Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 9069007, 15 pages
https://doi.org/10.1155/2022/9069007

Research Article

An Efficient and Lightweight Method for Human Ear Recognition Based on MobileNet

Xuebin Xu,[1,2] **Yibiao Liu**,[1,2] **Shuxin Cao**,[1,2] **and Longbin Lu**[1,2]

[1]*School of Computer Science and Technology, Xi'an University of Posts & Telecommunications, Xi'an Shaanxi 710121, China*
[2]*Shaanxi Key Laboratory of Network Data Analysis and Intelligent Processing, Xi'an University of Posts & Telecommunications, Xi'an Shaanxi 710121, China*

Correspondence should be addressed to Yibiao Liu; liuyibiao@stu.xupt.edu.cn

Received 22 June 2022; Revised 5 October 2022; Accepted 11 October 2022; Published 30 October 2022

Academic Editor: Xin Ning

In recent years, biometric recognition patterns have attracted the attention of many researchers, among which human ears, as a unique and stable biometric feature, have significant advantages in verifying personal identity. In the Internet era, a system with low computing cost and good real-time performance is more popular. Most of the existing ear recognition methods are based on a large parameter network model, which causes a large memory footprint and computational overhead. This paper proposes an efficient and lightweight human ear recognition method (ELERNet) based on MobileNet V2. Based on the MobileNet V2 model, dynamic convolution decomposition is introduced to enhance the representation ability of human ear features. Then, combined with the coordinate attention mechanism, the spatial features of human ear images are aggregated to locate the location information of the human ear features more accurately. We conducted experiments on AWE and EarVN1.0 human ear datasets. Compared with the MobileNet V2 model, the recognition accuracy of our method is significantly improved. Using less computing hardware resources, the ELERNet model achieves 83.52% and 96.10% Rank-1 (R1) recognition accuracy, respectively, which is better than other models. Finally, we provide a visual interpretation using GradCAM technology, and the results show that our method can learn specific and discriminative features in the ear images.

1. Introduction

Biometric recognition has developed rapidly in the last decade, and it uses common characteristics of individuals for recognition. For example, iris [1–3], face [4, 5], fingerprint [6–8], gait [9, 10], electrocardiogram (ECG) and electroencephalogram (EEG) [11–13], and voice [14, 15] are commonly used techniques for biometric recognition. Human ear recognition has unique advantages over other biometric recognition techniques and has recently received much attention. Unlike face recognition, it is not affected by changes in facial expressions, and it can be done without contact. Different ear structures are unique [16, 17]. The ear characteristics do not change much over time, so the human ear pattern can already be used by police as evidence for identification [18]. In modern forensic identification and criminal investigations, multimodal recognition such as ear, face, and palm print ensures foolproof identification [19].

Many experiments have been conducted on constrained and unconstrained human ear datasets in recent years. Among them, the constrained human ear dataset has a single shooting angle, illumination, background, and resolution and is less difficult to identify. On the contrary, unconstrained human ear datasets are relatively more difficult to recognize due to the large inter- and intraclass variation. Researchers proposed several ear recognition methods based on handcrafted features in the early days of human ear recognition. Most methods do not use a baseline ear database and standard evaluation metrics to assess the performance of the model, and there is a slight variation in the ear images in the database. When these methods experiment on unconstrained ear databases, the recognition performance degrades significantly and is much lower than that of the deep learning-based methods. Currently, the use of deep learning [20, 21] is becoming more and more common. Deep learning-based techniques are used in various fields,

such as human ear recognition and human activity recognition (HAR) [22–24]. So researchers have proposed many in-depth feature learning-based methods for human ear recognition and achieved good recognition performance in unconstrained ear databases. However, as deep neural networks continue to evolve, their drawbacks continue to be exposed. Most networks have a large number of parameters and high model complexity. Because they require extremely high hardware requirements, they are difficult to be applied to embedded devices and mobile terminals and can only be used in individual scenarios. The rapid development of the mobile Internet has led to the growing demand for lightweight networks and real-time performance. Therefore, Google has successively proposed MobileNet V1 [25], MobileNet V2 [26], and MobileNet V3 [27]. They are easy to deploy on embedded devices and mobile terminals with much-reduced computation and parameters while maintaining the recognition performance. Therefore, we use MobileNet V2 as a baseline network for ear recognition and propose a highly efficient and lightweight human ear recognition method based on MobileNet. In addition, our proposed method can be used for medical image analysis. For example, patients with suspected Hepatitis C Virus (HCV) [28] infection are classified into two categories, healthy and unhealthy, to help clinicians diagnose and treat HCV.

Our contributions can be summarized as follows: (1) we propose an efficient and lightweight human ear recognition method (ELERNet) based on MobileNet. The model consumes fewer computational hardware resources and is easy to apply to mobile and embedded devices. (2) To enhance the ear feature representation capability of the model, dynamic convolutional decomposition [29] is introduced to reduce the difficulty of ear feature extraction. (3) To enhance the feature robustness of the model, a coordinate attention mechanism [30] is introduced. The spatial features of the ear image are aggregated to precisely locate the location information of the spatial features of the ear, which improves the recognition performance of the model. (4) We conducted extensive experiments on two representative unconstrained human ear datasets, AWE [31–33] and EARVN1.0 [34], which showed excellent recognition performance. Compared with existing human ear recognition models, ELERNet has significantly higher recognition accuracy with a small memory footprint and computational overhead. (5) We used the Gradient-Weighted Class Activation Mapping (GradCAM) [35] technique to explain how MobileNet V2, as well as our predictions made by the proposed model ELERNet. The visualization highlights that our model can learn specific and discriminative features in the ear image.

2. Related Work

In the early days of human ear recognition, researchers based their recognition methods primarily on handcrafted features. In [36], the authors proposed ear recognition based on Scale-Invariant Feature Transform (SIFT) features and homography distance. The recognition performance of this method is better than Principal Component Analysis (PCA). It also shows excellent robustness under slight angle changes, background interference, and occlusion. The disadvantage of their method is that they do not use a standard benchmark database and do not use evaluation criteria to assess the model performance. In another study, the authors extracted ear boundary features using a wavelet approach [37]. The ear features were then saved to a database for matching. The disadvantages of this method are that no precise performance evaluation metrics were used, and the experiments were conducted on a small dataset. A 2D orthogonal filter-based human ear recognition method was proposed in [38]. The method first performs ear feature segmentation, and then ear features are extracted. The experimental results show that the 2D orthogonal filter has excellent recognition performance. The drawback of the method is that the ear images in the database it uses hardly change much. In [39], the authors designed a method for ear feature extraction using local binary patterns (LBP). The results show that LBP outperforms Principal Component Analysis (PCA). The drawback is that the experiments were evaluated on a database of images captured indoors. In [40], the authors performed a comparative analysis of human ear recognition based on the average and uniform variants of LBP. The method achieved a desirable recognition performance on constrained databases. However, when the experiments were performed on the unconstrained ear database, the recognition performance significantly decreased. In [41], the authors proposed a pattern recognition method that uses edge ear features to learn local ear features. The method is robust to small magnitudes of illumination and rotation. The recognition performance of the method is significantly better than other descriptor-based methods. The disadvantage is that the recognition performance on unconstrained databases needs to be improved. In [42], the authors first extracted global features using the Gabor-Zernike operator and then local features using the local phase quantization operator. The method was evaluated on three constrained datasets and achieved perfect recognition results. However, the recognition results of the method on the unconstrained datasets still fall short of the deep learning-based methods.

With the emergence of deep learning in recent years, especially the development of deep convolutional neural networks (CNN), it can solve most computer vision problems. Researchers have proposed many methods for human ear recognition based on deep feature learning and achieved good recognition performance. In [43], the authors modified common CNN architectures such as ResNet, VGG face, and GoogleNet to validate them on unconstrained datasets. To enable the network to learn multiscale information features, the authors use a spatial pyramid-pooling layer to replace the last pooling layer of the CNN model to add central loss during training. In addition, the authors provide a new database of images captured under challenging outdoor conditions USTB-HelloEar. Experimental results show that the VGG face model has the best recognition performance. The disadvantages of this approach are that no performance evaluation metrics are used to evaluate the model, and the model has a large memory footprint and high computational

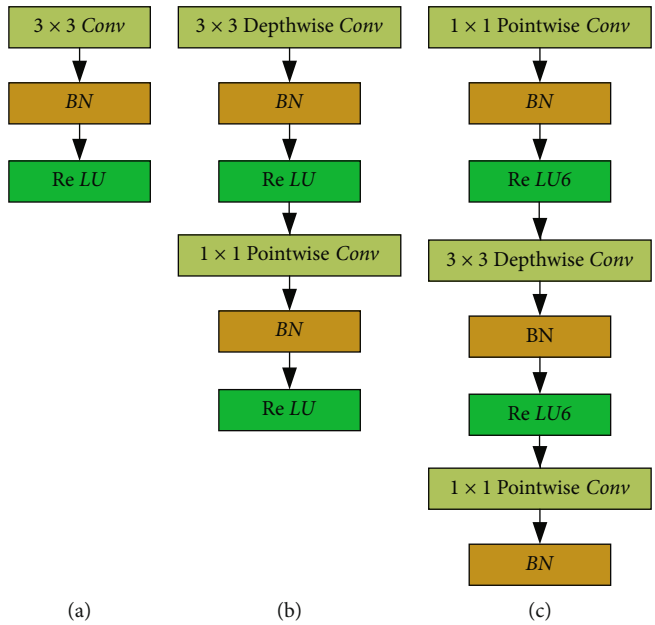


Figure 1: Comparison of the structure diagrams of the three convolutional approaches. (a) Ordinary convolution. (b) Depth-separable convolution DSC. (c) Improved depth-separable convolution IDSC.

cost. In [44], the authors first used RefiNet for ear detection and then ResNet for ear recognition. The method achieved good recognition performance on an unconstrained database, showing the advantages of deep learning-based methods. The disadvantage of the method is that ear detection is based on existing methods, and ear recognition has limited innovation. Moreover, the system needs to consume more computational hardware resources. In [45], the authors used integrated learning, feature extraction, and fine-tuning learning strategies based on models such as Inception, ResNext, and VGG. Good recognition results were achieved on publicly available unconstrained databases. The drawback of the method is that the performance is evaluated on only one database, which does not highlight the model's generalization ability. Moreover, the large number of model parameters makes it difficult to embed the model into mobile applications for specific ear recognition scenarios.

