Computer Modeling in
Engineering & Sciences

Tech Science Press

DOI: 10.32604/cmes.2024.049174

ARTICLE

A Hand Features Based Fusion Recognition Network with Enhancing Multi-Modal Correlation

Wei Wu*, **Yuan Zhang, Yunpeng Li, Chuanyang Li and Yan Hao**

School of Information Engineering, Shenyang University, Shenyang, 110044, China

*Corresponding Author: Wei Wu. Email: wuwei429@163.com

Received: 29 December 2023 Accepted: 06 February 2024 Published: 16 April 2024

ABSTRACT

Fusing hand-based features in multi-modal biometric recognition enhances anti-spoofing capabilities. Additionally, it leverages inter-modal correlation to enhance recognition performance. Concurrently, the robustness and recognition performance of the system can be enhanced through judiciously leveraging the correlation among multimodal features. Nevertheless, two issues persist in multi-modal feature fusion recognition: Firstly, the enhancement of recognition performance in fusion recognition has not comprehensively considered the inter-modality correlations among distinct modalities. Secondly, during modal fusion, improper weight selection diminishes the salience of crucial modal features, thereby diminishing the overall recognition performance. To address these two issues, we introduce an enhanced DenseNet multimodal recognition network founded on feature-level fusion. The information from the three modalities is fused akin to RGB, and the input network augments the correlation between modes through channel correlation. Within the enhanced DenseNet network, the Efficient Channel Attention Network (ECA-Net) dynamically adjusts the weight of each channel to amplify the salience of crucial information in each modal feature. Depthwise separable convolution markedly reduces the training parameters and further enhances the feature correlation. Experimental evaluations were conducted on four multimodal databases, comprising six unimodal databases, including multispectral palmprint and palm vein databases from the Chinese Academy of Sciences. The Equal Error Rates (EER) values were 0.0149%, 0.0150%, 0.0099%, and 0.0050%, correspondingly. In comparison to other network methods for palmprint, palm vein, and finger vein fusion recognition, this approach substantially enhances recognition performance, rendering it suitable for high-security environments with practical applicability. The experiments in this article utilized a modest sample database comprising 200 individuals. The subsequent phase involves preparing for the extension of the method to larger databases.

KEYWORDS

Biometrics; multi-modal; correlation; deep learning; feature-level fusion

1 Introduction

Attack methods targeting single-modal biometric identification are diverse and ceaseless, encompassing techniques such as 3D face replicas and sets of manipulated fingerprints. Multimodal fusion biometric recognition inherently possesses advantages in countering attacks while alleviating the demands on the performance of unimodal biometric recognition. Comprehensive utilization of complementary information from multimodal sources enriches the feature representation in multimodal fusion, substantially mitigating the impact of environmental variations on recognition performance, particularly in noisy or extreme conditions, thereby enhancing the robustness of the system.

Before the rise of multimodal biometrics, unimodal biometrics were already deeply integrated into everyday life, such as face recognition. Benamara et al. [1] introduced a multi-sensor face detector model built upon the advanced YOLO (You Only Look Once) v3 architecture. This model demonstrates the capability to detect faces captured in both visible and thermal images. To mitigate the modal gap between visible and thermal spectra, a novel CycleGAN structure is applied. Zheng et al. [2] proposed a converter-based cross-fertilization method that effectively integrates facial and image features to enhance focus on salient facial regions. Jiménez-Bravo et al. [3] designed a cost-effective face recognition system, incorporating techniques like edge computing and data augmentation. Sun et al. [4] employed global features learned from a global-based module and local features learned from a region-based module to create a representation of fused face features.

At this stage, the use of multimodality to accomplish tasks is beginning to rise in all industries. Especially for the field of biometric identification. Shi et al. [5] integrated a two-branch convolutional neural network with a Long Short-Term Memory (LSTM) network to extract spatio-temporal feature information from the original load pattern of network traffic. Additionally, a convolutional neural network was employed to extract feature information for multimodal information mining. Han et al. [6] employed an original traffic feature extraction method to reduce redundant features and expedite neural network convergence. Chen et al. [7] incorporated designed spatial context information to dynamically modulate and filter image features, thereby enhancing the dependency modeling of image tagging and improving the model's inference capabilities.

Wu et al. [8] utilized a deep hash network to extract binary templates for palm print and palm vein features, followed by fractional level fusion. Oldal et al. [9] conducted key point detection and main line extraction for hand geometry features and palm print features, recognizing them through template-based matching to detect corresponding points in palm print images. Ramachandran et al. [10] employed Log-Gabor transform, Histogram of Oriented Gradients (HOG), and Local Binary Pattern (LBP) to extract features from palmprint and iris images, concluding with score-level fusion.

In this paper, an improved DenseNet multimodal recognition network based on feature-level fusion is proposed to address the above two problems. The three-modal information is preprocessed and input into the RGB three-color channel for RGB-like fusion. And the inter-modal correlation is enhanced with channel correlation. Subsequently, the DenseNet network is improved. The weights of each channel are dynamically adjusted with the Efficient Channel Attention Network (ECA-Net). By adaptively choosing a convolution kernel size proportional to the channel dimension, the information interaction across channels is accomplished with convolution operation. The correlation information among the three modalities is fully utilized to improve the prominence of important information in the features of each modality. Meanwhile, the depth separable convolution drastically reduces the training parameters, avoids network overfitting, and again improves the feature correlation to ensure strong generalization ability in small sample data sets. In order to prove the effectiveness of the network, this paper selects palm prints, palm veins and finger veins as three modalities for fusion experiments. Fig. 1 shows the overall architecture of the proposed network in this paper. The main contributions of the paper are:

i. This study makes a notable contribution by addressing the frequently neglected inter-modality correlations in multi-modal feature fusion recognition. The proposed enhanced DenseNet network capitalizes on these correlations, thereby improving the robustness and overall performance of the recognition system.

ii. The article presents a solution to the issue of improper weight selection during modal fusion, a common challenge that can lead to diminished recognition performance. The Enhanced DenseNet network integrates the ECA-Net, dynamically adjusting weights to amplify crucial modal features, thereby optimizing fusion and enhancing overall recognition performance.

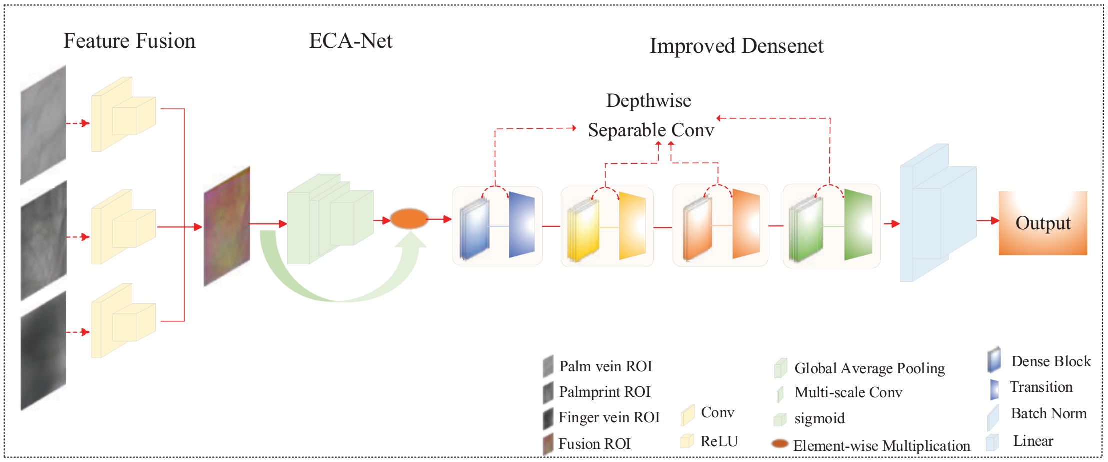


Figure 1: Network structural diagram

2 Related Work

Although multimodal biometrics has started to gradually enter our lives. Nonetheless, multimodal biometric recognition generally encounters the following two issues. Firstly, modality fusion fails to fully consider the correlation between modes, hindering the improvement of recognition performance. Secondly, improper weighting during modality fusion diminishes the significance of crucial modal features, resulting in an overall decrease in recognition performance.

