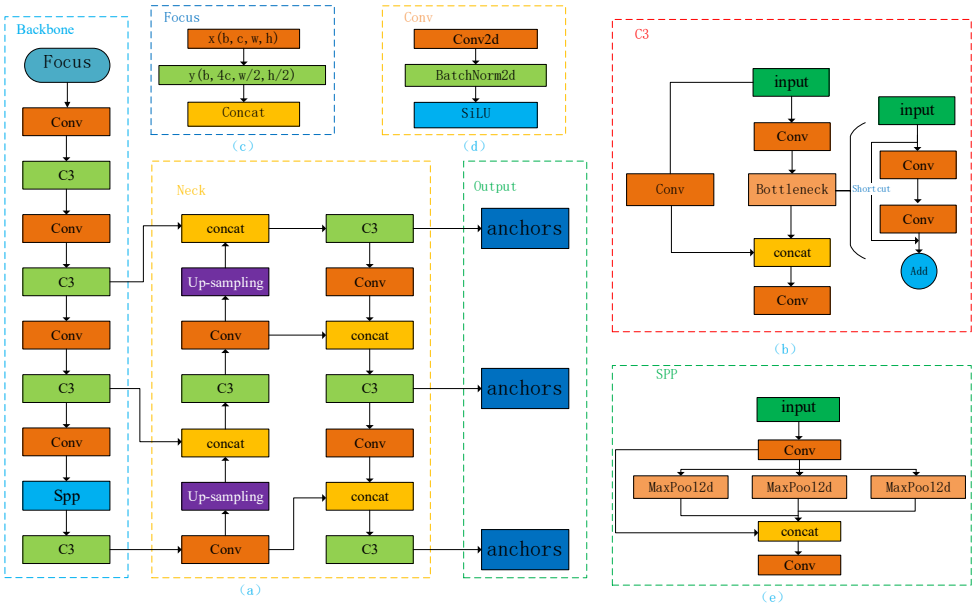

Figure 1. The network structure of YOLOv5. (**a**) The network structure of YOLOv5; (**b**) C3 module; (**c**) Focus module; (**d**) Conv module; (**e**) SPP module.

YOLOv5 has four versions: YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x. The main difference is that the depth multiple and width multiple of the model are different, as shown in Table 1.

Table 1. The depth multiple and width multiple of YOLOv5′s four versions.

Version	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
depth multiple (DM)	0.33	0.67	1.0	1.33
width multiple (WM)	0.50	0.75	1.0	1.25

In the YOLOv5 network, DM affects the number of network application bottleneck components, that is, the length of the network depth. The number of bottleneck components applied to different versions of the YOLOv5 network is different and needs to be determined according to the theoretical repetition number and DM of bottleneck components—that is, the actual repetition number = the theoretical repetition number of bottleneck × DM.

In the YOLOv5 network, WM affects the number of convolution kernels in the network convolution layer (Conv), that is, the number of channels. The number of convolution layer channels of different versions of the YOLOv5 network is different and needs to be determined according to the theoretical value and WM of the convolution kernel—that is, output channel = theoretical convolution kernel value × WM.

The actual running number of the bottleneck module in C3 and the actual channel number of convolutional layers in different versions of YOLOv5 are shown in Table 2. The number of C3 and Conv in the backbone network and neck network follows the principle of uniform number, first from left to right, and then from top to bottom. In C3-*n* and Conv-*m* in Table 2, *n* and *m* represent the theoretical repetitions of C3 and Conv, respectively.

Table 2. The actual running number of the bottleneck module in C3 and the actual channel number of convolutional layers of different versions of YOLOv5.

No.	Version	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x	Version	YOLOv5s	YOLOv5m	YOLOv5l	YOLOv5x
1st	C3-3	1	2	3	4	Focus (Conv-64)	32	48	64	80
2nd	C3-9	3	6	9	12	Conv-128	64	96	128	160
3rd	C3-9	3	6	9	12	Conv-256	128	192	256	320
4th	C3-3	1	2	3	4	Conv-512	256	384	512	640
5th	C3-3	1	2	3	4	Conv-1024	512	768	1024	1280
6th	C3-3	1	2	3	4	Conv-256	128	192	256	320
7th	C3-3	1	2	3	4	Conv-512	256	384	512	640
8th	C3-3	1	2	3	4	Conv-256	128	192	256	320

2.2. Lightweight Network

With the development of convolutional neural networks, neural networks have been widely used in industrial computer vision and have achieved success. However, due to the limitation of storage space and power consumption, convolutional neural networks still face great challenges. Lightweight networks are widely used in tasks such as computer vision because it reduces the number of parameters, model size, and calculation while maintaining the good performance of the network model. Common lightweight networks include MobileNetV3 [30], Ghostnet [31], and Shufflenetv2 [32].

2.2.1. MobileNetV3

MobileNetV3 is one of the lightweight networks of the MobileNet series, which has the best features.

First, MobileNetV3 uses the deep separable convolution in the MobileNetV1 network [33], which consists of a deep convolution and a pointwise convolution shown in Figure 2. Deep convolution is a single-channel calculation method—that is, the number of input feature maps, the number of convolution kernels, and the number of output feature maps are the same. Pointwise convolution uses a 1×1 convolution kernel to extract features for each element after the deep convolution. The relationship between the depthwise separable convolution and the ordinary convolution parameters is shown in Equation (1), and the computational relationship is shown in Equation (2):

$$\frac{D_k \times D_k \times 1 \times M + 1 \times 1 \times M \times N}{D_k \times D_k \times M \times N} = \frac{1}{N} + \frac{1}{D_k^2}, \tag{1}$$

$$\frac{D_k \times D_k \times 1 \times M \times D_F \times D_F + 1 \times 1 \times M \times N \times D_F \times D_F}{D_k \times D_k \times M \times N \times D_F \times D_F} = \frac{1}{N} + \frac{1}{D_k^2}, \tag{2}$$

From Equations (1) and (2), it can be seen that for the same image, the unique structure of the depthwise separable convolution makes the number of parameters and the amount of computation $1/N + 1/D_K^2$ times that of the ordinary convolution, so when N is larger, the amount of computation and parameters reduced by depthwise separable convolution will be more. However, it also loses key features while reducing the number of parameters and calculations. Therefore, MobileNetV3 uses the inverse residual structure in MobileNetV2 [34] and replaces the last ReLU6 activation function in each MobileNetV2-Block with a linear activation function, which can reduce the consumption of network features and make the subsequent network get more information.

Second, MobileNetV3 also introduces the attention mechanism SENet network and makes lightweight improvements to it. The attention mechanism is used to establish the relationship between channels to enhance the representation ability of the network. In order to improve the computational reasoning speed, the activation function in MobileNetV3

is replaced with the h-swish function. The final MobileNetV3 base network is shown in Figure 3.

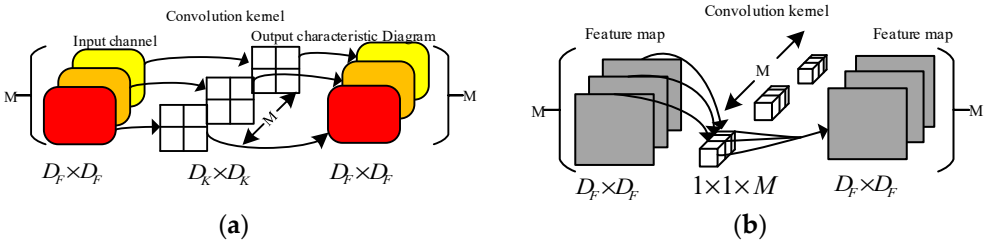

Figure 2. The deep separable convolution: (**a**) deep convolution; (**b**) pointwise convolution.

Based on the MobileNetV3-Block module, the MobileNetV3 network has two network structures, namely, the MobieNetV3-Large network structure and the MobileNetV3-Small network structure, which are shown in Tables 3 and 4. In the table, Bneck is used to represent the MobileNetV3-Block network structure, followed by the depth separable convolution size. SE represents whether the attention mechanism is introduced at this layer. In the column of SE, the √ sign represents SE, the-sign represents no SE. NL represents the type of nonlinear function used, where HS represents h-swish, RE represents ReLU, and S is the stride.

Table 3. MobieNetV3-Large network structure.