This paper proposes an efficient and lightweight human ear recognition method (ELERNet) based on MobileNet V2. The model is evaluated on two publicly available unconstrained datasets. The large intra- and interclass variation of the unconstrained human ear datasets leads to the difficulty of ear feature extraction. We introduced a dynamic channel fusion mechanism to reduce potential spatial features' dimensionality to implement the dynamic convolutional decomposition [29] and enhanced the ear feature representation. Considering that the unconstrained human ear dataset varies significantly regarding shooting angle, illumination, background, and resolution size, these factors increase recognition difficulty. Therefore, we introduced the coordinate attention mechanism [30]. It aggregates the spatial features of unconstrained human ear images to obtain a coordinate-aware ear feature map. Then, the location information of the spatial features of the ear is precisely located, which dramatically enhances the feature robustness of the model.

3. Method

3.1. MobileNet V2. Since AlexNet [46] won the ImageNet challenge, the deep convolutional neural network craze has been rekindled. Convolutional neural networks are found everywhere in computer vision tasks. In order to achieve higher accuracy, researchers have designed increasingly complex convolutional neural network models with a larger and larger number of parameters, leading to a significant decrease in operational efficiency. In some real-world scenarios, recognition tasks need to be performed promptly on computationally constrained platforms. An example is this paper's work related to human ear recognition. In order to solve the above problem, MobileNet V1 [25], a model with a small number of parameters and low latency, was proposed by Google. Its network idea is mainly to replace the standard convolutional operation with Deep Separable Convolution (DSC), which dramatically reduces the model parameters. 3×3 Depthwise Conv used by DSC generates the output channel after performing the convolution operation, and it has only one layer of thickness, which can be slid layer by layer over the input tensor. Then the thickness is adjusted using 1×1 Pointwise Conv. In order to solve the loss of DSC feature information, MobileNet V2 [26] was proposed, which improved the original DSC, which we call Improved DSC (IDSC). Figure 1 compares the ordinary convolution, the depth-separable convolution, and the improved depth-separable convolution.

3.2. Attention Module. Both datasets used in this paper are unconstrained human ear datasets with a significant

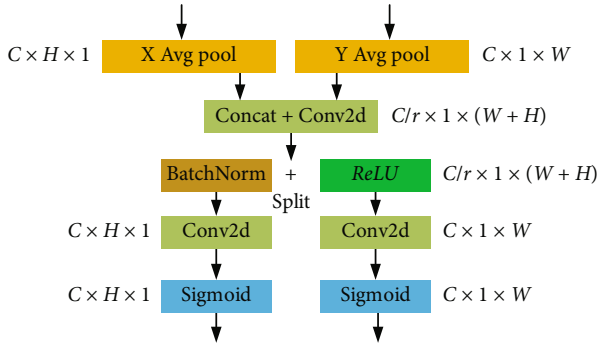


Figure 2: Coordinate attention module structure diagram.

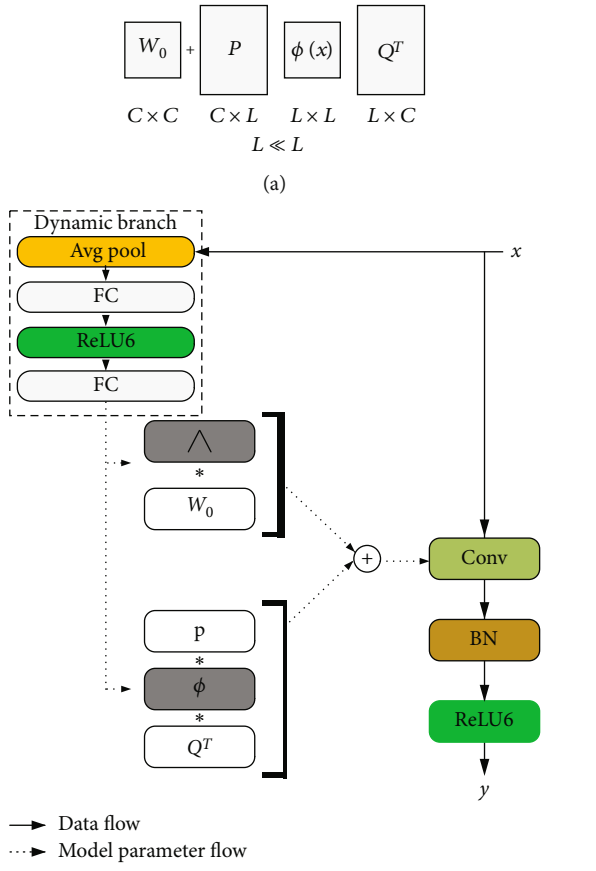


Figure 3: (a) Dynamic convolution via matrix decomposition. (b) Structural diagram of the dynamic convolution decomposition layer.

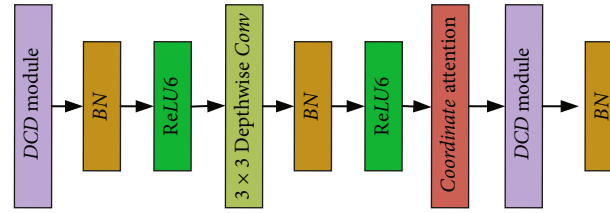


Figure 4: IDSCPlus block structure diagram.

Table 1: ELERNet body architecture.

Type	Filter shape	Output size
Conv	3 × 3 × 3 × 32	112 × 112 × 32
Conv dw	3 × 3 × 3 × 32 dw	112 × 112 × 32
Conv	1 × 1 × 32 × 16	112 × 112 × 16
IDSCPlus-1	—	56 × 56 × 24
IDSCPlus-2	—	56 × 56 × 24
IDSCPlus-3	—	28 × 28 × 32
IDSCPlus-4	—	28 × 28 × 32
IDSCPlus-5	—	28 × 28 × 32
IDSCPlus-6	—	14 × 14 × 64
IDSCPlus-7	—	14 × 14 × 64
IDSCPlus-8	—	14 × 14 × 64
IDSCPlus-9	—	14 × 14 × 64
IDSCPlus-10	—	14 × 14 × 96
IDSCPlus-11	—	14 × 14 × 96
IDSCPlus-12	—	14 × 14 × 96
IDSCPlus-13	—	7 × 7 × 160
IDSCPlus-14	—	7 × 7 × 160
IDSCPlus-15	—	7 × 7 × 160
IDSCPlus-16	—	7 × 7 × 320
Conv	1 × 1 × 320 × 1280	7 × 7 × 1280
AvgPool	Pool 1 × 1	1 × 1 × 1280
DCD-CLS	Classifier	1 × 1 × 100/164

intraclass variation. Interference features such as background and ear ornaments in the ear images can negatively affect the recognition performance. Reducing the negative impact of these interference features on the recognition performance makes the model focus mainly on the ear contour when feature extraction is performed on the ear images. We insert the coordinate attention module [30] behind the 3 × 3 Depthwise Conv layer in the IDSC module. Its structure diagram is shown in Figure 2. Unlike other attention mechanisms, it can embed location information into channel attention with almost no computational overhead. Coordinate attention can decompose channel attention into two one-dimensional feature encoding processes that aggregate features along two spatial directions. Thus, it enhances the extraction of features of interest in ear images.

The coordinate attention mechanism can be divided into coordinate information embedding and coordinate attention generation. The first part retains location information critical to recognizing performance, and the global pool is decomposed into two 1D feature codes. Given an input x, the pool core of the spatial scope $(H, 1)$ is used to encode the channel along with horizontal coordinates, and similarly $(1, W)$ is used to encode the channel along