Within current multimodal feature fusion methods, Yang et al. [11] employed a single device to capture multimodal data from the same hand area. To extract pertinent information, fingerprint and finger vein features are decomposed into shared and private components, enhancing complementarity. Rajasekar et al. [12] introduced a multimodal biometric recognition system based on deep learning methodologies. The approach incorporates three Convolutional Neural Network (CNN) architectures for iris, face, and fingerprint to integrate and construct the system. Additionally, it adopts a two-level fusion strategy involving feature-level fusion and score-level fusion. Daas et al. [13] proposed two multimodal architectures, feature-level fusion, and fractional-level fusion, based on the biometric features of finger-knuckleprint and finger-vein. Transfer learning with CNN is employed for feature extraction to bolster system security. Sasikala [14], the integration of two biometric features, namely fingerprint and retina, involves a combination of deep learning (DL) and hashing methods. A bidirectional gated recurrent unit (BiGRU) model is employed to discern the correlation among internal features within single-modal images. While some methods in the aforementioned context capture multimodal data from the same hand area, and others detect the correlation of internal features within single-modal images, they fall short of fully leveraging the correlation information between different modalities.

In [15], the Improved Principal Component Analysis (IPCA) method was employed for the extraction and dimensionality reduction of four modal databases, encompassing face, ear, palm, and finger. The four biometric features were fused with corresponding matching scores, and the resultant features were sorted. Despite the traditional method enhancing the salience of crucial features, achieving improved recognition performance remains challenging, especially when dealing with large-scale databases. In [16], a hybrid fusion model was introduced for iris, palm vein, and finger vein modalities. This model incorporates a multi-set structure to capture typical features, and the distribution information of scores is utilized to aid decision-making, thereby enhancing recognition accuracy and security. Abdullahi et al. [17], a spatial and temporal multimodal fingerprint and finger vein network, named FS-STMFPFV-Net, was introduced based on fingerprint and finger vein modalities. Image variability is enhanced by independently learning two channels, and ReliefFS is employed for feature selection. In [18], a Two-Stream Convolutional Neural Network was advocated to augment the information pertaining to Multimodal Facial Biometrics. The architecture comprises two successive components, each employing distinct fusion strategies, amalgamating three-color data and multimodal face biometric texture descriptors. To augment the prominence of crucial features, certain methodologies from the aforementioned employ multistage pretraining for refined feature weight selection, while others intensify the layers of the feature extraction network. However, collectively, these approaches amplify the intricacy of the network model.

In this research, we present an enhanced DenseNet multimodal recognition network that addresses critical issues related to insufficient consideration of inter-modality correlations and improper weight selection during fusion. The proposed network incorporates Efficient Channel Attention and depth-wise separable convolution, leading to a notable improvement in recognition performance.

This paper is summarized as follows. Section 3 describes the derivation of the methods mentioned in this article. Section 4 creates four multimodal databases and performs performance experiments on the method. Section 5 summarizes the article.

3 Method

During the image data preprocessing stage, we utilize the methodology outlined by Wu et al. [19] to implement denoising techniques on palmprint and palm vein images. This involves contour extraction, identification of intersection and valley points, resulting in the acquisition of regions of interest (ROI) images sized at $128\,pixels \times 128\,pixels$. Adhering to the methodology elucidated by Krishnan et al. [20], denoising procedures were implemented on the finger vein image, eliminating border elements, identifying finger edges, and demarcating the corresponding region. Subsequently, an ROI image sized at $128\,pixels \times 128\,pixels$ was derived.

The ROIs of palmprint, palm vein, and finger vein are individually fed into the RGB three-color channels, resulting in a three-modal RGB feature map denoted as X, possessing dimensions of $3 \times 128 \times 128$. The dimension of X is denoted as $C \times H \times W$, where C signifies the number of channels, and H and W symbolize the height and width of the feature map, respectively. Initially, a global average pooling operation is executed on X to generate a feature map incorporating global average information from the three modes, as illustrated in Eq. (1),

$$Y_c = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} X_{cij} \tag{1}$$

In this context, Y_c signifies the globally averaged pooled output on channel c. An adaptive approach is employed to ascertain the size (k) of the convolution kernel, which scales proportionally with the channel dimension C. This is depicted in Eq. (2),

$$k = \Psi(C) = \left|\frac{\log_2(C)}{\gamma} + \frac{b}{\gamma}\right|_{odd} \tag{2}$$

γ and b are the respective parameters of the mapping function. Additionally, $|n|$ *odd* represents the nearest odd number to n. This approach enables dynamic adjustments to the interaction coverage among diverse modalities based on the channel count. It guarantees comprehensive consideration of the correlation among the three modalities, as illustrated in Eq. (3),

$$Z_c = K * Y_c \tag{3}$$

Within this context, Z_c denotes the convolution output specific to channel c, with K representing the convolution kernel. Following the convolution operation, the resulting output undergoes processing by the activation function σ, as outlined in Eq. (4),

$$O_c = \sigma(Z_c) \tag{4}$$

Ultimately, the feature map O_c, following processing, is element-wise multiplied with the original input three-modal feature map X to yield the ultimate feature map. This methodology allows ECA-Net to fully exploit the correlation information among the three-modal features. The detailed process is illustrated in Fig. 2.

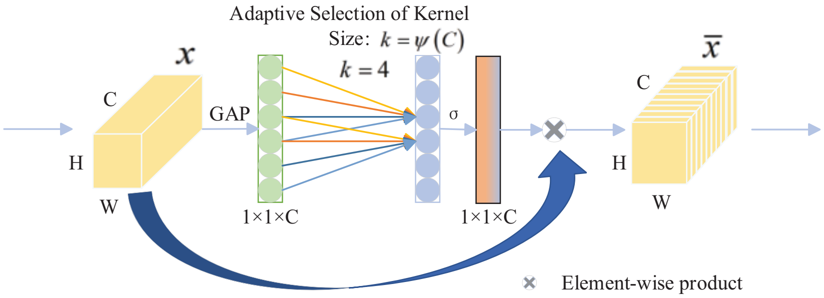


Figure 2: Efficient channel attention module

In the preceding steps, ECA-Net dynamically determines the size (k) of the convolution kernel using an adaptive method. This size delineates the scope of local cross-channel interaction. By performing convolution operations on each channel with k neighbors, local cross-channel interaction information is acquired. Subsequently, an element-wise multiplication operation is executed to fully leverage the correlation information among channels, thereby enhancing the representational capacity of the three-modal feature map. The representation of the process when k is set to 4 is depicted in Fig. 2.

Subsequently, to facilitate a more comprehensive learning of the employed trimodal features and mitigate the risk of overfitting associated with a small sample database, enhancements are applied to the Dense Block module and the Transition module through the integration of deep separable convolution. Leveraging the densely connected nature inherent in the Dense Block module enables the network to more effectively recycle features learned in earlier layers. This enhances the efficiency of employing trimodal features. Robust generalization capability is guaranteed, particularly on small sample databases. Concurrently, depth separable convolution further fortifies the network's feature representation capability, leading to a substantial reduction in the number of parameters. This renders the network more adaptive and robust. The core architecture of this network is depicted in Fig. 3.

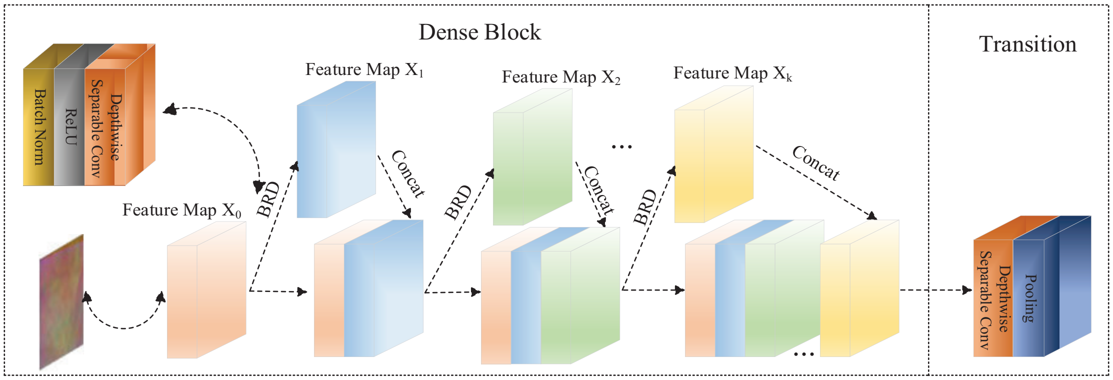


Figure 3: Structure diagram of proposed method network

The Dense Block module within this network stands as a pivotal constituent of the entire architecture. Comprising numerous dense layers, each layer is intricately linked through cross-channel concatenation. This thoughtful design ensures that each layer accepts the features from all preceding layers as input and transmits its feature output to all ensuing layers. Thus, this approach ensures the comprehensive utilization of information across the three modalities.