Layers	Operator	Exp Size	SE	NL	S	Layers	Operator	Exp Size	SE	NL	S
1	Conv2d	16	-	HS	2	11	Bneck, 3 × 3	184	-	HS	1
2	Bneck, 3 × 3	16	-	RE	1	12	Bneck, 3 × 3	480	√	HS	1
3	Bneck, 3 × 3	64	-	RE	2	13	Bneck, 3 × 3	672	√	HS	1
4	Bneck, 3 × 3	72	-	RE	1	14	Bneck, 5 × 5	672	√	HS	2
5	Bneck, 5 × 5	72	√	RE	2	15	Bneck, 5 × 5	960	√	HS	1
6	Bneck, 5 × 5	120	√	RE	1	16	Bneck, 5 × 5	960	√	HS	1
7	Bneck, 5 × 5	120	√	RE	1	17	conv2d, 1 × 1	-	-	HS	1
8	Bneck, 3 × 3	240	-	HS	2	18	Pool, 7 × 7	-	-	HS	1
9	Bneck, 3 × 3	200	-	HS	1	19	conv2d 1 × 1, NBN	-	-	-	1
10	Bneck, 3 × 3	184	-	RE	1	20	conv2d 1 × 1, NBN	-	-	HS	1

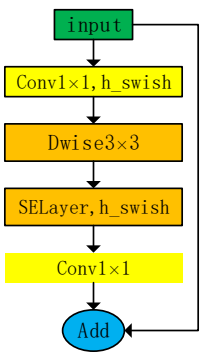

Figure 3. The network structure of MobileNetV3.

Table 4. MobieNetV3-Small network structure.

Layers	Operator	Exp Size	SE	NL	S	Layers	Operator	Exp Size	SE	NL	S
1	Conv2d, 3 × 3	16	-	HS	2	9	Bneck, 5 × 5	48	√	HS	1
2	Bneck, 3 × 3	16	√	RE	2	10	Bneck, 5 × 5	96	√	HS	2
3	Bneck, 3 × 3	24	-	RE	2	11	Bneck, 5 × 5	96	√	HS	1
4	Bneck, 3 × 3	24	-	RE	1	12	Bneck, 5 × 5	96	√	HS	1
5	Bneck, 5 × 5	40	√	HS	2	13	conv2d, 1 × 1	576	√	HS	1
6	Bneck, 5 × 5	40	√	HS	1	14	Pool, 7 × 7	-	-	HS	1
7	Bneck, 5 × 5	40	√	HS	1	15	conv2d 1 × 1, NBN	1024	-	HS	1
8	Bneck, 5 × 5	48	√	HS	1	16	conv2d 1 × 1, NBN	K	-	HS	1

2.2.2. ShuffleNetv2

In 2018, ShuffleNetv2 was a lightweight network proposed by the domestic Kuangshi technology team. There are three main improvements in ShuffleNetv2: (1) channel branch, which can replace the role of group convolution in ShuffleNetv1 [35] and reduce the amount of computation brought by group convolution; (2) deep convolution applied to greatly reduce the amount of computation; and (3) channel rearrangement, which can help the exchange of information between the branches and avoid the limitations of the branch structure. ShuffleNetv2 is divided into two basic units, as shown in Figure 4.

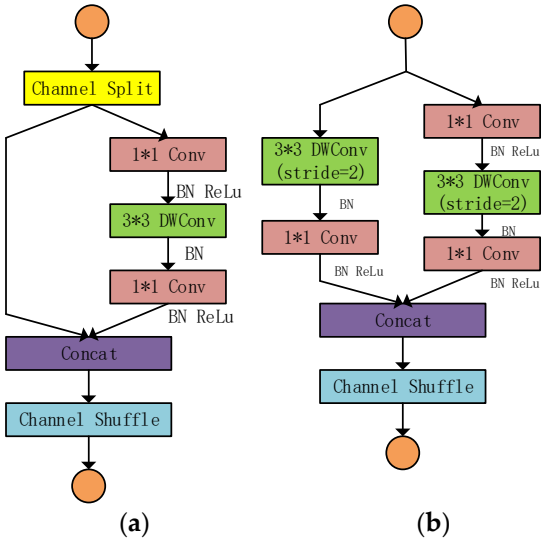

Figure 4. The basic unit of ShuffleNetv2 structure: (**a**) stride = 1; (**b**) stride = 2.

2.2.3. GhostNet

In 2020, Kai Han proposed the Ghostnet lightweight network. The Ghostnet network extracts more features with fewer parameters and effectively accepts redundant information in the network. The Ghostnet module turns the normal convolution operation into a two-step operation. The first step is the traditional convolution operation, but it reduces the application of the convolution kernel. The second step is a lightweight linear operation to generate redundant feature maps. The comparison between traditional convolution operation and Ghost module operation is shown in Figure 5.

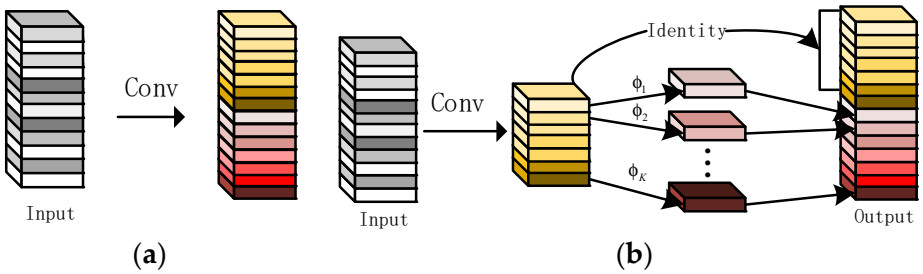

Figure 5. (**a**) Traditional convolution; (**b**) GhostConv module.

In Figure 5, when the size of the input feature map is $D_F \times D_F \times M$, the convolution kernel of the traditional convolution is $D_k \times D_k \times N$, and the computation amount is $D_k \times D_k \times M \times D_F \times D_F \times N$. The first step of the GhostConv module assumes that m feature maps are generated, and the computation amount is $D_k \times D_k \times M \times D_F \times D_F \times m$. In order to ensure the same size as the traditional convolution output, the second step of the GhostConv module is a lightweight linear operation on the feature map output by the first step, as shown in Equation (3).

$$y_{ij} = \phi_{ij}(y'_i), \forall_i = 1, \cdots, m; j = 1, \cdots, s, \tag{3}$$

where y'_i represents the ith feature map, y_{ij} denotes the jth feature map obtained by linear operation of the ith feature map, ϕ_{ij} is expressed as a linear operation. The GhostConv module can get N output feature maps, and $N = m \times s$. It can be seen from Figure 5b that only s-1 linear transformation that takes up computing resources is performed, so the computation amount of the GhostConv module is $D_K \times D_K \times M \times D_F \times D_F \times m + (s-1) \times D_K \times D_K \times D_F \times D_F$. Then the calculation relationship between the GhostConv module and traditional convolution is shown in Equation (4).

$$\frac{D_K \times D_K \times M \times D_F \times D_F \times N}{D_K \times D_K \times M \times D_F \times D_F \times m + (s-1) \times D_K \times D_K \times D_F \times D_F} \approx s, \tag{4}$$

According to Equation (4), the traditional convolution is s times as much as the GhostConv module in the calculation. Therefore, building a Ghostnet network based on the Ghostnet-Block network can greatly reduce the number of network parameters and the amount of computation. Using the idea of the Ghostnet network, the C3Ghost module is shown in Figure 6. Here, DWConv represents deep separable convolution, and stride is the length of the step.

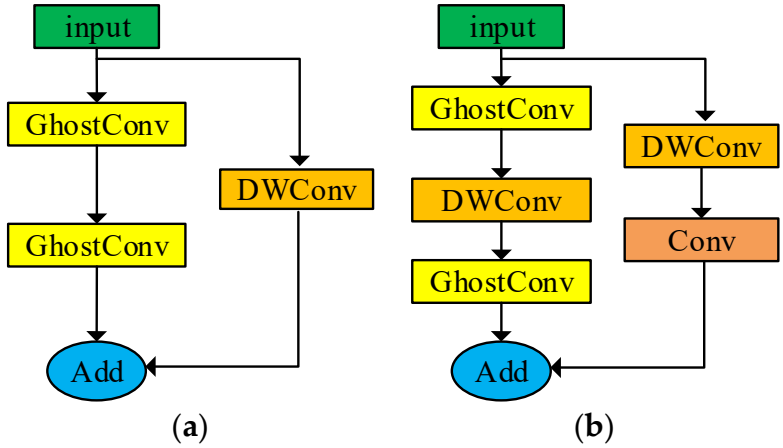

Figure 6. C3Ghost module: (**a**) stride = 1; (**b**) stride = 2.