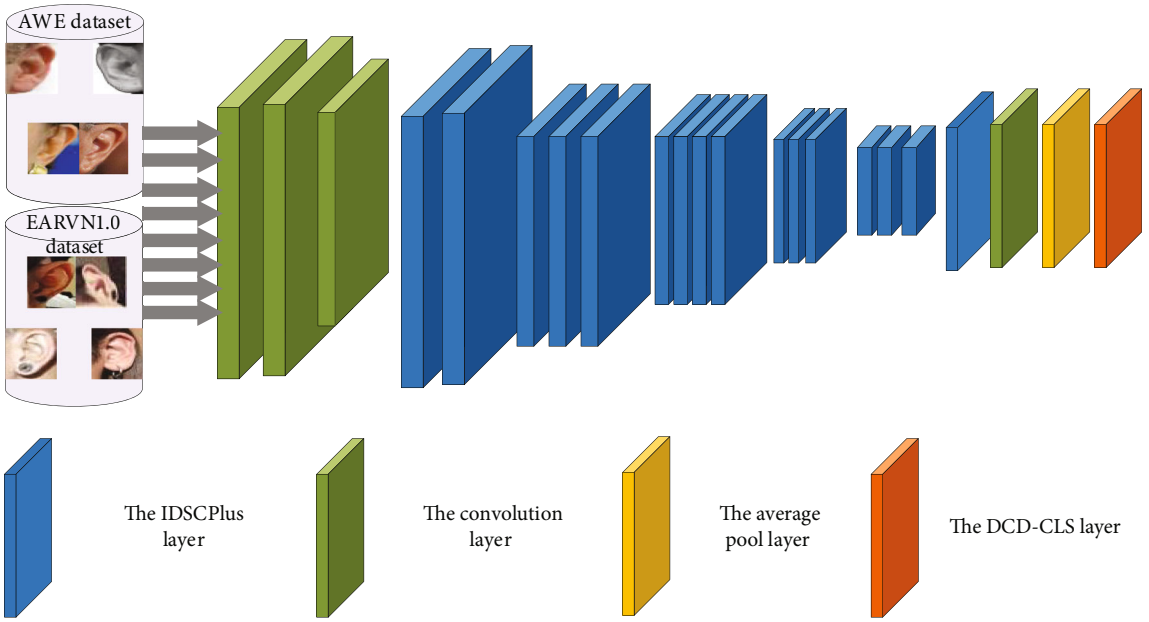
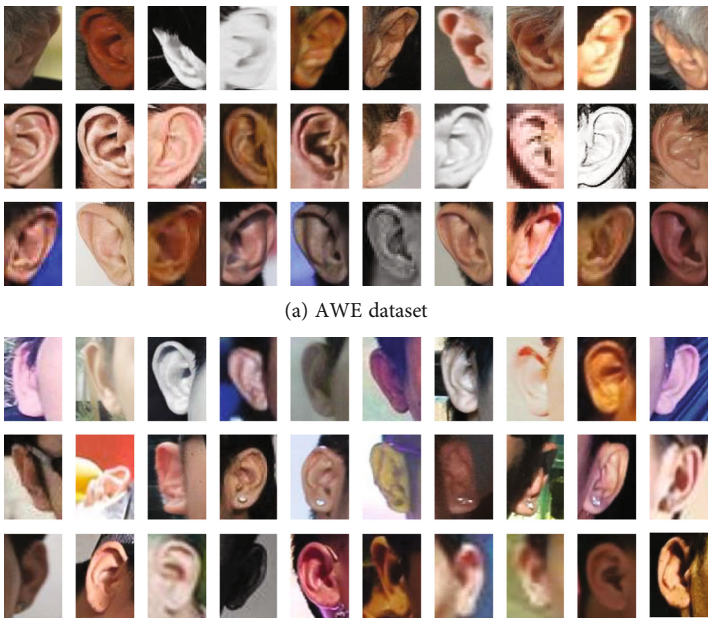


Figure 5: The basic framework of ELERNet.



(a) AWE dataset

(b) EARVN1.0 dataset

Figure 6: Ear images from three subjects in AWE and EARVN1.0. These ear images have considerable variations in resolution, background, illumination, angle, and occlusion.

with vertical coordinates. The output of the c-th channel at height H is

$$z_c^h(h) = \frac{1}{W} \sum_{0 \le i < W} x_c(h, i). \tag{1}$$

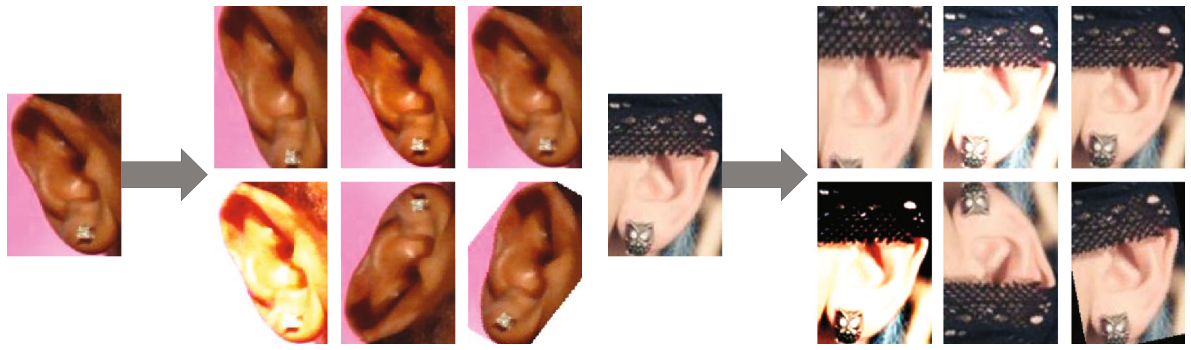
The output of the c-th channel of width W is

$$z_c^w(w) = \frac{1}{H} \sum_{0 \le j < H} x_c(j, w). \tag{2}$$

The second part is the generation of coordinate attention: we connect the aggregation feature map generated by equations (1) and (2) and then obtain equation (3) through F_1.

$$f = \delta\left(F_1\left(\left[z^h, z^w\right]\right)\right), \tag{3}$$

where f is the feature map, δ is the nonlinear activation function, F_1 is the 1×1 convolution transformation function, and $[\cdot, \cdot]$ denotes concatenation of spatial dimensions. To



(a) Augmented AWE dataset (b) Augmented EARVN1.0 dataset

Figure 7: Data augmentation operations on human ear images from AWE and EARVN1.0 training sets.

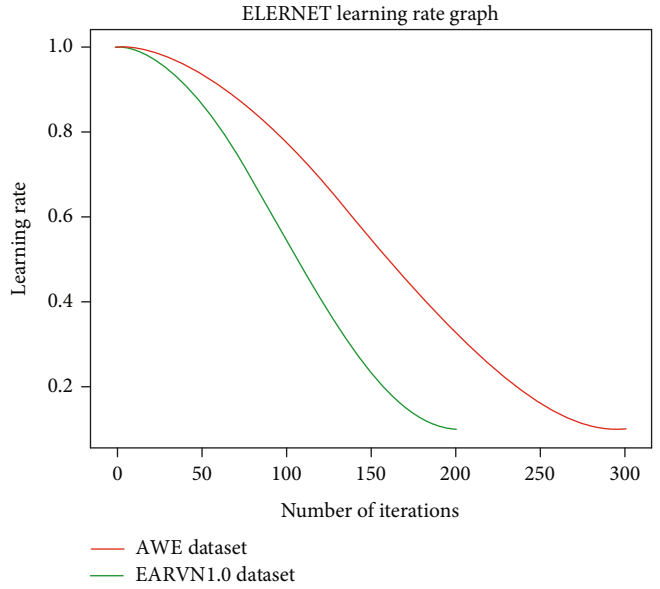


Figure 8: Graphs of learning rate changes under different data sets.

obtain the input x, we split f into two independent tensors, f^h and f^w. The number of channels of the two independent tensors is equal by the 1×1 convolutional transformation functions F^h and F^w. The specific process is

$$g^h = \sigma\left(F_h\left(f^h\right)\right), \tag{4}$$

$$g^w = \sigma(F_w(f^w)), \tag{5}$$

where g^h and g^w are the attention weights, σ is the sigmoid function. The output y of the coordinate attention block is

$$y_c(i,j) = x_c(i,j) \times g_c^h(i) \times g_c^w(j). \tag{6}$$

3.3. Dynamic Convolution Decomposition. Since the two human ear datasets used in this paper are both wild datasets, the samples of the same subject are pretty different. Most ear images have significant differences in angle, resolution, etc. It is not easy to use ordinary convolution to extract the features.

Table 2: Compare the number of model parameters, model computational complexity, and quantitative performance metrics R1 of MobileNet V2 with other models.

Model	Params	M-adds	AWE (R1)	EARVN1.0 (R1)
MobileNet V2	3.5 M	300.0 M	80.51%	91.09%
MobileNet V3-large	5.4 M	219.0 M	80.82%	91.48%
MobileNet V3-small	2.9 M	66 M	72.81%	80.62%
ShuffleNet V1 [47]	2.3 M	140 M	77.49%	88.17%
ShuffleNet V2 [48]	4.1 M	147 M	78.00%	88.75%

To adaptively extract the ear features of interest, we replace the 1×1 Pointwise Conv layer in the IDSC module with a dynamic convolutional decomposition [29] module. It fuses $\phi(x)$ by applying dynamic channels in the low-dimensional space ($Q^T x \in R^L, L \ll C$), as shown in Figure 3(a). Enhancing the learning of the corresponding channels of the high-

Table 3: Comparison of model parametric quantities, model computational complexity, and quantitative performance metrics (R1, R5, and AUC) for DCD at different layers. We highlight in bold the optimal values of the performance metrics.

DW	PW	CLS	Params	M-adds	AWE			EARVN1.0		
					R1	R5	AUC	R1	R5	AUC
			3.5 M	300.0 M	80.51%	91.29%	98.81%	91.09%	97.62%	98.83%
✓			4.2 M	301.6 M	82.01%	93.60%	98.84%	93.78%	98.47%	98.85%
	✓		5.0 M	323.5 M	82.71%	93.63%	98.86%	95.04%	98.72%	98.88%
		✓	3.9 M	300.6 M	81.41%	93.56%	98.82%	91.98%	97.94%	98.84%
✓		✓	4.6 M	302.2 M	82.27%	93.62%	98.83%	94.23%	98.51%	98.86%
✓	✓		5.8 M	325.1 M	82.62%	93.63%	98.86%	94.86%	98.58%	98.86%
	✓	✓	5.5 M	324.1 M	**83.02%**	**93.68%**	**98.89%**	**95.58%**	**98.85%**	**98.90%**
✓	✓	✓	6.1 M	325.7 M	82.82%	93.65%	98.87%	95.22%	98.77%	98.88%

dimensional potential space and reducing the dimensionality of the potential space make the model parametric number small and low complexity, which improves the feature expression of the model. The dynamic channel fusion mainly uses an L × L matrix $\phi(x)$ to achieve. $\phi_{i,j}(x)$ is the function of the input x. Through $\phi(x)$, the L channels dynamic fusion and then uses P to increase the number of output channels

$$W(x) = W_0 + P\Phi(x)Q^T = W_0 + \sum_{i=1}^{L}\sum_{j=1}^{L} p_i \Phi_{i,j}(x) q_j^T. \quad (7)$$

The dynamic convolution decomposition layer is shown in Figure 3(b). It uses dynamic branches to generate the coefficients of dynamic channel attention $\wedge(x)$ and dynamic channel fusion $\phi(x)$. Input x first passes through the average pool, then through the first fully connected layer (FC), using ReLU6 as the activation layer, and finally through the second fully connected layer (FC).