Within this network, the Dense Block module stands as a fundamental constituent and comprises several dense layers. These layers are connected to each other by cross-channels. Such a design permits each layer to accept features from all preceding layers as input and relay the feature output from this layer to all succeeding layers. Consequently, this approach ensures the comprehensive utilization of information among the three modalities.

In detail, the network incorporates four Dense Block modules, with each module comprising a specific number of layers (3, 6, 12, 8). The input to the ith layer of each module encompasses a cascade of feature maps from the initial layer up to layer i − 1. This implies that the input to layer i encompasses all features from layer 1 to layer i − 1. This relationship is depicted in Eq. (5),

$$x_i = H_i\left([x_0, x_1, \ldots, x_{i-1}]\right) \tag{5}$$

Here, x_i represents the output feature map of the ith layer, while H_i denotes the mapping function associated with the ith layer. The mapping function H_i receives the input feature maps $[x_0, x_1, \ldots, x_{i-1}]$ from all preceding layers and produces the output x_i for the ith layer. The notation $[x_0, x_1, \ldots, x_{i-1}]$ indicates the concatenation of feature maps from all preceding layers, resulting in a tensor that consolidates all the features of the prior layers.

The Transition layer serves as the linkage between two neighboring Dense Blocks. The primary role of the Transition layer is to decrease the dimension of the feature map. This, in turn, reduces the computational complexity of the model, making it easier to train the network and establish effective transitions between adjacent Dense Blocks. The Transition layer consists of a 1 × 1 convolutional layer and a 2 × 2 average pooling layer. Additionally, deep separable convolution decomposes the convolution operation into deep convolution and pointwise convolution. This results in a substantial reduction in the number of parameters. This aspect is especially crucial for small sample databases as it effectively mitigates the model's reliance on an extensive amount of training data, consequently minimizing the risk of overfitting.

In this network, the depthwise separable convolution can be deconstructed into two distinct steps: Depthwise Convolution and Pointwise Convolution.

Within the realm of depthwise convolution, a convolutional kernel is individually applied to each channel of the input. Given that the input comprises three channels, three 3 × 3 convolution kernels are employed to process these channels. This step ensures the independence of each channel without intermixing among channels. In pointwise convolution, a 1 × 1 convolution kernel linearly combines the output of depthwise convolution. This step is responsible for the linear combination between channels without considering spatial information. The depthwise separable convolution disentangles channel information and spatial information within the input data. This segregation leads to a reduction in computation workload and the number of parameters. The detailed architecture is illustrated in Fig. 4.

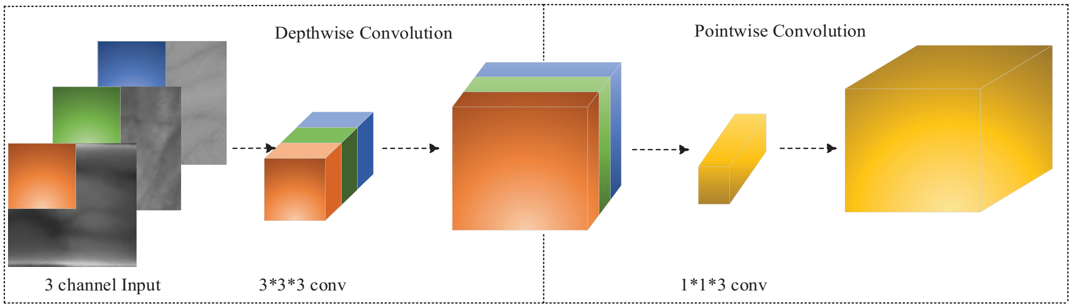


Figure 4: Depth separable convolution structure diagram

Fig. 4 depicts the initial step as the depthwise convolution operation. During this stage, the process unfolds in a two-dimensional plane. Each channel undergoes processing by a depthwise convolution kernel, resulting in a corresponding feature map. The number of depthwise convolution kernels corresponds to the channels in the preceding layer. In the case of a three-channel image input, the convolution produces only three feature maps, without an increase in their count. This ensures clarity and simplicity in feature map production.

Subsequently, the pointwise convolution stage amalgamates the feature maps acquired through deep convolution. The convolution kernel size is 1 × 1 × 3, with 3 representing the number of channels in the upper layer. Pointwise convolution serves the purpose of weighting and amalgamating feature maps along the depth direction, leading to the generation of novel feature maps. Each convolutional kernel yields an output feature map. By combining deep convolution and pointwise convolution, deep separable convolution achieves flexibility in processing feature information from various channels at the same spatial position, all the while preserving computational efficiency. This results in a more nuanced and specific feature representation.

Within this network, improvements are introduced to the Dense Block module and the Transition layer, achieved by incorporating deep separable convolution. The network's training performance is further enhanced through the design of a specific architecture with adjusted parameters. This design ensures optimized configurations for effective database training. The configurations include a growth rate set to 16, utilization of four Dense Block modules, and each block having a specified number of layers (3, 6, 12, 8). The initial convolution layer learns eight filters, and the batch normalization size is configured to 4. The choice of these parameters undergoes meticulous consideration and experimental verification to guarantee that the network attains commendable performance in specific tasks. The detailed parameters are presented in Table 1.

Table 1: The network architecture of the proposed method model

Output size	Proposed method
[3,28,28]	Dense block 1: $\begin{bmatrix} \text{Dense Layer}: \ \text{Depthwise\_Conv}.3 \times 3, 64 \\ \text{Conv}.1 \times 1, 64\ \text{Pointwise\_Conv}.1 \times 1, 16 \end{bmatrix} \times 3$ Transition: $\begin{bmatrix} \text{Depthwise\_Conv}.3 \times 3, 56 \\ \text{Pointwise\_Conv}.1 \times 1, 28 \end{bmatrix}$
[3,62,62]	Dense block 2: $\begin{bmatrix} \text{Dense Layer}: \ \text{Depthwise\_Conv}.3 \times 3, 64 \\ \text{Conv}.1 \times 1, 64\ \text{Pointwise\_Conv}.1 \times 1, 16 \end{bmatrix} \times 6$ Transition: $\begin{bmatrix} \text{Depthwise\_Conv}.3 \times 3, 124 \\ \text{Pointwise\_Conv}.1 \times 1, 62 \end{bmatrix}$
[3,127,127]	Dense block 3: $\begin{bmatrix} \text{Dense Layer}: \ \text{Depthwise\_Conv}.3 \times 3, 64 \\ \text{Conv}.1 \times 1, 64\ \text{Pointwise\_Conv}.1 \times 1, 16 \end{bmatrix} \times 12$ Transition: $\begin{bmatrix} \text{Depthwise\_Conv}.3 \times 3, 254 \\ \text{Pointwise\_Conv}.1 \times 1, 127 \end{bmatrix}$
[3,16,16]	Dense block 4: $\begin{bmatrix} \text{Dense Layer}: \ \text{Depthwise\_Conv}.3 \times 3, 64 \\ \text{Conv}.1 \times 1, 64\ \text{Pointwise\_Conv}.1 \times 1, 16 \end{bmatrix} \times 8$ Transition: $\begin{bmatrix} \text{Depthwise\_Conv}.3 \times 3, 64 \\ \text{Pointwise\_Conv}.1 \times 1, 16 \end{bmatrix}$

4 Experimental Results and Analysis

4.1 Database

Following statistical analysis, it was observed that there is currently no publicly available database encompassing various hand-based features simultaneously, such as palmprint, palm vein, fingerprint, knuckleprint, finger vein, and hand shape, all belonging to the same individual. Consequently, this paper employs two single-modal palm vein databases, two single-modal palmprint databases, and a finger vein database to establish four multimodal databases centered around hand features. Detailed information regarding the single-modal databases is outlined in Table 2. Sample images from the five single-modal databases are depicted in Fig. 5.