3. Proposed Method

The YOLOv5 target detection algorithm and three lightweight networks are introduced in Sections 2.1 and 2.2. From Table 2, we can see that compared with the other three YOLOv5 networks, YOLOv5s has the smallest number of parameters and calculations, so YOLOv5s is selected for ear recognition.

When studying three lightweight networks, we have found that each network has its own advantages. The MobileNetV3 network uses a large number of deep separable convolutions to fuse independently calculated feature information, which can effectively reduce the model size. ShuffleNetv2 network uses channel rearrangement technology to extract feature information, which reduces the amount of model computation, the number of parameters, and the size of the model but sacrifices too much accuracy. The Ghostnet network halves the convolution layer and retains most of the feature information, but it is not enough to reduce the amount of computation, the number of parameters, and the size of the model. Therefore, in this paper we propose to use two lightweight networks, MobileNetV3 and Ghostnet, to make YOLOv5s lightweight and improve its real-time performance. The improved network model is named YOLOv5s-MG, and the structure diagram is shown in Figure 7.

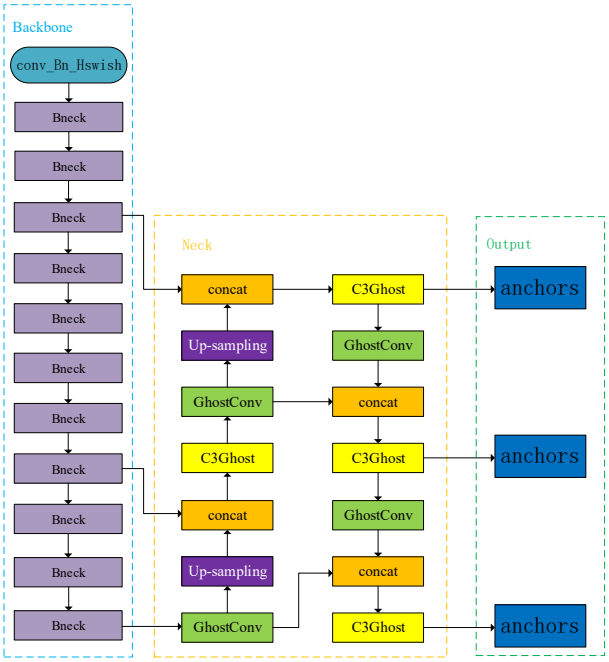

Figure 7. The framework of the proposed algorithm YOLOv5s-MG.

3.1. Lightweight of YOLOv5s Backbone Network Based on MobileNetV3

From Tables 3 and 4, it can be seen that the MobileNetV3-Small network has fewer layers than the MobileNetV3-Large network, with fewer convolution kernels applied, so the number of parameters and calculations are also reduced. Therefore, in this paper, we use the MobileNetV3-Small network to lightweight the YOLOv5s backbone network, as shown in the blue dotted box in Figure 7, which can reduce the computation amount of YOLOv5s backbone network feature extraction. The specific steps are as shown in Step 1:

Step1: Lightweight of YOLOv5s backbone network based on MobileNetV3

Step1.1: The feature extraction network from the first layer to the twelfth layer of the MobileNetV3-Small network is used to replace the YOLOv5s backbone network to extract the features of human ear images, and the depth separable convolution size in the backbone network of MobileNetV3-Small is kept unchanged;

Step1.2: The MobileNetV3-Block module of layer4, layer9, and layer12 in the backbone network of MobileNetV3-Small is used as the input of the neck layer of YOLOv5s for feature fusion. This can meet the needs of YOLOv5s's three output target frame sizes.

3.2. Lightweight of YOLOv5s Neck Network Based on Ghostnet

It can be seen from Section 2.2.3 that the traditional convolution is s times that of the GhostConv module. When the s value is larger, the computation amount of Ghostnet will be less. Therefore, the idea of the Ghostnet network is used to lightweight the YOLOv5s Neck network, as shown in the orange dotted box in Figure 7, which can reduce the computation amount of YOLOv5s neck network feature fusion. The specific steps are shown in Step 2:

Step2: Lightweight of YOLOv5s neck network based on Ghostnet

Step2.1: Replace the C3 module in the YOLOv5s neck network with the C3Ghost module;

Step2.2: Replace the Conv module in the YOLOv5s neck network with the GhostConv module.

Previously, only the lightweight network was used to improve the C3 or Conv of the backbone network of YOLOv5. In this paper, in addition to improving the backbone network, the neck network is also improved, and two lightweight networks MobileNetV3 and GhostNet are also adopted, which further improves the real-time performance of the network from feature extraction and feature fusion.

4. Experimental Results

4.1. Human Ear Datasets

In order to verify the effectiveness of the proposed YOLOv5s-MG lightweight human ear recognition method, the human ear dataset is needed to train and detect the model. In this paper, we use three distinctive human ear datasets: CCU-DE, USTB, and EarVN1.0.

4.1.1. CCU-DE

CCU-DE [36] is a dynamic and small sample human ear dataset established by the Human Ear Laboratory of Changchun University. The database takes into account all possible situations and influencing factors of human ears in practice, such as illumination, angle, movement, occlusion, etc., including pictures and videos. There are five human ear databases in total, as shown in Table 5.

Table 5. Human ear database information of CCU-DE.

Database	Shooting Situation
Eardata1	Static human ear fixed-point shooting
Eardata2	Video of human ears moving in translation with the human body when the human body is walking normally at different angles
Eardata3	Video of the photographed standing in the shooting center doing a 90° uniform rotation motion
Eardata4	Video of the photographed standing in the shooting center doing a 180° uniform rotation motion
Eardata5	Dynamic human ear video with interference information

In order to diversify the dataset and better verify the effectiveness of the ear recognition algorithm proposed in this paper, we selected Eardata3 and Eardata4 ear video samples for experiments. We selected No. 25 and No. 26 in Eardata3 and No. 1, No. 27 to No. 33 ear videos in Eardata4.

In the training of the human ear recognition model, feature extraction and training are carried out on one picture after another. Therefore, dynamic video is captured in this paper at the rate of 2 frames per second, and the CCU-DE human ear dataset is selected as shown in Table 6. There are 3274 pictures. According to the ratio of training set: test set:

verification set = 3:1:1, the training set has 1964 pictures, and the test set and verification set each have 655 pictures, respectively.

Table 6. The selection of CCU-DE human ear dataset.

Database	Category	Left Ear (NF)	Right Ear (NF)	Total (NF)	Size of the Picture
Eardata4	ear1	153	135	288	1280 × 720
Eardata3	ear25	111	139	250	
Eardata3	ear26	112	154	266	
Eardata4	ear27	249	156	405	
Eardata4	ear28	96	106	202	
Eardata4	ear29	163	188	351	
Eardata4	ear30	156	124	280	
Eardata4	ear31	203	186	389	
Eardata4	ear32	175	182	357	
Eardata4	ear33	235	251	486	
total		1653	1621	3274	

4.1.2. USTB

USTB [37] is a human ear dataset established by the Human Ear Laboratory of the University of Science and Technology, Beijing. There are three databases in USTB, namely, database1, database2, and database3. Because database3 has more variety of categories and postures than database1 and database2, it is selected to verify the ear recognition algorithm in this paper. There are 79 categories in database3, and 70 categories are selected in this paper, namely, No. 1 to No. 70. In order to make the human ear data better simulate the real situation, the data of the human ear pictures in the database was enhanced, including the changes of angle, saturation, color, and brightness. The sample dataset is shown in Table 7. There are 7700 pictures. According to the ratio of training set: test set: verification set = 3:1:1, the training set has 4620 pictures, and the test set and verification set each have 1540 pictures, respectively.

Table 7. The sample dataset of database3 in USTB.

Category	Attitude Change									Size
70	0°	5°	−5°	10°	−10°	20°	−20°	flip left and right	enhancement: contrast brightness color	768 × 576

4.1.3. EarVN1.0

EarVN1.0 [38] is a large-scale ear image dataset established by the Faculty of Computer Science of Ho Chi Minh City Open University in Vietnam in 2018. This is one of the largest public ear datasets in the research field, containing 28,412 color images of 164 participants, of which 98 men provided 17,571 male ear images and 66 women provided 10,841 female ear images.