3.4. ELERNet Introduction. In order to improve the feature representation capability of the model and better cope with the challenges posed by the considerable variation within the same category of the unconstrained human ear dataset. At the same time, the ear features of interest are extracted adaptively, and interference features are filtered to enhance the model's robustness. We improve the IDSC module and call it the IDSCPlus block, as shown in Figure 4. We replaced the 1 × 1 Pointwise Conv layer with the DCD module and inserted the CA attention module behind the 3 × 3 Depthwise Conv layer.

The structure of ELERNet is shown in Table 1. In this model, the input human ear image is first preliminarily extracted through a 3 × 3 standard convolution layer, a 3 × 3 Depthwise Conv layer, and a 1 × 1 standard convolution layer. Then, 16 IDSCPlus modules and a 1 × 1 standard convolution layer are successively used to extract depth features from ear images. Finally, the distinguishing features are obtained and classified through the AvgPool and DCD-CLS layers. The architecture of ELERNet is shown in Figure 5.

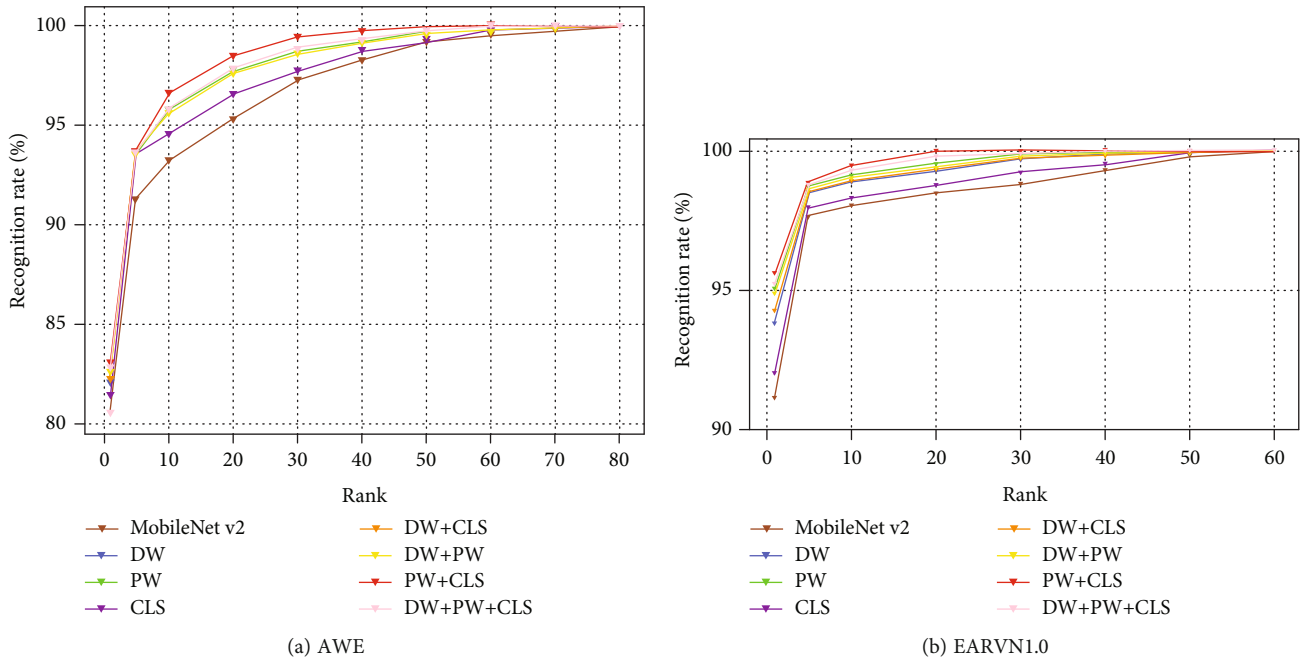
4. Experimental Results and Discussion

4.1. Dataset Introduction. Annotated Web Ears (AWE) [31–33] is a human ear dataset produced by the University of Ljubljana, with 1000 images, including 100 subjects. Each subject has 10 images, which belong to the unconstrained dataset. Some ear images are challenged by decoration and hair occlusions. EARVN1.0 [34] is a new unconstrained human ear dataset that contains 164 subjects with a total of 28,412 images that have undergone significant changes in lighting, scale, and pose. These images have significant variations in lighting, resolution, pose, etc. Most of the images also face challenges such as decoration and background occlusion. Figure 6 shows the ear images of three of the subjects. Since there are many images in the EARVN1.0 dataset, we randomly select 10 images of each subject for display.

4.2. Data Augmentation. During the training of the model, too small a sample size can lead to overfitting of the model. To avoid this phenomenon, adopting aggressive data expansion is a good choice. This way, the model gets different images during the training process, which can significantly improve the model's generalization ability. Figure 7 shows the expanded images.

4.3. Parameter Settings. This paper proposes a human ear recognition method based on the Pytorch open-source framework. The experiment is completed on the NVIDIA Tesla V100 SXM2 16G server. We set up the cosine scheduler and defined the learning rate decay. The specific change curve of the learning rate is shown in Figure 8. We set the number of training iterations to 300 rounds on the AWE dataset in the experiment. We set the number of training iterations to 200 rounds in the experiment on the EARVN1.0 data set. We choose stochastic gradient descent (SGD) as the optimizer of this experiment. The parameters are set to support the learning decay rate, Nesterov momentum, and support momentum parameter, and the batch size is set to 16 for all experiments.

4.4. Evaluation Metrics. The cumulative matching feature (CMC) curve is biometric recognition's most famous performance evaluation metric. We have plotted cumulative

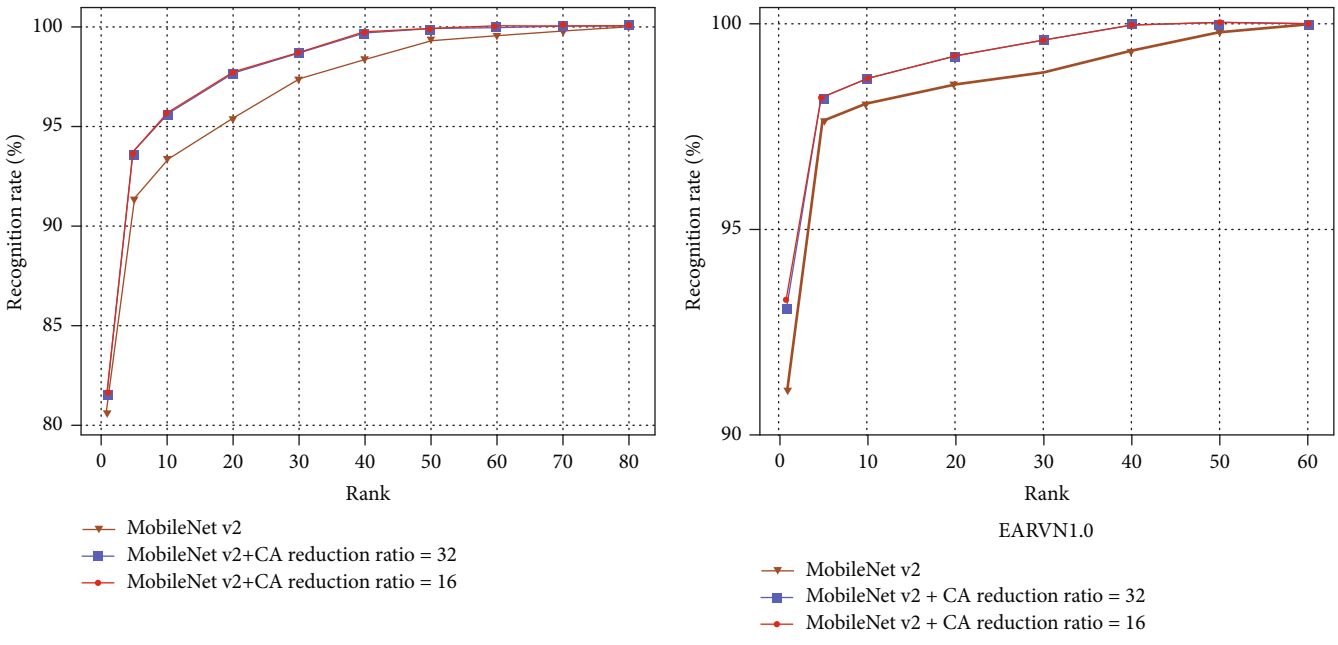


(a) AWE (b) EARVN1.0

Figure 9: The CMC curves compare the recognition performance of DCD at different layers.

Table 4: Comparing the number of model parameters, model computational complexity, and quantitative performance metrics (R1, R5, and AUC) at different reduction ratios. We highlight the optimal values of the performance metrics in bold.