Table 2: Hand-based single-modal database description

Databases	Traits	Subject	Sample	Total
CASIA-P (Zhou et al. [21])	Palmprint	200	6	1200
Tongji-V (Zhang et al. [22])	Palm-vein	600	20	12000
Tongji-P (Zhang et al. [22])	Palmprint	600	20	12000
PolyU-NIR (Zhang et al. [23])	Palm-vein	250	6	1500
SDUMLA-HMT (Yin et al. [24])	Finger-vein	106	18	1908

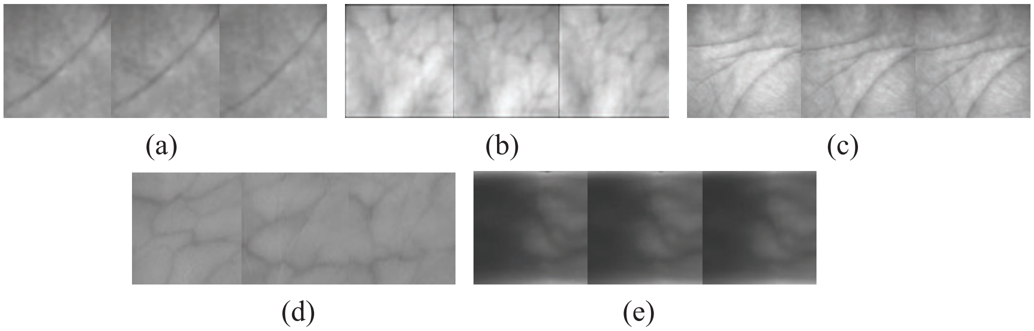


Figure 5: ROI samples of single-modal database. (a) CASIA-P; (b) Tongji-V; (c) Tongji-P; (d) Poly U-NIR; (e) SDUMLA-HMT

The multimodal database TCSD includes three publicly available single-modal databases: Tongji-V, CASIA-P, and SDUMLA-HMT. The Tongji University Palm Vein database (Tongji-V) collects palm vein images in a non-contact manner with a light source wavelength of 940 nm. It comprises 12,000 palm vein image samples from 600 individuals aged between 20 and 50 years. The Chinese Academy of Sciences Palmprint database (CASIA-P) is derived from the CASIA Multispectral Palmprint database. It captures palmprint features using a light source with a wavelength of 460 nm, including left and right hand features of the same individual. In this paper, these are treated as features of two different individuals, resulting in 1,200 palmprint image samples from 200 individuals. The Shandong University Machine Learning and Data Mining Laboratory Finger Vein database (SDUMLA-HMT) provides finger vein images for the index, middle, and ring fingers of both hands for each individual. Similar to CASIA-P, left and right hand features of the same individual, as well as finger vein images from different fingers of the same hand, are treated as features of different individuals. Therefore, the database comprises 3,816 finger vein image samples from 636 individuals. For each of the three databases, 200 individuals were selected, resulting in a total of 600 images, and the ROI size is $128\,pixels \times 128\,pixels$.

The multimodal database NIR-TSD comprises three publicly available single-modal databases: PolyU-NIR, Tongji-P, and SDUMLA-HMT. The PolyU Multispectral Palmprint database (PolyU-NIR) collects palmprint images under near-infrared illumination. The acquisition equipment includes a CCD camera and a high-power halogen light source for contact-based collection. The near-infrared illumination is used to capture palm vein images. This database contains 1,500 palm vein image samples from 250 individuals. The Tongji University Palmprint database (Tongji-P) collects palmprint images through contact-based acquisition and includes 12,000 palmprint image samples from 600 individuals. From the mentioned databases, including SDUMLA-HMT, 200 individuals were selected from each, resulting in a total of 600 image samples. The ROI size is $128\,pixels \times 128\,pixels$.

The multimodal database NIR-CSD consists of three publicly available single-modal databases: PolyU-NIR, CASIA-P, and SDUMLA-HMT. 200 individuals were selected from each of these three databases, resulting in a total of 600 image samples. The ROI size is $128\,pixels \times 128\,pixels$.

The multimodal database TTSD comprises three publicly available single-modal databases: Tongji-V, Tongji-P, and SDUMLA-HMT. A total of 600 images were selected, with 200 individuals chosen from each of the three databases. The ROI size is $128\,pixels \times 128\,pixels$. Table 3 presents detailed information about the utilized multimodal databases, and Fig. 6 displays sample images from the four multimodal databases.

Table 3: Hand-based multimodal database description

Databases	Subject	Sample	Total	Image size
TCSD	200	6	1200	128 × 128
NIR-TSD	200	6	1200	128 × 128
NIR-CSD	200	6	1200	128 × 128
TTSD	200	6	1200	128 × 128

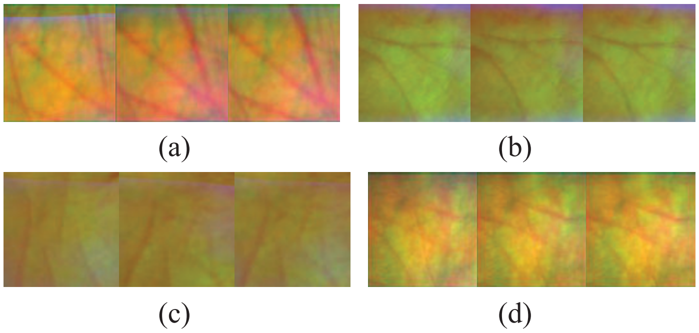

(a) (b) (c) (d)

Figure 6: Multimodal database ROI samples. (a) TCSD; (b) NIR-TSD; (c) NIR-CSD; (d) TTSD

4.2 Performance Indicators

To evaluate the accuracy of the proposed multimodal biometric recognition method, this study divided the data into training and validation sets in a ratio of 8:2 on four multimodal databases. The individuals in the training and validation sets are non-overlapping.

The recognition performance is assessed using metrics such as True Positive Rate (TPR), False Positive Rate (FPR), EER, Recall, Precision, Accuracy curves, Loss curves, Macro Precision-Recall (PR) curves, and Receiver Operating Characteristic (ROC) curves, as outlined in Eqs. (6)–(9),

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

$$FPR = \frac{FP}{FP + TN} \tag{7}$$

$$Precision = \frac{TP}{TP + FP} \tag{8}$$

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{9}$$

where TP represents the number of True Positives, TN represents the number of True Negatives, FP represents the number of False Positives, and FN represents the number of False Negatives.

The ROC curve is used to assess the classification performance of a binary classi-fication model at different thresholds. The Macro PR curve is a way to evaluate the performance of a model in multi-class problems. It aggregates the predictive results of multiple classes into a binary classification problem and then calculates the macro Precision and macro Recall, providing a comprehensive performance assessment.

4.3 Ablation Experiments

For the proposed multi-modal fusion recognition method in this paper, experimental verification is conducted from two aspects: modal ablation and network module ablation.

4.3.1 Modal Ablation

In modal ablation experiments, the proposed method is applied to train and recognize networks in single-modal, dual-modal, and triple-modal forms. During single-modal training, 1,200 palm vein image samples from 200 individuals in the PolyU-NIR database are used for training and validation. For dual-modal training, a combination of 1200 image samples from 200 individuals in both the PolyU-NIR and Tongji-P databases is used for training and validation after fusion. In the case of triple-modal training, 1,200 image samples from 200 individuals in the TTSD database are utilized for training and validation. Figs. 7 and 8 show the intuitive performance in terms of accuracy and loss across the three modalities. It can be observed from the figures that with an increase in modalities, the accuracy and loss curves of the validation set fit more quickly and approach the optimal values.