This dataset is different from the previous dataset. The original photo was taken under the condition that the camera system, illumination conditions were not limited, and the human ear image was cropped under the condition that the original image had great changes in posture, scale, and illumination. Each subject provided at least about 100 left or right ear images. The resolution of human ear images is also uneven, and some images have very low resolution (less than 25 × 25 pixels).

In this paper, we selected the first 15 kinds of human ear samples from the dataset of EarVN1.0 for experiments. There are 3201 photos. According to the ratio of training set: test set: verification set = 3:1:1, the training set has 1921 pictures, and the test set and verification set each have 640 pictures, respectively.

From Sections 4.1.1–4.1.3, we can see that the three ear datasets in this experiment have their own characteristics, as shown in Table 8.

Table 8. Comparison of three ear datasets.

Human Ear Dataset	Data Size	Attitude Change	Resolution	Category	Status	Gender
CCU-DE	3274	richest	highest	10	dynamic	female and male
USTB	7700	richer	higher	70	static	female and male
EarVN1.0	3201	richest	uneven	15	static	female and male

The above three datasets have different characteristics in this paper. Compared with the other two ear datasets, the CCU-DE ear dataset is dynamic, with the least number of categories, the smallest amount of data, rich attitude changes, and the highest image resolution. Compared with the other two ear datasets, the USTB ear dataset is static, with the largest number of categories, the largest amount of data, rich attitude changes, and medium image resolution. Compared with the other two ear datasets, the EarVN1.0 ear dataset is static, with fewer categories and fewer data, but with rich attitude changes, uneven image resolution, and poor parts.

4.2. Experimental Setting

In order to verify the effectiveness and feasibility of the proposed algorithm, two sets of experiments are set up. The first group compares four YOLOv5 models in three ear datasets. The second group compares the improved method YOLOv5s-MG in this paper with other methods, such as YOLOv5s-V3 (Lightweight YOLOv5s backbone network using MobileNetV3 [39]), YOLOv5s-V2 (Lightweight YOLOv5s backbone network using ShuffleNetv2 [40]), YOLOv5s-G (The C3 and Conv modules of the backbone and neck networks of YOLOv5s were replaced by the C3Ghost module and the GhostConv module [41]), YOLOv7, and YOLOv5s.

The experimental platform used in the experiment is shown in Table 9, and the parameter settings of network training are shown in Table 10.

Table 9. The experimental platform.

Name	Configuration
CPU	IntelI CoreIi5-10400F CPU@2.90GHZ
Memory	16 GB
GPU	NVIDIA GeFORCE RTX 3060
GPU-accelerated library	CUDA11.1.134, CUDNN8.0.5
Operating system	Windows10(64bit)
Software environment	pytorch1.8, python3.8
Dataset partitioning	Training set:validation set:test set = 3:1:1

Table 10. The parameter settings of network training.

Name	Configuration
Initial learning rate	0.01
Learning rate reduction coefficient	0.2
Weight attenuation coefficient	0.0005
Momentum coefficient	0.937
Optimizer	SGD with momentum
Batch size	16

4.3. Evaluation Indicators

In order to evaluate the performance of the ear recognition algorithm proposed in this paper, evaluation indicators are needed. The evaluation indexes used in this experiment

mainly include mean average precision (mAP), model size (/MB), computation amount (GFLOPS/G), and model parameter number (params/M).

4.3.1. mAP

1. Classification target division result;

The test result will divide the classification target into true positives (TP), false positives (FP), false negatives (FN), and true negatives (TN). TP is the number of true targets that are considered correct targets. FP is the number of false targets considered correct targets. FN is the number of true targets considered false targets. FP is the number of false targets considered false targets.

2. Precision and Recall;

Precision and Recall are defined according to the target classification results, as shown in Equations (5) and (6), respectively. From Equation (5), we can see that precision is the proportion of samples that the classifier considers to be positive and is indeed positive to all the samples that the classifier considers to be positive. Recall is the proportion of all actual positive samples considered to be positive from Equation (6).

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP}), \tag{5}$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN}), \tag{6}$$

3. Average precision (AP) and mean average precision (mAP);

AP is single-class average precision as shown in Equation (7). mAP is the average accuracy of all test images for all categories as shown in Equation (8). In Equation (7), N is the total number of single-class targets. In Equation (8), C is the total category of all detected objects.

$$AP_i = \sum_1^N Precision/N, \tag{7}$$

$$mAP = \sum_1^C AP_i/C, \tag{8}$$

At present, there are two commonly used mAPs: mAP@0.5 and mAP@0.5:0.95. mAP@0.5 is the AP when the threshold of intersection and union (IOU) of each category is set to 0.5, and then mAP values are calculated for all categories. mAP@0.5:0.95 is the AP when the threshold range of IOU of each category is set to 0.5–0.95 and the step size is 0.5, and then the mAP values are calculated for all categories. mAP@0.5 is chosen in this experiment.

4.3.2. Model Parameter Quantity (Params/M)

The model parameters are the sum of all the parameters in the model. Convolutional neural networks are mainly composed of the weights of the convolution layer and the fully connected layer. Therefore, the parameters are mainly related to the network structure of the model. The more complex the model is, the deeper the number of network layers is and the greater the parameters of the model are. Generally, for a convolution kernel with N input channels and M output channels and the size of $H \times W$, its parameter calculation is shown in Equation (9):

$$\text{params} = (N \times H \times W + 1) \times M, \tag{9}$$

The unit of parameter quantity is generally expressed in millions (M). Therefore, the number of parameters does not directly affect the reasoning performance of the model. However, the size of the parameters will affect both memory occupation and program initialization time.

4.3.3. Amount of Calculation (GFLOPS/G)

The amount of calculation is a standard to measure the complexity of the algorithm, which is usually expressed by 1 billion floating-point operations per second (GFLOPS). 1GFLOPS = 10^9 FLOPs. FLOPs stands for floating-point operations per second.

4.3.4. Model Size (MB)

Model size refers to the memory size of the trained model file, and the unit is MB. The model size is the same as the model parameter quantity, which will not affect the performance of the model. However, when there are multiple concurrent tasks on the same platform, such as the reasoning server, vehicle platform, and mobile APP, the memory occupation is often required to be controllable.

The above four indicators are mainly evaluated from the accuracy and complexity of the model, which reflects the effectiveness and real-time performance of the model.

4.4. Ear Recognition Experiments of Four YOLOv5 Models on Three Datasets

In this experiment, four YOLOv5 models—YOLOv5s, YOLOv5m, YOLOv5l, and YOLOv5x—were used to conduct experiments on CCU-DE, USTB, and EarVN1.0 human ear datasets. Figure 8 is the training curve of different versions of the YOLOv5 network model using the training set and validation set in three datasets, the abscissa is the training epoch, and the ordinate is the accuracy (mAP@0.5). It can be seen from Figure 8 that as the number of epochs gradually increases, the value of mAP@0.5 gradually tends to be stable, that is, the model converges. However, for different human ear datasets, the epoch of convergence is different because of the difference in the pose, resolution, image size, and number. The epoch of model convergence on the CCU-DE human ear dataset is about 40, the epoch of model convergence on the USTB human ear dataset is about 50, and the epoch of model convergence on the EarVN1.0 human ear dataset is about 150. In this experiment, epoch 150 was selected as a quantitative to verify the difference between different versions of YOLOv5.

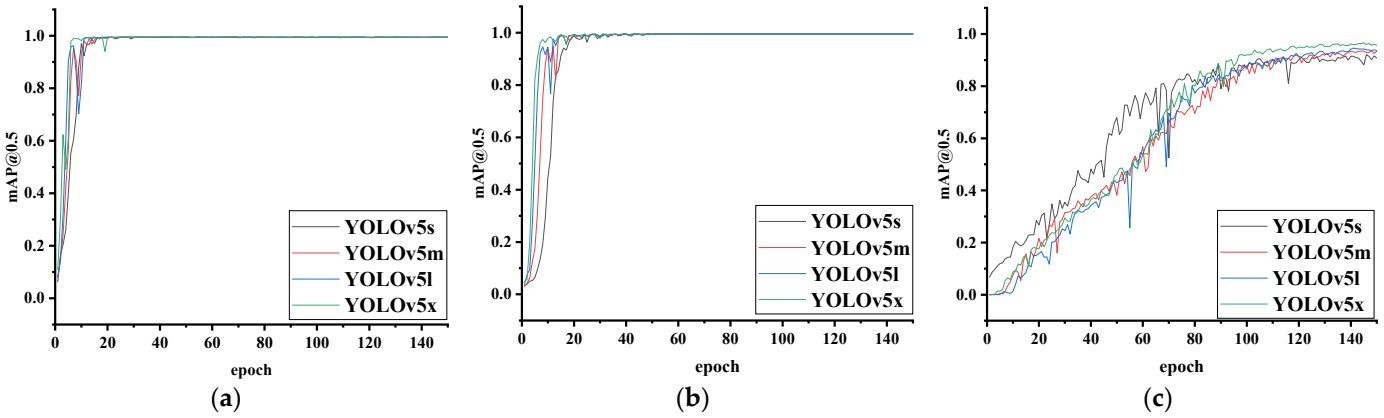

Figure 8. The mAP@0.5 value of four YOLOv5 models trained on three human ear datasets: (**a**) CCU-DE ear dataset; (**b**) USTB ear dataset; (**c**) EarVN1.0 ear dataset.