Model	ϒ	Params	M-adds	AWE			EARVN1.0		
				R1	R5	AUC	R1	R5	AUC
MobileNet V2	—	3.50 M	300.0 M	80.51%	91.29%	98.81%	91.09%	97.62%	98.83%
+CA	32	3.95 M	310.0 M	81.52%	93.57%	98.82%	93.08%	98.17%	98.84%
+CA	16	4.37 M	310.0 M	**81.70%**	**93.58%**	**98.83%**	**93.29%**	**98.21%**	**98.84%**

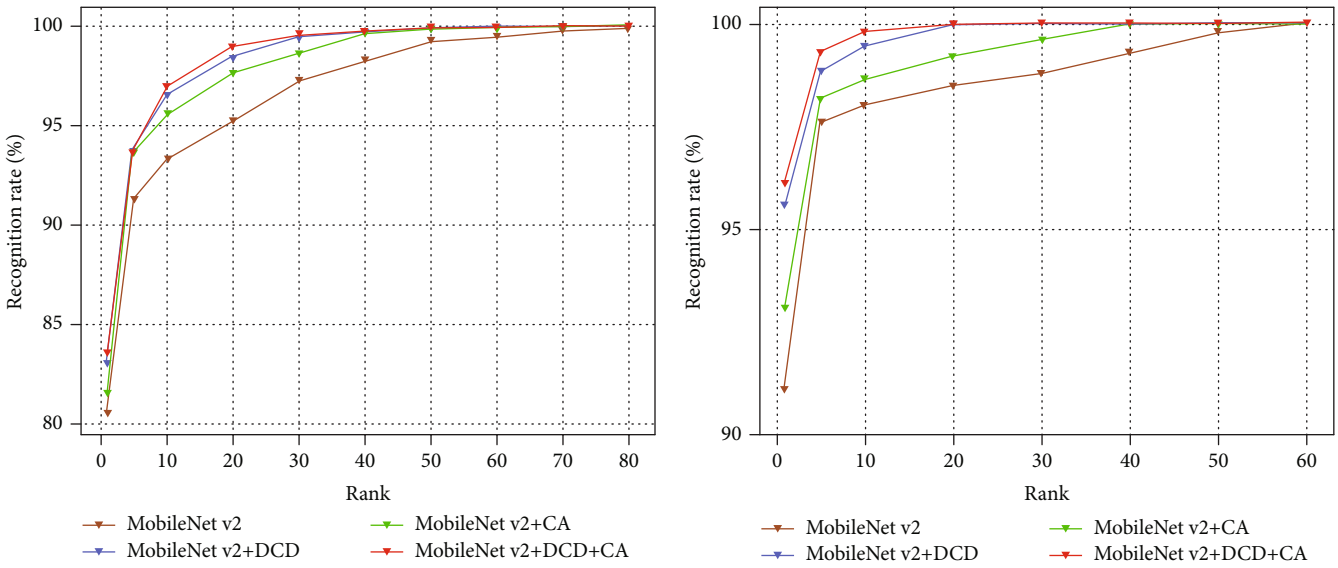


(a) AWE (b) EARVN1.0

Figure 10: The CMC curves compare the impact of changes in the reduction ratio on the recognition performance.

Table 5: We compare the number of model parameters, model computational complexity, and quantitative performance metrics (R1, R5, and AUC) under different ablation experiments. We highlight the optimal values of the performance metrics in bold.

DCD	CA	Params	M-adds	AWE			EARVN1.0		
				R1	R5	AUC	R1	R5	AUC
		3.50 M	300.0 M	80.51%	91.29%	98.81%	91.09%	97.62%	98.83%
✓		5.50 M	324.1 M	83.02%	93.68%	98.89%	95.58%	98.85%	98.90%
	✓	3.95 M	310.0 M	81.52%	93.57%	98.82%	93.08%	98.17%	98.84%
✓	✓	5.95 M	336.0 M	**83.52%**	**93.71%**	**98.97%**	**96.10%**	**99.28%**	**98.92%**



(a) AWE (b) EARVN1.0

Figure 11: Comparing the CMC curves of different recognition models in the ablation experiments.

matching feature (CMC) curves for recognition experiments and evaluated the performance of the recognition models using three quantitative metrics. We briefly describe each metric below.

Cumulative Matching Characteristics (CMC) Curve: this is the probability that a recognition model returns the correct identity within the first k ($k \leq N$) ranks, N being the number of individuals in the entire gallery.

Rank-1 (R1) Recognition Rate: this is the percentage of the most matched probe images in the gallery that are recognized as correct identities.

Rank-1 (R5) Recognition Rate: this is the percentage of correct identities found as the gallery's top five matching probe images.

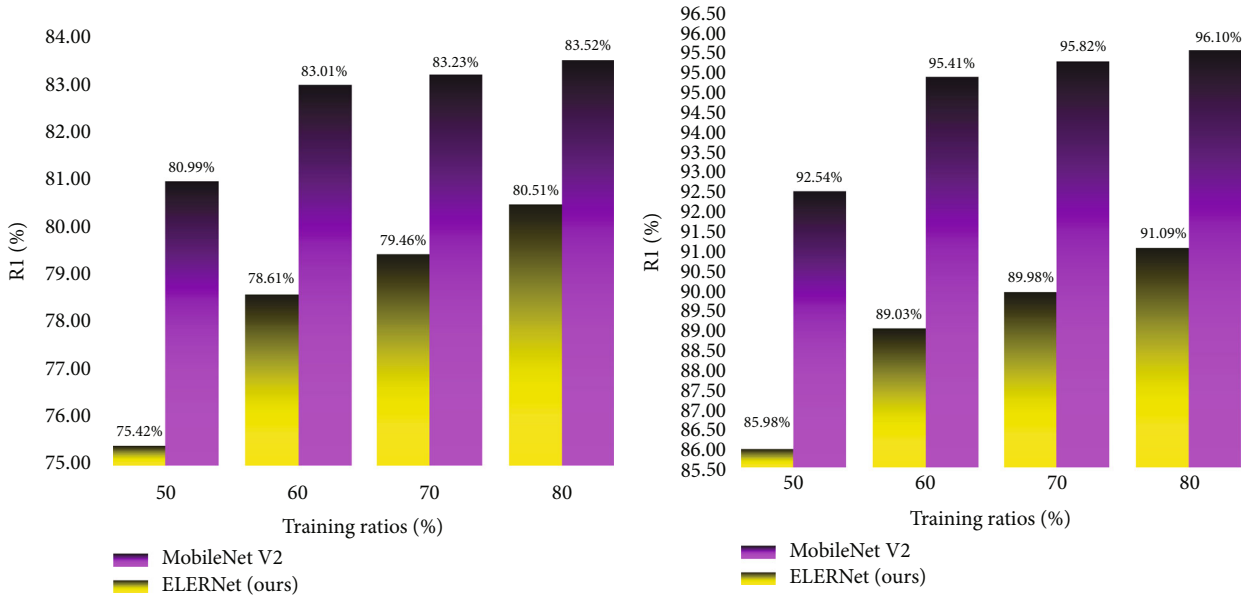
The Area Under the CMC Curve (AUC): Based on the CMC curve, the area under the curve is calculated. A high AUC score indicates a strong model classification performance and a critical evaluation index of the model recognition performance.

4.5. Model Exploration. We selected MobileNet V2, MobileNet V3-Large, MobileNet V3-Small, ShuffleNet V1 [47], and ShuffleNet V2 [48], and five advanced lightweight network models evaluated on AWE and EARVN1.0 ear datasets.

Table 6: Comparison of quantitative performance metrics R1 for MobileNet V2 and ELERNet with different training ratios.

Training ratios	Method	Database	
		AWE (R1)	EARVN1.0 (R1)
50%	MobileNet V2	75.42%	85.98%
	ELERNet (ours)	80.99%	92.54%
60%	MobileNet V2	78.61%	89.03%
	ELERNet (ours)	83.01%	95.41%
70%	MobileNet V2	79.46%	89.98%
	ELERNet (ours)	83.23%	95.82%
80%	MobileNet V2	80.51%	91.09%
	ELERNet (ours)	83.52%	96.10%

Table 2 shows their number of model parameters, model computational complexity, and quantitative performance metrics R1. The experimental results show that MobileNet V3-Small has a small number of model parameters and computational complexity. However, it performs the worst on the AWE and EARVN1.0 ear datasets, with performance metrics R1 of 72.81% and 80.62%, respectively. This is 7.70% and 10.47% lower than that of MobileNet V2. MobileNet V2



(a) AWE (b) EARVN1.0

FIGURE 12: The histogram visually compares the quantitative performance metrics R1 for MobileNet V2 and ELERNet at different training ratios.

has only 0.31% and 0.39% lower performance metrics R1 on the AWE and EARVN1.0 ear datasets compared to MobileNet V3-Large. However, the number of model parameters of MobileNet V2 is 1.9 M smaller than that of MobileNet V3-Large. Models with many parameters are not convenient to deploy to mobile terminals or embedded devices and thus cannot be adapted to specific ear recognition scenarios. ShuffleNet V1 has the smallest number of model parameters and moderate model computational complexity, with performance metrics R1 of 77.49% and 88.17%, which are 3.02% and 2.92% lower than MobileNet V2, respectively. ShuffleNet V2 has 0.6 M more model parameters than MobileNet V2, moderate model computational complexity, and performance metrics R1 of 78.00% and 88.75%, respectively, which are 2.51% and 2.34% lower than those of MobileNet V2.

4.6. The Impact of DCD at Different Layers. Table 3 shows the results of inserting DCD into three different layers, including (1) Depthwise conv (DW), (2) Pointwise conv (PW), and (3) fully connected classifier (CLS). According to the experimental results, the model recognition performance can be improved by using DCD in DW, PW, and CLS layers. The experimental results on AWE dataset are (DW+1.5%, PW+2.2%, and CLS+0.9%), and the experimental results on EARVN1.0 dataset are (DW+2.69%, PW +3.95%, and CLS+0.89%). The results show that the optimal recognition performance can be obtained by combining DCD with PW and CLS simultaneously. To show the differences in recognition performance, Figure 9 plots the CMC curves for DCD at different layers.