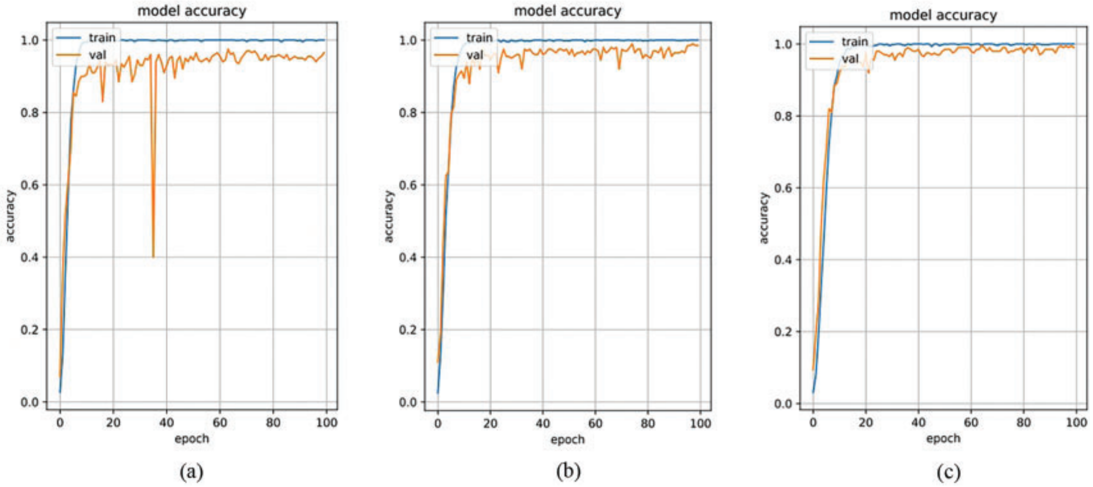


Figure 7: Three modal training/verification accuracy curves. (a) Single-mode accuracy; (b) Dual-modal accuracy; (c) Three-modal accuracy

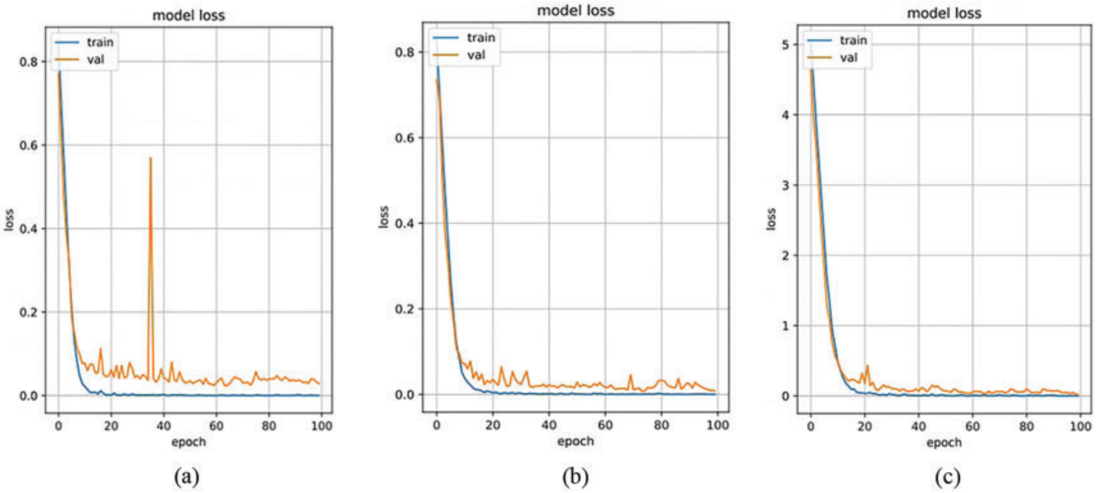


Figure 8: Three modal loss rate curves. (a) Single-mode loss rate; (b) Dual-modal loss rate; (c) Three-modal loss rate

4.3.2 Network Module Ablation

In the network module ablation experiments, the TTSD database, comprising 1200 images from 200 individuals, was used for training and validation. Performance tests were conducted on the original DenseNet program, the program with only the addition of ECA-Net, the program with only the introduction of depth separable convolution, and the completely improved program. Figs. 9 and 10 intuitively demonstrate the program's performance in terms of accuracy and loss under these four conditions. From the figures, it can be observed that the proposed improvements are effective, leading to enhanced performance in each scenario compared to the original program. The accuracy and loss curves of the validation set for the fully improved program fit more quickly and approach optimal values.

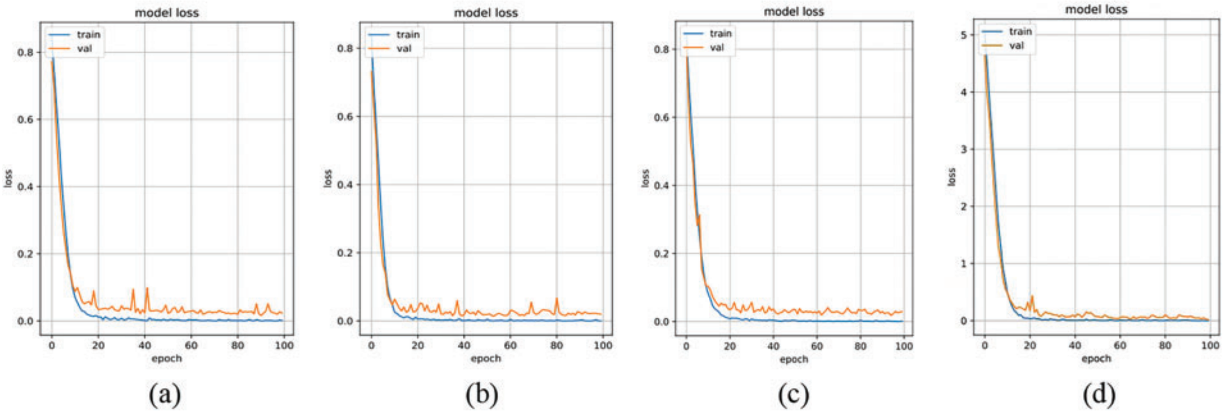
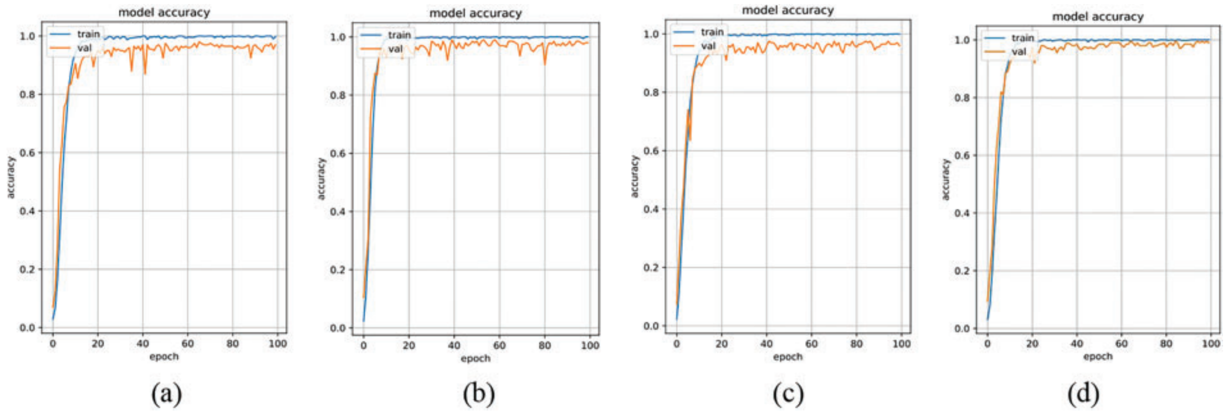


Figure 9: Program accuracy in four cases. (a) DenseNet; (b) ECA-DenseNet; (c) DW-conv; (d) Proposed method



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

Figure 10: Program loss rate in four cases. (a) DenseNet; (b) ECA-DenseNet; (c) DW-conv; (d) Proposed method