When the epoch value is 150, the experimental results of the four evaluation indicators on the testing set of the three human ear datasets are shown in Table 11 and Figure 9. Figure 9a is the accuracy (mAP@0.5) comparison result, Figure 9b is the parameter quantity (params) comparison result, Figure 9c is the calculation amount (GFLOPS) comparison result, and Figure 9d is the model size comparison result.

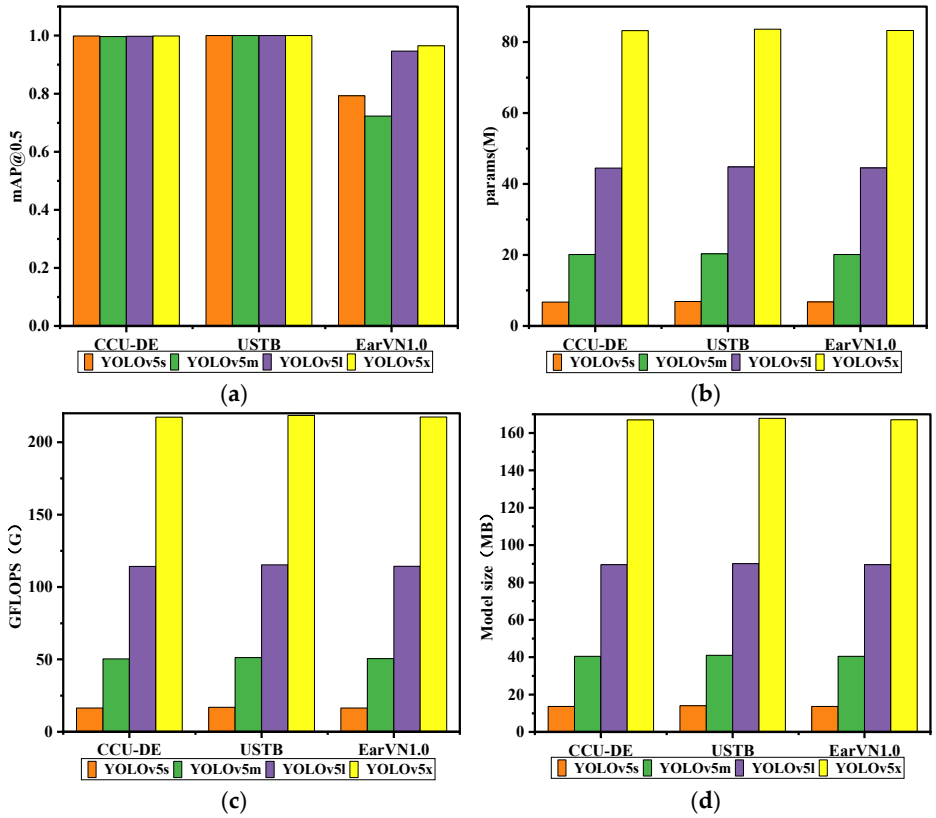

Figure 9. The comparison results of four evaluation indexes of five methods on three datasets: (**a**) mAP@0.5; (**b**) the parameter quantity (params); (**c**) the computation amount (GFLOPS); and (**d**) the model size.

Table 11. Quantitative comparison results of YOLOv5s, YOLOv5m, YOLOv5m, and YOLOv5x.

Human Ear Dataset	Model	Params (M)	GFLOPS(G)	Model Size (MB)	mAP@0.5
CCU-DE (epoch = 150)	YOLOv5s	6.75	16.4	13.7	0.999
	YOLOv5m	20.1	50.4	40.55	0.997
	YOLOv5l	44.49	114.2	89.5	0.998
	YOLOv5x	83.22	217.3	167	0.999
USTB (epoch = 150)	YOLOv5s	6.9	16.9	14.07	1
	YOLOv5m	20.32	51.2	41.03	1
	YOLOv5l	44.79	115.3	90.09	1
	YOLOv5x	83.60	218.6	167.85	1
EarVN1.0 (epoch = 150)	YOLOv5s	6.76	16.4	13.7	0.793
	YOLOv5m	20.12	50.5	40.5	0.723
	YOLOv5l	44.51	114.3	89.5	0.947
	YOLOv5x	83.25	217.4	167.1	0.965

From Table 11 and Figure 9a, it can be seen that on the USTB ear dataset, the four YOLOv5 models have the highest ear recognition rate: the mAP@0.5 value is 1, CCU-DE follows, and mAP@0.5 is above 0.99; on the EarVN1.0 dataset, the mAP@0.5 values of the four YOLOv5 models vary greatly, with the highest value above 0.9 and the lowest value above 0.7.

The difference in mAP@0.5 values is mainly related to three factors. First, related to the model, the YOLOv5x model has the largest depth and width, and the strongest feature extraction ability, so the mAP@0.5 value is the largest. The YOLOv5s model has the smallest depth and width, and the weakest feature extraction ability, so the mAP@0.5 value is the smallest. Second, it is related to whether the model converges. It can be seen from Figure 8 that the four models on the CCU-DE and USTB human ear datasets are completely convergent, so the accuracy of human ear recognition is high, that is, the mAP@0.5 value is large. On the EarVN1.0 dataset, when the epoch is 150, the model begins to converge. Because the model has not yet fully converged, the mAP@0.5 value will be small. Third, it is related to the human ear dataset. The human ear dataset with high resolution, a small amount of data, and relatively simple posture has a high ear recognition rate, that is, the mAP@0.5 value is large on the USTB human ear dataset, and the mAP@0.5 value on the EarVN1.0 dataset is small.

From Table 11 and Figure 9b–d, it can be seen that the params, GFLOPS, and model size of the four models on the three human ear datasets increase in turn. The params, GFLOPS, and model size of YOLOv5s are the smallest, but the params, GFLOPS, and model size of YOLOv5x are the largest. The params, GFLOPS, and model size of YOLOv5s are 8.1%, 7.5%, and 8.2% of YOLOv5xs, respectively, on CCU-DE and EarVN1.0, and the params, GFLOPS, and model size of YOLOv5s are 8.3%, 7.7%, and 8.4% of YOLOv5xs, respectively, on USTB.

The difference in params, GFLOPS, and model size is mainly related to two factors. First, it is related to network depth and width. The deeper and wider the network, the more complex the network, so the params and GFLOPS are greater and the model is larger. Second, it is related to the number of samples in the human ear dataset. From Table 8, we can see that the number of samples of CCU-DE and EarVN1.0 is 3274 and 3201, so the params, GFLOPS, and model size are approximately equal. However, the number of samples of USTB is 7700, so the params, GFLOPS, and model size increase slightly.

Through experimental and theoretical analysis, it can be seen that the four YOLOv5 models can be used for ear recognition, but when the mAP@0.5 value is the same, the params, GFLOPS, and model size of the YOLOv5s network are the smallest, so the YOLOv5s network is selected for ear recognition.

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

In order to verify the feasibility and effectiveness of the proposed lightweight network YOLOv5s-MG in this paper and to verify the similarities and differences between the improved lightweight network, the original YOLOv5s model, and the YOLOv7 model, we trained and tested YOLOv5s-MG, YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, YOLOv7, and YOLOv5s networks with the same parameters on CCU-DE, USTB, and EarVN1.0 human ear datasets. The experimental platform and experimental parameters are shown in Tables 9 and 10.