4.7. The Impact of Reduction Ratio Υ. We investigate the effect of the reduction ratio on model performance by reducing the size of the reduction ratio and observing the changes in model performance before and after changing the reduction ratio. As shown in Table 4, when we reduce the reduction ratio by half from the original size of 32 to 16, the number of model parameters increases, but the model performance improves. This shows from the side that the robustness of the model to changes in the reduction ratio and adding more parameters by reducing the reduction ratio is beneficial to improve the model performance. When the coordinate attention mechanism was inserted behind the Depthwise Conv layer of the baseline network and the reduction ratios were set to normal 32, the model recognition performance was improved. Experimental results in the AWE dataset were (baseline+1.01%), experimental results on EARVN1.0 dataset (baseline+1.99%).

Figure 10 plots the CMC curves for different reduction ratios.

4.8. Ablation Experiment. In this part, we mainly conducted ablation experiments to prove the influence of dynamic convolution decomposition (DCD) and coordinate attention (CA) on model recognition performance. The experimental results are presented in Table 5. According to the experimental results in the table, the model recognition performance will improve when the DCD module or CA module is inserted separately. Nevertheless, when the DCD module and CA module are added to the baseline network simultaneously, the model performance will be optimized, and the optimal results have been highlighted. Figure 11 plots the CMC curves for the ablation experiments.

4.9. The Impact of Different Training Ratios. In this section, we discuss the robustness of the model to the training set and test set partitioning. We divide the training images at

Table 7: The number of model parameters and model computational complexity of existing human ear recognition methods are compared with our proposed method. We highlight the optimal values in bold.

Model	Params	M-adds
AlexNet	57.70 M	363.50 M
VGG16	134.25 M	8000 M
VGG19	140.30 M	10000 M
Inception V3	22.10 M	3000 M
ResNet18	11.69 M	790 M
ResNet34	21.80 M	1720 M
ResNet50	23.8 M	1930 M
ResNet101	42.80 M	3630 M
ResNet152	60.42 M	5660 M
ResNeXt50	23.30 M	2000 M
ResNeXt101	87.10 M	4000 M
ELERNet (ours)	5.95 M	336 M

different scales and then conduct relevant experiments to evaluate the model recognition performance and partitioning robustness of the baseline network and ELERNet when dealing with the training and testing human ear images at different scales. Table 6 shows the recognition performance of two human ear databases (AWE and EARVN1.0) under different training ear image proportions. For AWE and EARVN1.0 human ear datasets, the proportion of ear images in the training set was randomly divided into 50%, 60%, 70%, and 80%. The experimental results show that, with the increase in the proportion of training ear images, the recognition performance of both baseline network and ELERNet on two-ear data sets is significantly improved. However, ELERNet achieved the best performance at the same training ear ratios. It is worth noting that ELERNet was better at training ratio = 50% than baseline network at training ratio = 80%. We used the histogram to show the comparison results more intuitively, as shown in Figure 12.

4.10. Model Parameters and Complexity Comparison. In the introduction, we cited many pieces of literature and discussed many existing ear recognition methods. It is worth noting that most of the methods that emerged in recent years are based on the models listed in Table 7 to build ear recognition models and propose various transfer learning strategies to solve the problem of ear recognition. Table 7 compares the number of parameters and the complexity of different models. It can be seen that the number of model parameters and complexity of the proposed method are the lowest.

4.11. Compared with Other Methods. As shown in Tables 8 and 9, we compared the proposed method with the methods using the AWE and EARVN1.0 human ear data set for human ear recognition in recent years. According to the comparison results, it can be concluded that the proposed method has the best recognition performance.

Table 8: The quantitative performance metrics R1, R5, and AUC of our proposed model (ELERNet) on the AWE ear dataset are compared with the methods that have been proposed in the literature. The optimal values of the performance metrics are highlighted in bold.

Method	AWE		
	R1	R5	AUC
Hassaballah et al. [40]	49.60%	—	—
Emeršič et al. [32]	49.60%	—	—
Dodge et al. [49]	56.35%	74.80%	—
Dodge et al. [49]	68.50%	83.00%	—
Dodge et al. [49]	80.03%	93.48%	—
Zhang et al. [43]	50.00%	70.00%	—
Emeršič et al. [50]	62.00%	80.35%	95.51%
Khaldi et al. [51]	50.53%	76.35%	80.97%
Hassaballah et al. [41]	54.10%	—	—
Khaldi et al. [52]	48.48%	—	—
Khaldi et al. [53]	51.25%	—	—
Wang et al. [54]	82.90%	—	—
Alshazly et al. [55]	67.25%	84.00%	96.03%
Korichi et al. [56]	82.00%	—	—
Omara et al. [57]	78.13%	—	—
Regouid et al. [58]	43.00%	—	—
Kacar and Kirci [59]	47.80%	72.10%	95.80%
Sajadi and Fathi [42]	53.50%	—	—
Omara et al. [60]	72.22%	—	—
Hansley et al. [61]	75.60%	90.60%	97.20%
Aiadi et al. [62]	82.50%	—	—
ELERNet (ours)	**83.52%**	**93.71%**	**98.97%**

Table 9: The quantitative performance metrics R1, R5, and AUC of our proposed model (ELERNet) on the EARVN1.0 ear dataset are compared with methods already proposed in the literature. The optimal values of the performance metrics are highlighted in bold.

Method	EARVN1.0		
	R1	R5	AUC
Mewada et al. [63]	78.88%	—	—
Ramos-Cooper et al. [64]	92.58%	97.88%	97.61%
Alshazly et al. [45]	93.45%	98.42%	99.18%
Alejo [65]	95.31%	—	—
ELERNet (ours)	**96.10%**	**99.28%**	98.92%

4.12. Visual Explanations. In this part of visual interpretation, we use Gradient-weighted Class Activation Mapping (GradCAM) [35]. It allows visual interpretation of the classification recognition (i.e., provides class differentiation interpretation by locating the region of interest in the ear image with class-specific gradient information) and helps us to understand MobileNet V2 and the predictions made by our method ELERNet. We provide some cases where MobileNet V2 makes wrong predictions on subjects, but ELERNet makes correct predictions on subjects. The original

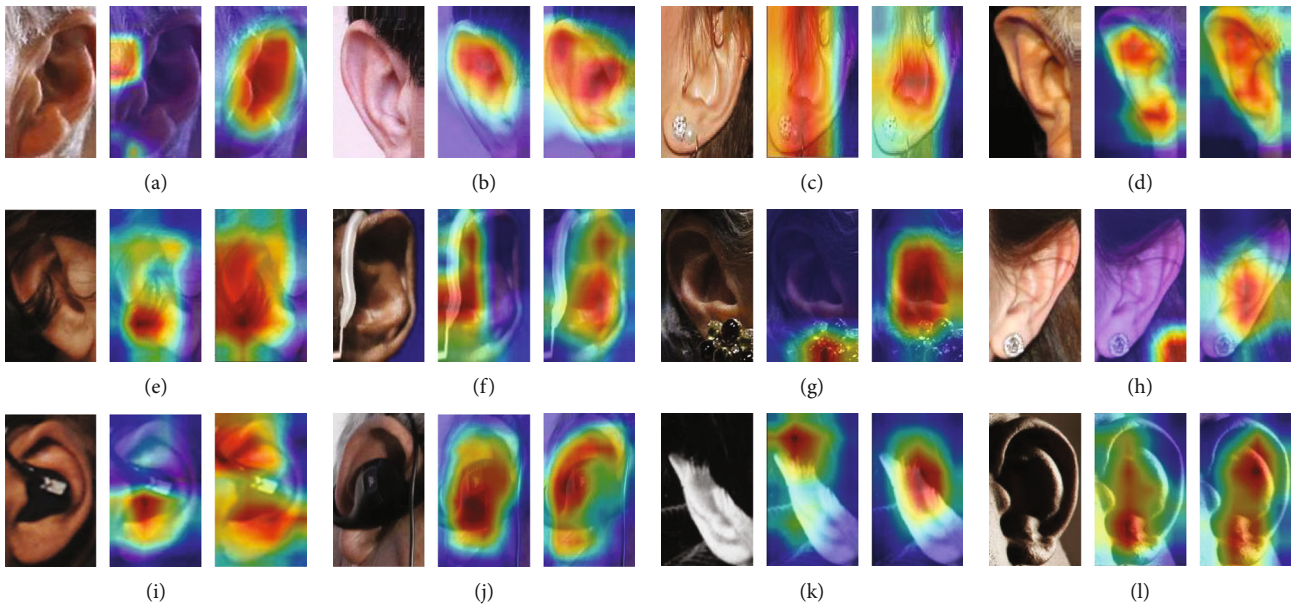


Figure 13: Visual interpretation of subject ear category differentiation by GradCAM for MobileNet V2 and ELERNet on the AWE dataset. The original image is shown on the left, the visualization results of the MobileNet V2 model are shown in the middle, and the visualization results of the ELERNet model are shown on the right.