4.4 Recognition Performance

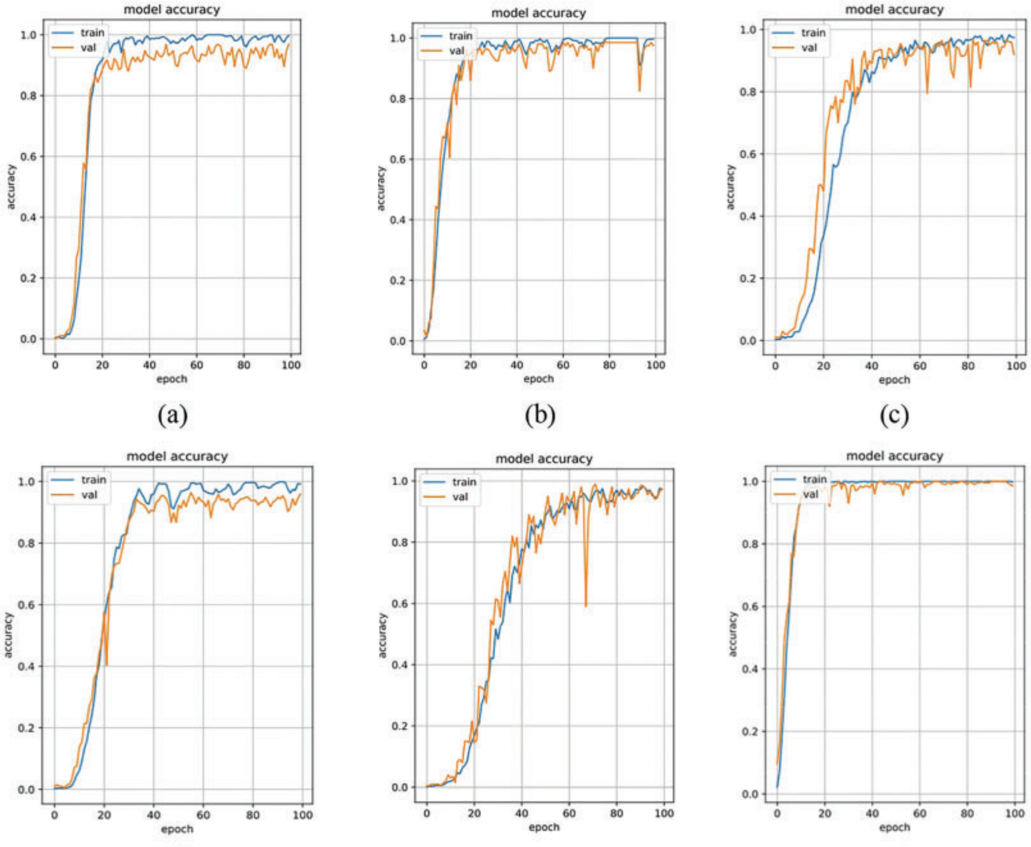
To validate the proposed network approach, experiments on recognition performance were conducted on a multimodal database. Several classical network models in the field of biometric recognition were selected for comparison with the proposed method. These models include VGG16 (Aleidan et al. [25]), ResNet18 (Aldjia et al. [26]), InceptionV3 (Jin et al. [27]), ShuffleNetV2 (Qian et al. [28]), and MobileNetV2 (Tapia et al. [29]). Standardization procedures were applied to the different modal features for each method, followed by their fusion into new RGB image features. Subsequently, the recognition performance of the models was evaluated.

In order to better adapt to high-security environments, this paper conducted experiments using a small sample database. Specifically, 200 classes were selected, with each class containing 6 images. The training and validation sets were divided in an 8:2 ratio, meaning that 5 images from each class were used for training and 1 image for validation. This design resulted in evaluation curves with an approximate curvature of 0, although it is still evident that the proposed method enhances performance.

The experimental results are presented in Table 4. The curves depicting the performance of various methods in relation to recognition across the four multimodal fusion databases are shown in Figs. 11 to 14. Figs. 11 and 12 illustrate the accuracy/loss curves on the NIR-TSD database, while Figs. 13 and 14 display the PR/ROC curves obtained from experiments on the four multimodal databases. It can be observed that the proposed method outperforms other methods in terms of recognition performance.

Table 4: Comparison of equal err rate of multiple methods

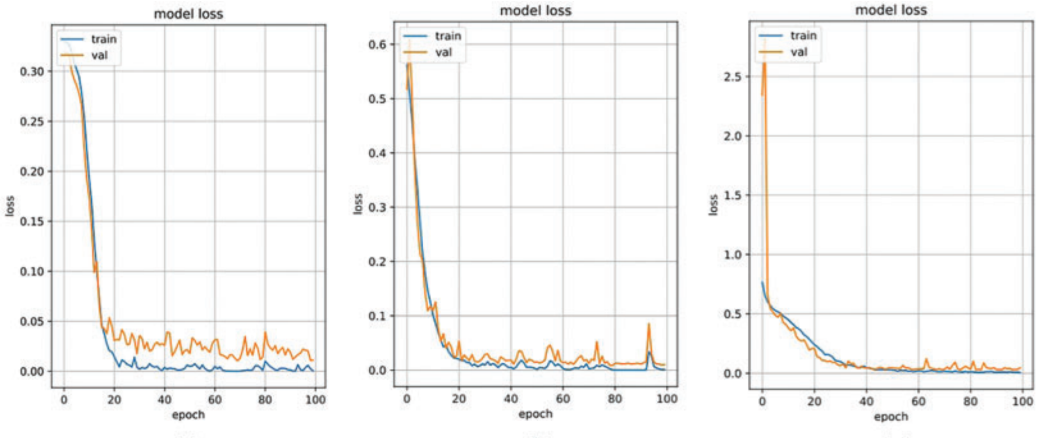
	VGG16	ResNet18	InceptionV3	ShuffleNetV2	MobileNetV2	Proposed method
TCSD	0.0947	0.0647	0.0199	0.0573	0.0597	0.0149
NIR-TSD	0.0698	0.0248	0.0497	0.0598	0.0249	0.0150
NIR-CSD	0.1196	0.0348	0.0398	0.1344	0.0237	0.0099
TTSD	0.0649	0.0398	0.0647	0.0548	0.0547	0.0050



(a) (b) (c)

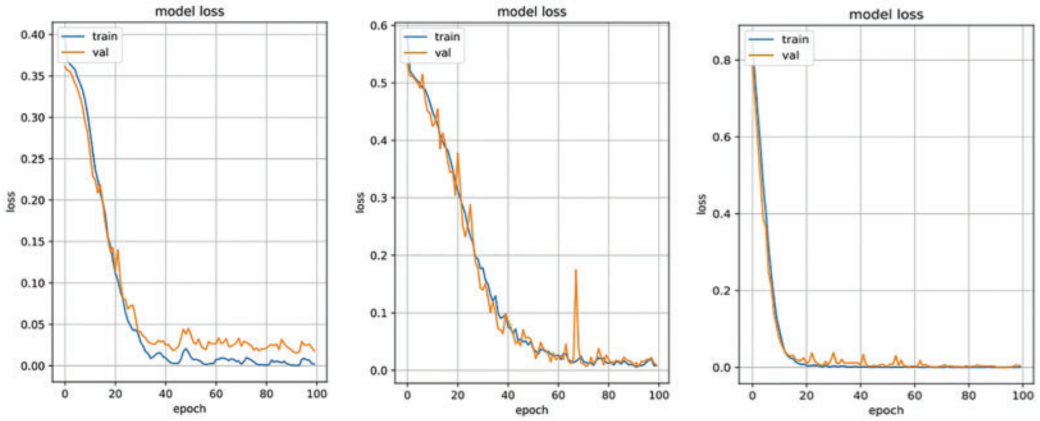
(d) (e) (f)

Figure 11: Accuracy of NIR-TSD database. (a) VGG16; (b) ResNet18; (c) InceptionV3; (d) ShuffleNetV2; (e) MobileNetV2; (f) Proposed method