Figure 10 shows the training curves of YOLOv5s-MG, YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, YOLOv7, and YOLOv5s networks using the training set and validation set on three datasets. The abscissa is the training epoch, and the ordinate is the accuracy (mAP@0.5). It can be seen from Figure 10 that as the number of epochs gradually increases, the value of mAP@0.5 gradually tends to be stable, that is, the model converges. However, for different human ear datasets, the epoch of convergence is different because of the difference in the pose, resolution, image size, and number. The epoch of model convergence on the CCU-DE ear dataset is about 40, as shown in Figure 10a. The epoch of model convergence on the USTB ear dataset is about 50, as shown in Figure 10b. The convergence time epoch of the model on the EarVN1.0 ear dataset is about 500, as shown in Figure 10d. Figure 10c shows the case of the model not converging when epoch = 150 on the EarVN1.0 ear dataset. Among the four lightweight models, the YOLOv5s-MG model converges faster and the curve is smooth, while YOLOv5s-V2 converges slower. Among the six methods, the YOLOv7 model has the slowest convergence speed and more oscillations during training.

In this experiment, epoch = 150 and epoch = 1000 were selected as quantities to verify the difference between the four lightweight YOLOv5s models and the YOLOv5s model. The experimental results on the testing set of the three human ear datasets are shown in Table 12 and Figure 11, respectively.

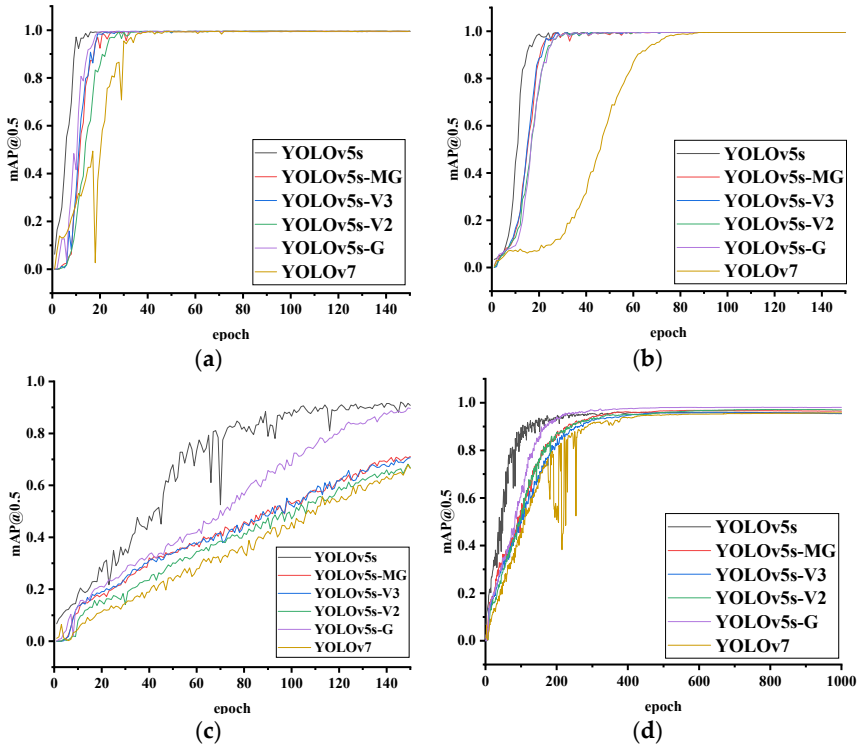

Figure 10. The mAP@0.5 value of YOLOv5s-MG (proposed), YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, YOLOv7, and YOLOv5s models trained on three human ear datasets: (**a**) CCU-DE ear dataset; (**b**) USTB ear dataset; (**c**) EarVN1.0 ear dataset (no convergence); (**d**) EarVN1.0 ear dataset (convergence).

From Table 12, we can see that the difference in mAP@0.5 values is mainly related to three factors. First, it is related to whether the model converges. It can be seen from Figure 10 that the six models on the CCU-DE and USTB human ear datasets are completely convergent in the case of 150 epochs, so the accuracy of human ear recognition is high, that is, the mAP@0.5 value is large. However, on the EarVN1.0 dataset, when the epoch is 150, the model begins to converge. Because the models have not yet fully converged, the mAP@0.5 value will be small. Compared with epoch = 150 and epoch = 1000 on the EarVN1.0 ear dataset in Figures 10c,d and 11a, and Table 12, mAP@0.5 of all methods increased, and especially in the four lightweight methods, mAP@0.5 increased greatly. mAP@0.5 of the YOLOv5s-V3 method increased by 42.2% from 0.389 to 0.811. mAP@0.5 of the YOLOv5s-V2 method increased by 49.6% from 0.322 to 0.818. mAP@0.5 of the YOLOv5s-G method increased by 34.2% from 0.543 to 0.885. mAP@0.5 of the YOLOv5s-MG (the proposed) method increased by 49.9% from 0.356 to 0.855. mAP@0.5 of the YOLOv5s method increased by 8.9% from 0.793 to 0.882. Second, it is related to the human ear dataset. From Table 8, the three ear datasets of CCU-DE, USTB, and EarVN1.0 have their own characteristics. From the experimental results shown in Figures 10d and 11a, and Table 12, the six networks except YOLOv7 have the highest ear recognition rate on the USTB ear dataset, and the mAP@0.5 value is 1. CCU-DE follows, and mAP@0.5 is above 0.99. On the EarVN1.0 dataset, the human ear recognition rate is the lowest, and the mAP@0.5 value is the smallest. When epoch = 1000, mAP@0.5 is above 0.81. It is proved theoretically and experimentally that the human ear dataset with high resolution, a small amount of data, and relatively simple posture has a high ear recognition rate, that is, the mAP@0.5 value is large on the CCU-DE and USTB human ear datasets, and the mAP@0.5 value on the

EarVN1.0 dataset is small. Third, it is related to the model. This could be expected as with the complexity reduction of the method, there is a decrease in precision. From Table 12 and Figure 11b–d, it can be seen that the params, GFLOPS, and model size of the improved YOLOv5s-MG model are much smaller than those of the other three lightweight models and YOLOv5s models, but the proposed model has slightly lower mAP@0.5 performance compared with the other methods for the observed dataset.

Table 12. Quantitative comparison results of the improved YOLOv5s-MG and other methods.

Human Ear Dataset	Model	Params (M)	GFLOPS (G)	Model Size (MB)	mAP@0.5
CCU-DE (epoch = 150)	YOLOv5s	6.75	16.4	13.7	0.999
	YOLOv5s-V3 [39]	3.39	6.3	7.04	0.998
	YOLOv5s-V2 [40]	3.63	8	7.56	0.999
	YOLOv5s-G [41]	3.2	8.9	6.8	0.999
	YOLOv5s-MG	2.05	3.7	4.3	0.997
	YOLOv7	34.84	103.5	71.41	0.993
USTB (epoch = 150)	YOLOv5s	6.9	16.9	14.07	1
	YOLOv5s-V3 [39]	3.55	6.8	7.37	1
	YOLOv5s-V2 [40]	3.79	8.5	7.88	1
	YOLOv5s-G [41]	3.4	9.4	7.14	1
	YOLOv5s-MG	2.2	4.2	4.6	1
	YOLOv7	35.14	104.5	72.04	0.904
EarVN1.0 (epoch = 150)	YOLOv5s	6.76	16.4	13.7	0.793
	YOLOv5s-V3 [39]	3.4	6.4	7.05	0.389
	YOLOv5s-V2 [40]	3.6	8	7.57	0.322
	YOLOv5s-G [41]	3.2	8.9	6.82	0.543
	YOLOv5s-MG	2.06	3.7	4.34	0.356
	YOLOv7	34.86	103.6	71.45	0.303
EarVN1.0 (epoch = 1000)	YOLOv5s	6.76	16.4	13.7	0.882
	YOLOv5s-V3 [39]	3.4	6.4	7.05	0.811
	YOLOv5s-V2 [40]	3.6	8	7.57	0.818
	YOLOv5s-G [41]	3.2	8.9	6.82	0.885
	YOLOv5s-MG	2.06	3.7	4.34	0.855
	YOLOv7	34.86	103.6	71.45	0.869

On the CCU-DE ear dataset, the params, GFLOPS, and model size of the improved YOLOv5s-MG model are 30.37%, 22.56%, and 31.39% of YOLOv5s, respectively. On the USTB human ear dataset, the params, GFLOPS, and model size of the improved YOLOv5s-MG model are 31.88%, 24.85%, and 32.69% of YOLOv5s, respectively. On the EarVN1.0 ear dataset, the params, GFLOPS, and model size of the improved YOLOv5s-MG model are 30.47%, 22.56%, and 31.68% of YOLOv5s, respectively. From Table 12 and Figure 11, we can also see that the params, GFLOPS, and model size of YOLOv5s MG are much smaller than those of YOLOv7.