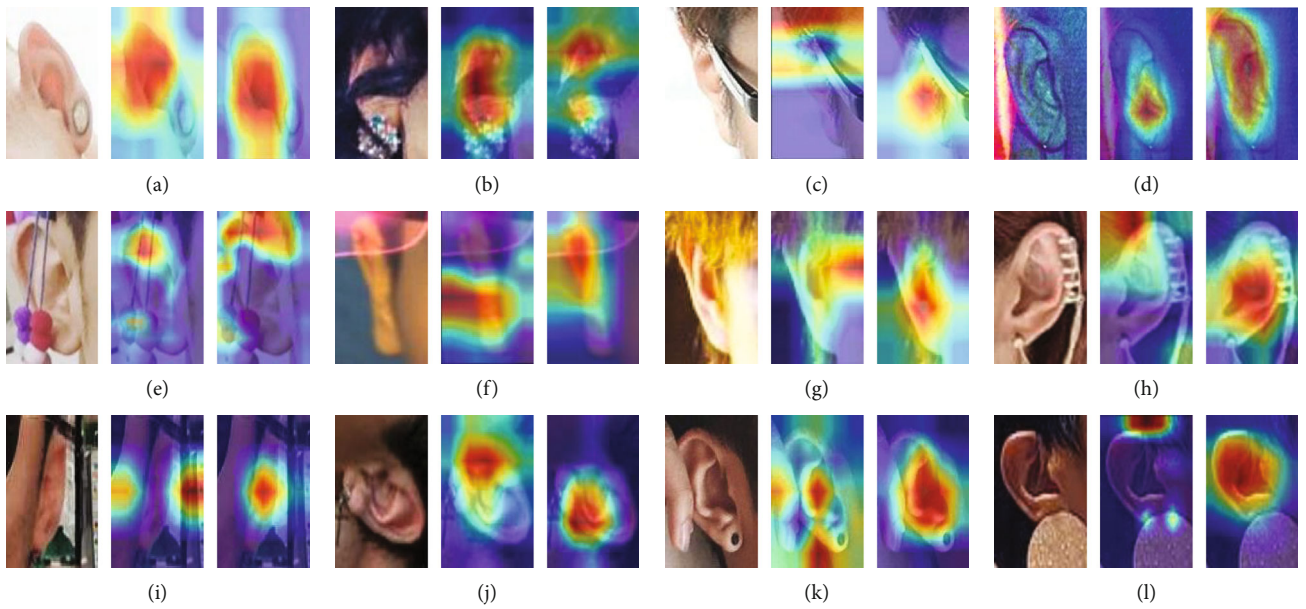


Figure 14: Visual interpretation of subject ear category differentiation by GradCAM for MobileNet V2 and ELERNet on the EARVN1.0 dataset. The original image is shown on the left, the visualization results of the MobileNet V2 model are shown in the middle, and the visualization results of the ELERNet model are shown on the right.

image, MobileNet V2 localization results, and ELERNet localization results are shown in Figure 13 (AWE) and Figure 14 (EARVN1.0). From the results, we can conclude that an essential prerequisite for making correct predictions is to take the ear's geometry as the most discriminative region, ignoring all distracting factors such as background and hair. First, we analyze the visualization results in Figure 13: (a) MobileNet V2 only pays attention to a piece of background hair features, ignoring the ear contour, which leads to wrong predictions. (b) Only pays attention to the upper half of the ear contour features. (c) The scope of attention is too large, and attention is paid to both ear studs and hair features. (d) The features of the middle part of the ear contour are ignored. (e) Excessive attention is paid to the hair-blocking part. (f) Only the earphone pendant is concerned, ignoring the ear's contours. (g) Focus only on distractor ear studs. (h) Focus too much on hair background features. (i) Focus only on earlobes and earplugs. (j) Focus

too much on earplugs and ignore ear contours. (k) Focus only on hair features. (l) Only focus on the earlobe part. Next, we analyze the visualization results in Figure 14: (a) only pay attention to the upper half of the ear contour. (b) Also, pay attention to the features of the hair occluded part. (c) Pay too much attention to the glasses frame. (d) Ignore the ear features in the upper part. (e) Pay attention to the incomplete ear contour under the premise of an auxiliary judgment of the occluder. (f) Pay too much attention to the features of the background part. (g) Pay attention to the facial features in the image. (h) Only focus on hair features, and ignore ear contours. (i) Focus on features such as windows in the background. (j) Focus only on earlobe features. (k) Focus too much on background features and occluder fingers. (l) Focus on hair features and decorations, ignoring ear features.

5. Conclusions

Most existing ear recognition methods are based on network models with high parameters and high model complexity. To address this problem, an efficient and lightweight human ear recognition method (ELERNet) based on MobileNet is proposed in this paper. We consider that the unconstrained human ear dataset has substantial intraclass and interclass differences, making feature extraction difficult. We introduce dynamic convolution decomposition and coordinate attention mechanism to enhance the model's feature robustness, learn discriminative ear features, and improve the recognition performance. Our method has been tested on both AWE and EARVN1.0, which are public unconstrained human ear datasets, and has achieved better recognition performance than the existing methods. Finally, using the GradCAM technology to explain our model performance visualization highlights that the model predicted results had a decisive impact area. According to the visualization results, we can conclude that the overall ear outline for predicting results is essential. At the same time, our model can be excellent for filtering out the background, earrings, earplugs, and hair, such as interference characteristics. Besides, illumination, angle, contrast, resolution, and other aspects have little influence on model performance, except in extreme cases. We will continue to optimize our approach for subsequent deployment to mobile devices or embedding it into small Linux systems. This will significantly aid identity confirmation in financial security, surveillance security, and other fields.

Data Availability

The data are available from the corresponding author upon reasonable request.

Conflicts of Interest

There is no conflict of interest regarding the publication of this paper.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 61673316); by the Scientific Research Project of Education Department of Shaanxi Province (21JK0921); by the Key Research and Development Projects of Shaanxi Province, under Grant No. 2017GY-071; by the Technical Innovation Guidance Special Project of Shaanxi Province, under Grant No. 2017XT-005; and by the Research Program of Xianyang City under Grant No. 2017 K01-25-3. Thanks are due to my teachers and classmates for giving me guidance on my studies.

References

[1] C. S. Hsiao, C. P. Fan, and Y. T. Hwang, "Design and analysis of deep-learning based iris recognition technologies by combination of U-Net and EfficientNet," in *2021 9th International Conference on Information and Education Technology (ICIET)*, pp. 433–437, Okayama, Japan, 2021.

[2] H. D. Rafik and M. Boubaker, "A multi biometric system based on the right iris and the left iris using the combination of convolutional neural networks," in *2020 Fourth International Conference On Intelligent Computing in Data Sciences (ICDS)*, pp. 1–10, Fez, Morocco, 2020.

[3] S. D. Shirke and C. Rajabhushnam, "Biometric personal iris recognition from an image at long distance," in *2019 3rd International Conference on Trends in Electronics and Informatics (ICOEI)*, pp. 560–565, Tirunelveli, India, 2019.

[4] M. Sahu and R. Dash, "Study on face recognition techniques," in *2020 International Conference on Communication and Signal Processing (ICCSP)*, pp. 613–616, Chennai, India, 2020.

[5] A. A. Sukmandhani and I. Sutedja, "Face recognition method for online exams," in *2019 International Conference on Information Management and Technology (ICIMTech)*, vol. 1, pp. 175–179, Jakarta/Bali, Indonesia, 2019.

[6] S. Shi, J. Cui, X. L. Zhang, Y. Liu, J. L. Gao, and Y. J. Wang, "Fingerprint recognition strategies based on a fuzzy commitment for cloud-assisted IoT: a minutiae-based sector coding approach," *IEEE Access*, vol. 7, pp. 44803–44812, 2019.

[7] I. Elzein and M. Kurdi, "Analysis of embedded fingerprint biometric recognition system algorithm," in *2021 12th International Symposium on Advanced Topics in Electrical Engineering (ATEE)*, pp. 1–4, Bucharest, Romania, 2021.

[8] M. H. Hersyah, D. Yolanda, and H. Sitohang, "Multiple Laboratory Authentication System Design Using Fingerprints Sensor and Keypad Based on Microcontroller," in *2020 International Conference on Information Technology Systems and Innovation (ICITSI)*, pp. 14–19, Bandung, Indonesia, 2020.

[9] E. M. Owaidah, K. S. Aloufi, and J. H. Alkhatib, "Gait recognition for Saudi Costume using Kinect skeletal tracking," in *2019 2nd international Conference on Computer Applications & Information Security (ICCAIS)*, pp. 1–5, Riyadh, Saudi Arabia, 2019.

[10] H. M. L. Aung and C. Pluempitiwiriyawej, "Gait biometric-based human recognition system using deep convolutional neural network in surveillance system," in *2020 Asia Conference on Computers and Communications (ACCC)*, pp. 47–51, Singapore, Singapore, 2020.

[11] R. Srivastva, A. Singh, and Y. N. Singh, "PlexNet: a fast and robust ECG biometric system for human recognition," *Information Sciences*, vol. 558, pp. 208–228, 2021.

[12] M. Wang, K. Kasmarik, A. Bezerianos, K. C. Tan, and H. Abbass, "On the channel density of EEG signals for reliable biometric recognition," *Pattern Recognition Letters*, vol. 147, pp. 134–141, 2021.

[13] W. Cui, Z. Wang, and Y. Li, "ECG-based biometric recognition under exercise and rest situations," *Biomedical Engineering Advances*, vol. 2, article 100008, 2021.

[14] R. Giorgi, N. Bettin, S. Ermini, F. Montefoschi, and A. Rizzo, "An iris+voice recognition system for a smart doorbell," in *2019 8th Mediterranean Conference on Embedded Computing (MECO)*, pp. 1–4, Budva, Montenegro, 2019.

[15] O. Tymchenko, B. Havrysh, O. O. Tymchenko, O. Khamula, B. Kovalskyi, and K. Havrysh, "Person voice recognition methods," in *2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP)*, pp. 287–290, Lviv, Ukraine, 2020.

[16] A. Bertillon, *La photographie judiciaire: avec un appendice sur la classification et l'identification anthropométriques*, Gauthier-Villars, París, 1890.

[17] A. Iannarelli, *Ear Identification*, Forensic Identification Series, 1989.

[18] L. Feenstra and C. Van der Lugt, "Ear witness," *The Journal of Laryngology & Otology*, vol. 114, no. 7, pp. 497–500, 2000.

[19] A. Pflug and C. Busch, "Ear biometrics: a survey of detection, feature extraction and recognition methods," *IET Biometrics*, vol. 1, no. 2, p. 114, 2012.