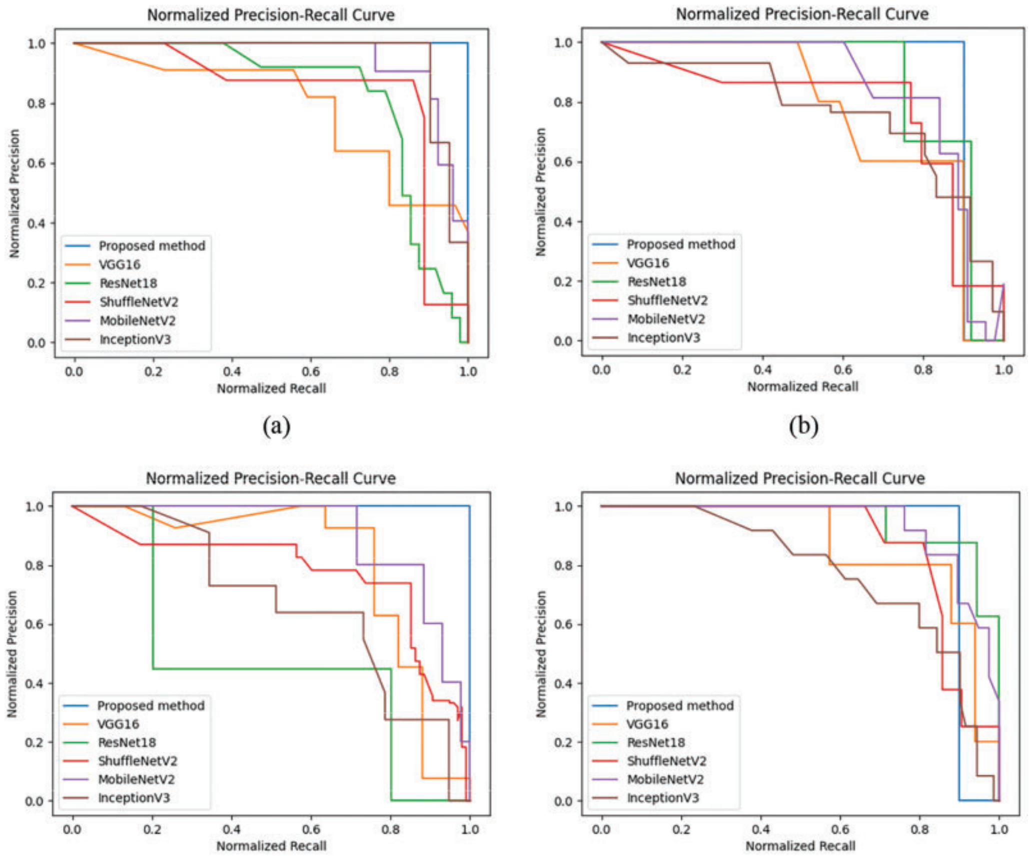
(a) (b) (c)

Figure 12: (Continued)



(d) (e) (f)

Figure 12: NIR-TSD database loss rate. (a) VGG16; (b) ResNet18; (c) InceptionV3; (d) ShuffleNetV2; (e) MobileNetV2; (f) Proposed method



(a) (b)

(c) (d)

Figure 13: PR curves of four multimodal databases. (a) TCSD; (b) NIR-TSD; (c) NIR-CSD; (d) TTSD

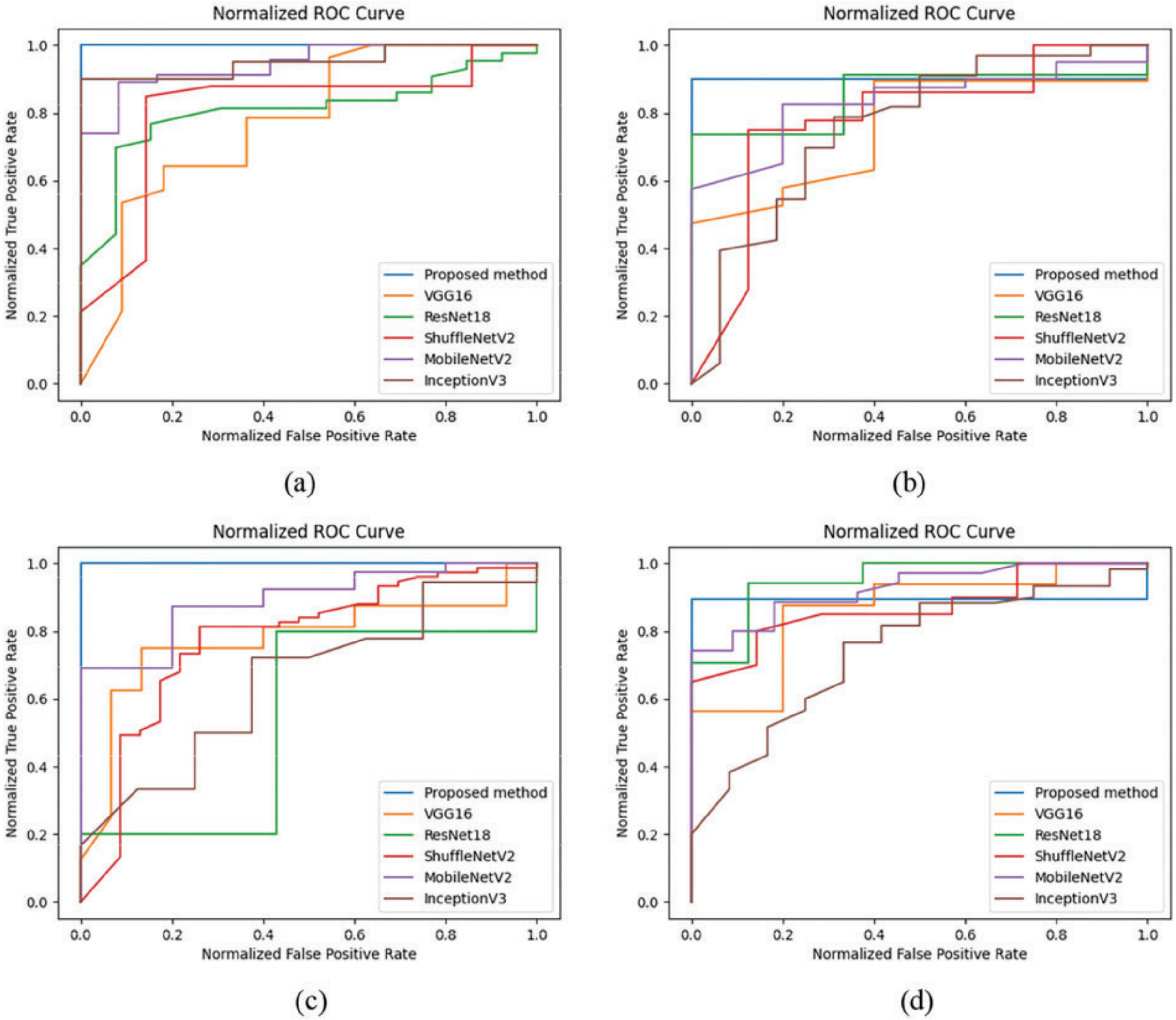


Figure 14: ROC curves of four multimodal databases. (a) TCSD; (b) NIR-TSD; (c) NIR-CSD; (d) TTSD

5 Conclusions

This study aims to improve feature-level fusion recognition for multimodal hand features. It achieves this by employing an enhanced DenseNet deep learning network that leverages inter-modal correlations. Multimodal biometrics recognition encounters two challenges in practical applications. First, the fusion recognition often fails to adequately consider the correlation between modalities. Second, improper weight selection during modal fusion can diminish the saliency of key features, leading to a degradation in overall recognition performance. To address these problems, an improved DenseNet multimodal recognition network based on feature-level fusion is proposed in this paper. The fusion of the three modalities is similar to RGB fusion, where the correlation between modalities is enhanced by inputting channel correlation information to the network. In the improved DenseNet network, the ECA network dynamically adjusts the weights of each channel to increase the salience of important information in each modal feature. In addition, the deep separable convolution greatly reduces the training parameters while further enhancing the feature correlation.