Therefore, the experimental results show that in the three human ear datasets, the YOLOv5s-MG network is superior to the YOLOv5s-V3, YOLOv5s-V2, and YOLOv5s-G lightweight networks in reducing the number of parameters, computation, and model size on the basis of maintaining the recognition accuracy at the same level.

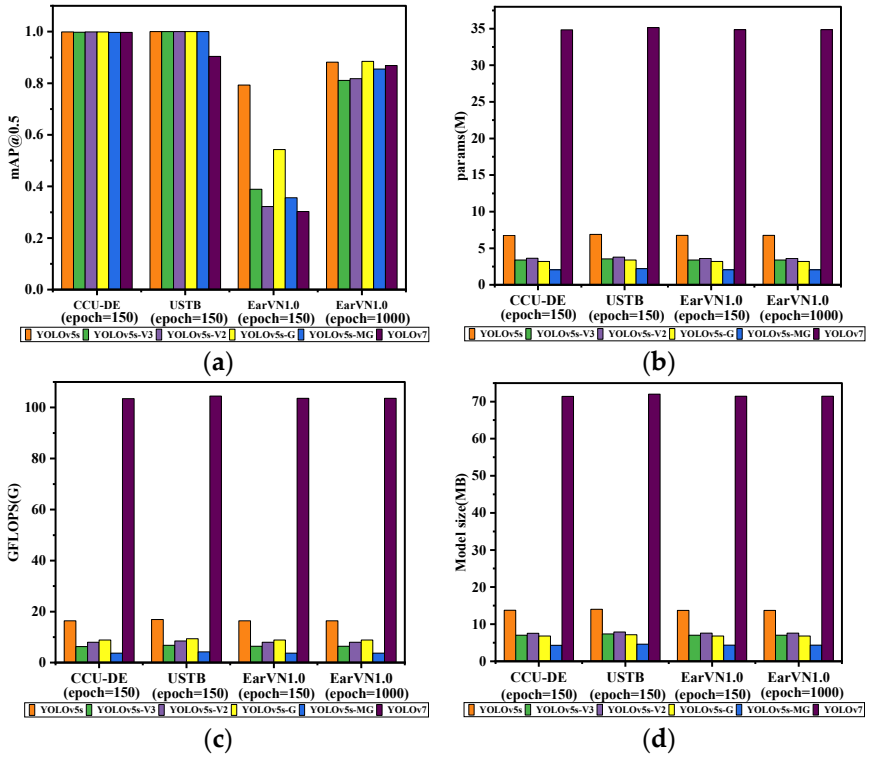

Figure 11. The comparison results of YOLOv5s-MG (proposed), YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, YOLOv7, and YOLOv5s models on three datasets: (**a**) mAP@0.5; (**b**) the parameter quantity (params); (**c**) the calculation amount (GFLOPS); and (**d**) the model size.

4.6. The Computational Complexity Analysis

In order to analyze the computational complexity, we calculate the number of ear pictures detected per second when the batch size is 1 (namely, FPS) and the calculation formula is shown in Equation (10).

$$\text{FPS} = 1/\text{average inference time per image}, \tag{10}$$

All experiments were performed under the same conditions and the FPS of all methods are shown in Table 13. The best and second-best running times are displayed in red and blue, respectively.

Table 13. FPS of the six methods on the testing set of three datasets.

Human Ear Dataset	Model					
	YOLOv5s	**YOLOv5s-V3**	**YOLOv5s-V2**	**YOLOv5s-G**	**YOLOv5s-MG**	**YOLOv7**
CCU-DE (epoch = 150)	6.7	12	10.5	8	14.5	1
USTB (epoch = 150)	4.6	7.5	6.2	5.5	8.5	1.3
EarVN1.0 (epoch = 150)	5.7	9.6	8.1	6.7	11.2	1.1
EarVN1.0 (epoch = 1000)	5.8	9.7	8.3	6.8	11.4	1.3

From Table 13, we can see that the proposed YOLOv5s-MG method has the highest detection speed. On the CCU-DE ear dataset, FPS of the improved YOLOv5s-MG model increased by 116.4%, 20.8%, 38.1%, 81.3%, and 1350% compared with that of YOLOv5s,

YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, and YOLOv7, respectively. On the USTB ear dataset, FPS of the improved YOLOv5s-MG model increased by 84.8%, 13.3%, 37.1%, 54.5%, and 553.8% compared with that of YOLOv5s, YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, and YOLOv7 respectively. On the USTB ear dataset when epoch = 1000, FPS of the improved YOLOv5s-MG model increased by 96.6%, 17.5%, 37.1%, 37.3%, and 776.9% compared with that of YOLOv5s, YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, and YOLOv7, respectively.

5. Conclusions

In this paper, we propose a new lightweight ear recognition method named YOLOv5s-MG, which uses the lightweight network MobileNetV3 and Ghostnet to lightweight YOLOv5s. Due to the improved YOLOv5s-MG method, the feature extraction of the backbone network of YOLOv5s and the feature fusion of the neck network of YOLOv5s are lightweight at the same time, which greatly reduces the number of parameters, calculations, and model size of the network, which can greatly improve the real-time performance of the method. The proposed YOLOv5s-MG method is tested on the testing set of three distinctive human ear datasets—CCU-DE, USTB, and EarVN1.0—and evaluated by four performance indicators: mAP@0.5, params, GFLOPS, and model size. Quantitative results show that the proposed method is superior to the other five methods in terms of real-time performance. Compared with YOLOv5s, the params, GFLOPS, and model size of this method on the CCU-DE dataset are increased by 69.63%, 77.44%, and 68.61%, respectively; the same three parameters on the USTB dataset are increased by 68.12%, 75.15%, and 67.31%, respectively; the same three parameters on EarVN1.0 dataset are increased by 69.53%, 77.44%, and 68.32%, respectively. Compared with the best results of the YOLOv5s-V3, YOLOv5s-V2, and YOLOv5s-G methods, the params, GFLOPS, and model size of this method on the CCU-DE dataset are increased by 35.94%, 41.27%, and 36.76%, respectively; the same three parameters on the USTB dataset are increased by 35.29%, 38.24%, and 35.57%, respectively; the same three parameters on EarVN1.0 dataset are increased by 35.63%, 42.19%, and 36.36%, respectively. For the recognition accuracy, that is, the mAP@0.5 value, there are different performances for different human ear datasets. In the case of complete convergence of the model, the recognition rate of the six methods is above 99% on the testing set of CCU-DE, but the mAP@0.5 value of the YOLOv5s-MG method is 0.2% lower than that of YOLOv5s; on the testing set of USTB, the recognition rate of the six methods except YOLOv7 is 100%; on the testing set of EarVN1.0, the ear recognition rate of the six methods is above 81% and the mAP@0.5 value of the YOLOv5s-MG method is 85.5%, which is 2.7% lower than that of YOLOv5s and 3% lower than that of the YOLOv5s-G method. For FPS, the proposed method YOLOv5s-MG in this paper is superior to the other five methods. The qualitative results demonstrate that YOLOv5s-MG has the same or slightly decreased ear recognition accuracy as YOLOv5s, YOLOv5s-V3, YOLOv5s-V2, YOLOv5s-G, and YOLOv7 methods, but the params, GFLOPS, and model size of this method are much smaller than those of other methods. The improved YOLOv5s-MG lightweight network proposed in this paper has good real-time performance and can meet the real-time requirements in the actual human ear recognition application system. However, for human ear datasets, such as EarVN1.0, when the real-time performance is improved, the human ear recognition rate may decrease. The next step will be to further study it to improve the human ear recognition rate.