[20] Y. Tang, L. Zhang, F. Min, and J. He, "Multi-scale deep feature learning for human activity recognition using wearable sensors," *IEEE Transactions on Industrial Electronics*, vol. 70, no. 2, pp. 2106–2116, 2022.

[21] X. Cheng, L. Zhang, Y. Tang, Y. Liu, H. Wu, and J. He, "Real-time human activity recognition using conditionally parametrized convolutions on mobile and wearable devices," *IEEE Sensors Journal*, vol. 22, no. 6, pp. 5889–5901, 2022.

[22] W. Huang, L. Zhang, H. Wu, F. Min, and A. Song, "Channel-Equalization-HAR: a light-weight convolutional neural network for wearable sensor based human activity recognition," *IEEE Transactions on Mobile Computing*, p. 1, 2022.

[23] W. Gao, L. Zhang, W. Huang, F. Min, J. He, and A. Song, "Deep neural networks for sensor-based human activity recognition using selective kernel convolution," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–13, 2021.

[24] C. Han, L. Zhang, Y. Tang, W. Huang, F. Min, and J. He, "Human activity recognition using wearable sensors by heterogeneous convolutional neural networks," *Expert Systems with Applications*, vol. 198, article 116764, 2022.

[25] A. G. Howard, M. Zhu, B. Chen et al., "Mobilenets: efficient convolutional neural networks for mobile vision applications, Honolulu, HI, USA, 2017, arXiv preprint arXiv:1704.04861.

[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L. C. Chen, "Mobilenetv2: inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510–4520, Salt Lake City, UT, USA, 2018.

[27] A. Howard, M. Sandler, G. Chu et al., "Searching for mobilenetv3," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1314–1324, Seoul, Korea, 2019.

[28] M. F. Alsaffar, "Elevation of some biochemical and immunological parameters in hemodialysis patients suffering from hepatitis C virus infection in Babylon Province," *Indian Journal of Forensic Medicine & Toxicology*, vol. 15, no. 3, p. 2355, 2021.

[29] Y. Li, Y. Chen, X. Dai et al., "Revisiting dynamic convolution via matrix decomposition," 2021, arXiv preprint arXiv:2103.08756.

[30] Q. Hou, D. Zhou, and J. Feng, "Coordinate attention for efficient mobile network design," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13713–13722, 2021.

[31] Ž. Emeršič, B. Meden, P. Peer, and V. Štruc, "Evaluation and analysis of ear recognition models: performance, complexity and resource requirements," *Neural Computing and Applications*, vol. 32, no. 20, pp. 15785–15800, 2020.

[32] Ž. Emeršič, V. Štruc, and P. Peer, "Ear recognition: more than a survey," *Neurocomputing*, vol. 255, pp. 26–39, 2017.

[33] Ž. Emeršič, L. L. Gabriel, V. Štruc, and P. Peer, "Convolutional encoder–decoder networks for pixel-wise ear detection and segmentation," *IET Biometrics*, vol. 7, no. 3, pp. 175–184, 2018.

[34] V. T. Hoang, "EarVN1.0: a new large-scale ear images dataset in the wild," *Data in Brief*, vol. 27, article 104630, 2019.

[35] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: visual explanations from deep networks via gradient-based localization," in *Proceedings of the IEEE international conference on computer vision*, pp. 618–626, Venice, Italy, 2017.

[36] J. D. Bustard and M. S. Nixon, "Toward unconstrained ear recognition from two-dimensional images," *IEEE transactions on systems, man, and cybernetics-Part A: Systems and Humans*, vol. 40, no. 3, pp. 486–494, 2010.

[37] B. Arbab-Zavar and M. S. Nixon, "On guided model-based analysis for ear biometrics," *Computer Vision and Image Understanding*, vol. 115, no. 4, pp. 487–502, 2011.

[38] T. S. Chan and A. Kumar, "Reliable ear identification using 2-D quadrature filters," *Pattern Recognition Letters*, vol. 33, no. 14, pp. 1870–1881, 2012.

[39] N. B. Boodoo-Jahangeer and S. Baichoo, "LBP-based ear recognition," in *IEEE international conference on bioinformatics and bioengineering*, pp. 1–4, Chania, Greece, 2013.

[40] M. Hassaballah, H. A. Alshazly, and H. A. Ali, "Ear recognition using local binary patterns: a comparative experimental study," *Expert Systems with Applications*, vol. 118, no. 14, pp. 182–200, 2019.

[41] M. Hassaballah, H. A. Alshazly, and H. A. Ali, "Robust local oriented patterns for ear recognition," *Multimedia Tools and Applications*, vol. 41, no. 78, pp. 31183–31204, 2020.

[42] S. Sajadi and A. Fathi, "Genetic algorithm based local and global spectral features extraction for ear recognition," *Expert Systems with Applications*, vol. 159, article 113639, 2020.

[43] Y. Zhang, Z. Mu, L. Yuan, and C. Yu, "Ear verification under uncontrolled conditions with convolutional neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 185–198, 2018.

[44] Ž. Emeršič, J. Križaj, V. Štruc, and P. Peer, "Deep ear recognition pipeline," *Recent Advances in Computer Vision*, vol. 804, pp. 333–362, 2019.

[45] H. Alshazly, C. Linse, E. Barth, and T. Martinetz, "Deep convolutional neural networks for unconstrained ear recognition," *IEEE Access*, vol. 8, pp. 170295–170310, 2020.

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, 2012.

[47] X. Zhang, X. Zhou, M. Lin, and J. Sun, "Shufflenet: an extremely efficient convolutional neural network for mobile devices," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 6848–6856, Salt Lake City, UT, USA, 2018.

[48] N. Ma, X. Zhang, H. T. Zheng, and J. Sun, "Shufflenet v2: practical guidelines for efficient cnn architecture design," in *Proceedings of the European conference on computer vision (ECCV)*, pp. 116–131, Munich, Germany, 2018.

[49] S. Dodge, J. Mounsef, and L. Karam, "Unconstrained ear recognition using deep neural networks," *IET Biometrics*, vol. 7, no. 3, pp. 207–214, 2018.

[50] Ž. Emeršič, D. Štepec, V. Štruc, and P. Peer, "Training convolutional neural networks with limited training data for ear recognition in the wild," 2017, arXiv preprint arXiv:1711.09952.

[51] Y. Khaldi and A. Benzaoui, "A new framework for grayscale ear images recognition using generative adversarial networks under unconstrained conditions," *Evolving Systems*, vol. 12, no. 4, pp. 923–934, 2021.

[52] Y. Khaldi and A. Benzaoui, "Region of interest synthesis using image-to-image translation for ear recognition," in *2020 International Conference on Advanced Aspects of Software Engineering (ICAASE)*, pp. 1–6, Constantine, Algeria, 2020.

[53] Y. Khaldi, A. Benzaoui, A. Ouahabi, S. Jacques, and A. Taleb-Ahmed, "Ear recognition based on deep unsupervised active learning," *IEEE Sensors Journal*, vol. 21, no. 18, pp. 20704–20713, 2021.

[54] Z. Wang, X. Gao, J. Yang, Q. Yan, and Y. Zhang, "Local feature fusion and SRC-based decision fusion for ear recognition," *Multimedia Systems*, vol. 28, no. 3, pp. 1117–1134, 2022.

[55] H. Alshazly, C. Linse, E. Barth, S. A. Idris, and T. Martinetz, "Towards explainable ear recognition systems using deep residual networks," *IEEE Access*, vol. 9, pp. 122254–122273, 2021.

[56] A. Korichi, S. Slatnia, and O. Aiadi, "TR-ICANet: a fast unsupervised deep-learning-based scheme for unconstrained ear recognition," *Arabian Journal for Science and Engineering*, vol. 47, no. 8, pp. 9887–9898, 2022.

[57] I. Omara, A. Hagag, G. Ma, F. E. Abd el-Samie, and E. Song, "A novel approach for ear recognition: learning Mahalanobis distance features from deep CNNs," *Machine Vision and Applications*, vol. 32, no. 1, pp. 1–14, 2021.

[58] M. Regouid, M. Touahria, M. Benouis, L. Mostefai, and I. Lamiche, "Comparative study of 1D-local descriptors for ear biometric system," *Multimedia Tools and Applications*, vol. 81, pp. 1–27, 2022.

[59] U. Kacar and M. Kirci, "ScoreNet: deep cascade score level fusion for unconstrained ear recognition," *IET Biometrics*, vol. 8, no. 2, pp. 109–120, 2019.

[60] I. Omara, H. Zhang, F. Wang, A. Hagag, X. Li, and W. Zuo, "Metric learning with dynamically generated pairwise constraints for ear recognition," *Information*, vol. 9, no. 9, p. 215, 2018.

[61] E. E. Hansley, M. P. Segundo, and S. Sarkar, "Employing fusion of learned and handcrafted features for unconstrained ear recognition," *IET Biometrics*, vol. 7, no. 3, pp. 215–223, 2018.

[62] O. Aiadi, B. Khaldi, and C. Saadeddine, "MDFNet: an unsupervised lightweight network for ear print recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–14, 2022.

[63] H. K. Mewada, A. V. Patel, J. Chaudhari, K. Mahant, and A. Vala, "Wavelet features embedded convolutional neural network for multiscale ear recognition," *Journal of Electronic Imaging*, vol. 29, no. 4, article 43029, 2020.

[64] S. Ramos-Cooper, E. Gomez-Nieto, and G. Camara-Chavez, "VGGFace-Ear: an extended dataset for unconstrained ear recognition," *Sensors*, vol. 22, no. 5, p. 1752, 2022.

[65] M. B. Alejo, "Unconstrained ear recognition using transformers," *Jordanian Journal of Computers and Information Technology*, vol. 7, no. 4, pp. 326–336, 2021.