The experimental results indicate that the proposed method achieves remarkably low EERs of 0.0149%, 0.0150%, 0.0099%, and 0.0050% on the four multimodal databases, demonstrating its practical application value in identity recognition for scenarios with high security requirements. In the next stage of research, further optimization of the model structure and parameter configuration will be considered to enhance generalization performance under different environmental conditions and reduce the demand for computational resources.

In the conducted experiments in this article, all the databases were processed following the criteria of the finger vein database. The resulting multimodal databases are all small-sample databases with limited capacity. Consequently, the next phase of this study aims to overcome the limitations of the databases, enabling experimentation on more extensive databases. This ensures that the program maintains robustness even when applied to larger databases.

Acknowledgement: The authors would like to thank the anonymous reviewers and journal editors for providing useful comments and suggestions.

Funding Statement: This research was funded by the National Natural Science Foundation of China (61991413), the China Postdoctoral Science Foundation (2019M651142), the Natural Science Foundation of Liaoning Province (2021-KF-12-07), and the Natural Science Foundations of Liaoning Province (2023-MS-322).

Author Contributions: The authors confirm contribution to the paper as follows: study conception and design: Wei Wu, Yuan Zhang; data collection: Yunpeng Li, Chuanyang Li; analysis and interpretation of results: Wei Wu, Yan Hao, Yuan Zhang; draft manuscript preparation: Wei Wu, Yuan Zhang. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: In this article, the CASIA Multi-Spectral Palmprint Database (CASIA-P), the Polyu database of the Hong Kong Polytechnic University (PolyU-NIR), the Palmprint and Palm Vein database of Tongji University (Tongji-P, Tongji-V), and the Finger Vein Database of Shandong University (SDUMLA-HMT) were used to conduct the experiments. And the corresponding literature sources of the databases are listed in Table 2.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

1. Benamara, N. K., Zigh, E., Stambouli, T. B., Keche, M. (2022). Towards a robust thermal-visible heterogeneous face recognition approach based on a cycle generative adversarial network. *International Journal of Interactive Multimedia and Artificial Intelligence, 7(4),* 132–145.
2. Zheng, C., Mendieta, M., Chen, C. (2023). POSTER: A pyramid cross-fusion transformer network for facial expression recognition. *2023 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3138–3147. Paris, France.
3. Jiménez-Bravo, D. M., Murciego, Á.L., Mendes, A. S., Silva, L. A., Iglesia, D. H. D. L. (2022). Edge face recognition system based on one-shot augmented learning. *International Journal of Interactive Multimedia and Artificial Intelligence, 7(6),* 31–44.

4. Sun, Z., Zhang, H. H., Bai, J. T., Liu, M. Y., Hu, Z. P. (2023). A discriminatively deep fusion approach with improved conditional GAN (im-cGAN) for facial expression recognition. *Pattern Recognition, 135,* 109157.
5. Shi, S. X., Han, D. Z., Cui, M. M. (2023). A multimodal hybrid parallel network intrusion detection model. *Connection Science, 35(1),* 2227780.
6. Han, D. Z., Zhou, H. X., Weng, T. H., Wu, Z. D., Han, B. et al. (2023). LMCA: A lightweight anomaly network traffic detection model integrating adjusted mobilenet and coordinate attention mechanism for IoT. *Telecommunication Systems, 84(4),* 549–564.
7. Chen, C. Q., Han, D. Z., Chang, C. C. (2024). MPCCT: Multimodal vision-language learning paradigm with context-based compact transformer. *Pattern Recognition, 147,* 110084.
8. Wu, T., Leng, L., Khan, M. K., Khan, F. A. (2021). Palmprint-palmvein fusion recognition based on deep hashing network. *IEEE Access, 9,* 135816–135827.
9. Oldal, L. G., Kovács, A. (2020). Hand geometry and palmprint-based authentication using image processing. *2020 IEEE 18th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 125–130. Subotica, Serbia.
10. Ramachandran, C., Sankar, D. (2020). Score level based fusion method for multimodal biometric recognition using palmprint and Iris. *2020 Advanced Computing and Communication Technologies for High Performance Applications (ACCTHPA)*, pp. 281–286. Cochin, India.
11. Yang, W., Huang, J., Luo, D., Kang, W. (2024). Efficient disentangled representation learning for multimodal finger biometrics. *Pattern Recognition, 145,* 109944.
12. Rajasekar, V., Saracevic, M., Hassaballah, M., Karabasevic, D., Stanujkic, D. et al. (2023). Efficient multimodal biometric recognition for secure authentication based on deep learning approach. *International Journal on Artificial Intelligence Tools, 32(3),* 2340017.
13. Daas, S., Yahi, A., Bakir, T., Sedhane, M., Boughazi, M. et al. (2020). Multimodal biometric recognition systems using deep learning based on the finger vein and finger knuckle print fusion. *IET Image Processing, 14(15),* 3859–3868.
14. Sasikala, T. S. (2024). A secure multi-modal biometrics using deep ConvGRU neural networks based hashing. *Expert Systems with Applications, 235,* 121096.
15. Gona, A., Subramoniam, M. (2022). Convolutional neural network with improved feature ranking for robust multi-modal biometric system. *Computers and Electrical Engineering, 101,* 108096.
16. Zhou, C., Huang, J., Yang, F., Liu, Y. (2020). A hybrid fusion model of Iris, palm vein and finger vein for multi-biometric recognition system. *Multimedia Tools and Applications, 79(39–40),* 29021–29042.
17. Abdullahi, S. B., Bature, Z. A., Chophuk, P., Muhammad, A. (2020). Sequence-wise multimodal biometric fingerprint and finger-vein recognition network (STMFPFV-Net). *Intelligent Systems with Applications, 19,* 200256.
18. Tiong, L. C. O., Kim, S. T., Ro, Y. M. (2020). Multimodal facial biometrics recognition: Dual-stream convolutional neural networks with multi-feature fusion layers. *Image and Vision Computing, 102,* 103977.
19. Wu, W., Wang, Q., Yu, S., Luo, Q., Lin, S. et al. (2021). Outside box and contactless palm vein recognition based on a wavelet denoising ResNet. *IEEE Access, 9,* 82471–82484.
20. Krishnan, A., Thomas, T., Mishra, D. (2021). Finger vein pulsation-based biometric recognition. *IEEE Transactions on Information Forensics and Security, 16,* 5034–5044.
21. Zhou, Y., Kumar, A. (2011). Human identification using palm-vein images. *IEEE Transactions on Information Forensics and Security, 6(4),* 1259–1274.
22. Zhang, D., Guo, Z. H., Lu, G. M., Zhang, L., Zuo, W. M. (2010). An online system of multispectral palmprint verification. *IEEE Transactions on Instrumentation and Measurement, 59(2),* 480–490.
23. Zhang, L., Li, L., Yang, A., Shen, Y., Yang, M. (2017). Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach. *Pattern Recognition, 69,* 199–212.

24. Yin, Y., Liu, L., Sun, X. (2011). SDUMLA-HMT: A multimodal biometric database. In: *Biometric recognition,* pp. 260–268. Berlin, Heidelberg: Springer Berlin Heidelberg.

25. Aleidan, A. A., Abbas, Q., Daadaa, Y., Qureshi, I., Perumal, G. (2023). Biometric-based human identification using ensemble-based technique and ECG signals. *Applied Sciences, 13(16),* 9454.

26. Aldjia, B., Leila, B. (2021). Sensor level fusion for multi-modal biometric identification using deep learning. *Proceedings of the 2021 International Conference on Recent Advances in Mathematics and Informatics (ICRAMI)*, pp. 1–5. Tebessa, Algeria.

27. Jin, H., Chen, S. (2023). Biometric recognition based on recurrence plot and inceptionV3 model using eye movements. *IEEE Journal of Biomedical and Health Informatics, 27(11),* 5554–5563.

28. Qian, H., Zhou, Y., Ding, P., Feng, S. (2023). *ConShuffleNet: An efficient convolutional neural network based on ShuffleNetV2,* pp. 948–955. Singapore: Springer Nature Singapore.

29. Tapia, J. E., Gonzalez, S., Busch, C. (2022). Iris liveness detection using a cascade of dedicated deep learning networks. *IEEE Transactions on Information Forensics and Security, 17,* 42–52.