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.; 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 in part by the funds of the Science Technology Department of Jilin Province [NO: 2020LY402L10] and in part by the funds of the Education Department of Jilin Province [NO: 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. Jain, A.K.; Pankanti, S.; Prabhakar, S.; Lin, H.; Ross, A. Biometrics: A grand challenge. In Proceedings of the International Conference on Pattern Recognition, Cambridge, UK, 26 August 2004; Volume 2, pp. 935–942.
2. Wang, S.N. Research on Ear Recognition Based on Deep Learning. Master's Thesis, University of Science and Technology Liaoning, Anshan, China, 2018.
3. Ding, Y.M. Research on Ear Recognition Based on Improved Sparse Representation. Master's Thesis, Harbin University of Science and Technology, Harbin, China, 2020.
4. Zhang, C.P.; Su, G.D. Summary of Face Recognition Technology. *J. Image Graph.* **2000**, *11*, 7–16.
5. Sakthimohan, M.; Rani, G.E.; Navaneethkrihnan, M.; Janani, K.; Nithva, V.; Pranav, R. Detection and Recognition of Face Using Deep Learning. In Proceedings of the 2023 International Conference on Intelligent Systems for Communication, IoT and Security (ICISCoIS), Coimbatore, India, 9–11 February 2023; pp. 72–76.
6. Ji, S.D. Research on the Fingerprint Identification Technology and Attendance System Application. Master's Thesis, Nanjing University of Posts and Telecommunications, Nanjing, China, 2017.
7. Dong, M. Overview of Fingerprint Identification Technology Development. *China Sci. Technol. Inf.* **2011**, *13*, 70.
8. Li, J.L.; Wang, H.B.; Tao, L. Multi-feature recognition of palm vein and palm print in single near-infrared palm image. *Comput. Eng. Appl.* **2018**, *54*, 156–164+236.
9. Chen, Y.H. Research and Implementation of Iris Recognition Key Problems. Master's Thesis, Jilin University, Changchun, China, 2015.
10. Jiao, X.H. Research and Implementation of Iris Authentication Technology Based on Embedded. Master's Thesis, Heilongjiang University, Harbin, China, 2018.
11. Zhang, Y. Ear Detection and Recognition under Uncontrolled Conditions Based on Deep Learning Algorithm. Ph.D. Thesis, University of Science and Technology Beijing, Beijing, China, 2008.
12. Kumar, A.; Zhang, D. Ear authentication using log-Gabor wavelets. Biometric Technology for Human Identification IV. *Int. Soc. Opt. Photonics* **2007**, *6539*, 65390A.
13. AsmaaSabet, A.; Kareem Kamal A, G.; Hesham, E. Human Ear Recognition Using SIFT Features. In Proceedings of the 2015 Third World Conference on Complex Systems (WCCS), Marrakech, Morocco, 23–25 November 2015; pp. 1–6.
14. Nosrati, M.S.; Faez, K.; Faradji, F. Using 2D wavelet and principal component analysis for personal identification based on 2D ear structure. In Proceedings of the 2007 International Conference on Intelligent and Advanced Systems, Kuala Lumpur, Malaysia, 25–28 November 2007; pp. 616–620.
15. Omara, I.; Li, X.M.; Xiao, G.; Adil, K.; Zuo, W. Discriminative local feature fusion for ear recognition problem. In Proceedings of the 2018 8th International Conference on Bioscience, Biochemistry and Bioinformatics (ICBBB 2018), Association for Computing Machinery, New York, NY, USA, 18–21 January 2018; pp. 139–145.
16. Xie, C.X.; Mu, Z.C.; Xie, J.J. Multi-pose ear recognition based on LLE. *J. Intell. Syst.* **2008**, *4*, 321–327.
17. Qian, Y.L.; Gai, S.Y.; Zheng, D.L. Fast 3D ear recognition based on local and global information. *J. Instrum.* **2019**, *40*, 99–106.
18. Susan, E. Ear Detection in the Wild using Faster R-CNN. In Proceedings of the 2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM), Barcelona, Spain, 28–31 August 2018; pp. 1124–1130.
19. Lei, Y.M.; Du, B.W.; Qian, J.R.; Feng, Z.B. Research on Ear Recognition Based on SSD_MobileNet_v1 Network. In Proceedings of the 2020 Chinese Automation Congress (CAC), Shanghai, China, 6–8 November 2020; pp. 4371–4376.
20. Qian, J.R. Research on Dynamic Human Ear Recognition Method Based on Deep Learning. Master's Thesis, Chang Chun University, Changchun, China, 2020.
21. Qi, J. Research on Target Identification Method Based on Human Ear Detection Technology. Master's Thesis, ChangChun University, Changchun, China, 2020.
22. Redmon, J.; Divvala, S.; Girshick, R.; Farhadi, A. You only look once: Unified real-time object detection. In Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 779–788.
23. Redmon, J.; Farhadi, A. YOLO9000: Better, Faster, Stronger. In Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 21–26 July 2017; pp. 6517–6525.
24. Redmon, J.; Farhadi, A. YOLOv3: An Incremental Improvement. *arXiv* **2018**, arXiv:1804.02767.
25. Bochkovskiy, A.; Wang, C.Y.; Liao, H.Y.M. YOLOv4: Optimal Speed and Accuracy of Object Detection. *arXiv* **2020**, arXiv:2004.10934.
26. Jirarat, I.; Surapon, N.C.; Suchart, Y. Deep Learning-based Face Mask Detection Using YoloV5. In Proceedings of the 2021 9th International Electrical Engineering Congress, Pattaya, Thailand, 10–12 March 2021; pp. 428–431.
27. Li, C.; Li, L.; Jiang, H.; Weng, K.; Geng, Y.; Li, L.; Ke, Z.; Li, Q.; Cheng, M.; Nie, W.; et al. YOLOv6: A single-stage object detection framework for industrial applications. *arXiv* **2022**, arXiv:2209.02976.

28. Wang, C.Y.; Bochkovskiy, A.; Liao, H.Y.M. YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors. *arXiv* **2022**. [CrossRef]
29. Ju, R.Y.; Cai, W.M. *Fracture Detection in Pediatric Wrist Trauma X-ray Images Using YOLOv8 Algorithm*; Springer: Berlin/Heidelberg, Germany, 2023.
30. Howard, A.; Sandler, M.; Chu, G.; Chen, L.C.; Chen, B.; Tan, M.X.; Wang, W.J.; Zhu, Y.K.; Pang, R.M.; Vasudevan, V. Searching for MobileNetV3. International Conference on Computer Vision. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul, Republic of Korea, 27 October–2 November 2019; pp. 1314–1324.
31. Han, K.; Wang, Y.; Tian, Q.; Guo, J.; Xu, C.; Xu, C. GhostNet: More Features from Cheap Operations. In Proceedings of the 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Seattle, WA, USA, 13–19 June 2020; pp. 1577–1586.
32. Ma, N.N.; Zhang, X.Y.; Zheng, H.-T.; Sun, J. ShuffleNet V2: Practical Guidelines for Efficient CNN Architecture Design. *Comput. Vis. ECCV* **2018**, *2018*, 122–138.
33. Howard, A.G.; Zhu, M.; Chen, B.; Kalenichenko, D.; Wang, W.; Weyand, T.; Andreetto, M.; Adam, H. MobileNets: Efficient Convolutional Neural Networks for Mobile Vision Applications. *arXiv* **2017**, arXiv:1704.04861.
34. Sandler, M.; Howard, A.; Zhu, M. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 4510–4520.
35. Zhang, X.; Zhou, X.; Lin, M. ShuffleNet: An Extremely Efficient Convolutional Neural Network for Mobile Devices. In Proceedings of the 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 6848–6856.
36. Lei, Y.; Qian, J.; Pan, D.; Xu, T. Research on Small Sample Dynamic Human Ear Recognition Based on Deep Learning. *Sensors* **2022**, *22*, 1718. [CrossRef]
37. Available online: http://www1.ustb.edu.cn/resb/visit/visit.htm (accessed on 30 December 2021).
38. Hoang, V.T. EarVN1.0: A new large-scale ear images dataset in the wild. *Sci. Direct* **2019**, *27*, 104630. [CrossRef] [PubMed]
39. Liu, C.H.; Pan, L.H.; Yang, F.; Zhang, R. Improved YOLOv5 lightweight mask detection algorithm. *Comput. Eng. Appl.* **2023**, *59*, 232–241.
40. Chen, K.; Liu, X.; Ja, L.J.; Fang, Y.L.; Zhao, C.X. Insulator Defect Detection Based on Lightweight Network and Enhanced Multi-scale Feature. *High Volt. Eng.* **2023**, 1–14. [CrossRef]
41. Zou, P.; Yang, K.J.; Liang, C. Improved YOLOv5 algorithm for real-time detection of irregular driving behavior. *Comput. Eng. Appl.* **2023**, 1–9. Available online: http://kns.cnki.net/kcms/detail/11.2127.TP.20230206.1311.003.html (accessed on 6 February 2023).