

Received 15 November 2022, accepted 15 December 2022, date of publication 6 February 2023, date of current version 23 February 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3242984

## RESEARCH ARTICLE

# Multimodal Finger Recognition Based on Asymmetric Networks With Fused Similarity

YIWEI HUANG<sup>1</sup>, HUI MA<sup>1,2</sup>, AND MINGYANG WANG<sup>1</sup>

<sup>1</sup>College of Electronic Engineering, Heilongjiang University, Harbin 150080, China

<sup>2</sup>Key Laboratory of Information Fusion Estimation and Detection, Heilongjiang University, Harbin 150080, China

Corresponding author: Hui Ma (mahui929@126.com)


This work was supported in part by the Heilongjiang Provincial Natural Science Foundation under Grant LH2022F047, and in part by the Special Fund of Fundamental Scientific Research Business Expense for Higher School of Heilongjiang Province under Grant 2021-KYYWF-0002.

**ABSTRACT** Multimodal biometric system has received increasing interest as it offers a more secure and accurate authentication solution than unimodal systems. However, existing biometric fusion methods are still inadequate in dealing with correlations and redundancy of multimodal features simultaneously, causing bottlenecks in performance improvement. To overcome the above problem, this paper proposes an end-to-end multimodal finger recognition model that incorporates attention mechanisms into a similarity-aware encoder for accurate recognition results. Firstly, due to the different distribution of fingerprint and finger vein images, we propose a finger asymmetric backbone network (FAB-Net) for extracting discriminative intra-modal features, which reduces the network width by efficient utilization of feature maps. Then, a novel attention-based encoder fusion network (AEF-Net) with fused similarity performs dimensionality reduction-based fusion on multimodal multilevel features to alleviate performance degradation due to information redundancy. We also introduce channel attention in AEF-Net, which differs from the traditional attention mechanism by considering interdependencies between modalities to further improve performance. Extensive recognition experiments are conducted on three multimodal finger databases to verify the effectiveness of our method compared to state-of-the-art methods. Detailed ablation studies have also been carried out, which demonstrated that encoder-based reconstruction of redundant information can improve recognition performance.

**INDEX TERMS** Multimodal biometric recognition, feature fusion, autoencoder, deep learning.

## I. INTRODUCTION

With the booming development of smart technologies and biometric recognition, personal identification has become a public social service to meet diversified social needs. Biometric identification refers to the use of the face [1], finger veins [2] and [3], fingerprints [4], or other human characteristics [5] of a person to be authenticated for recognition, and it has grown rapidly because of its exceptional convenience and effectiveness. However, unimodal biometrics recognition has limitations in terms of security and accuracy, such as spoof attacks [6] and intra-class variations, and unimodal data is often hard to monitor completely and comprehensively

The associate editor coordinating the review of this manuscript and approving it for publication was Wenming Cao .

for changes in the acquisition environment [7]. Multimodal biometric systems provide more modes of authentication and a higher level of security, making them a promising form of identity recognition. Among them, multimodal finger biometrics have received the most attention for their user-friendliness and outstanding accessibility in acquisition.

Due to the outstanding security and accuracy advantages of multimodal systems, multimodal finger recognition technology has many real-world applications, such as e-commerce, healthcare, forensics and surveillance, military security systems, etc.

Multimodal biometric recognition is getting increasingly popular as it demonstrates to be a powerful method for extracting complementing data from multimodal databases, but the fusion algorithm is still a key issue in multimodal

recognition technology. The feasibility of multimodal fusion recognition has been demonstrated, and the performance of the bimodal system built by Brunelli et al. [8] has been shown to be better than that of the subsystem. According to the concept of multi-source information fusion proposed by Ross and Jain [9], multimodal fusion algorithms can be classified as feature-level fusion, score-level fusion, and decision-level fusion. Several researchers have attempted to conduct comparative experiments between several different levels of fusion strategies. In the literature [10], feature-level fusion strategies were compared with score-level fusion strategies and decision-level fusion strategies, it was shown that feature fusion has the potential to exhibit higher accuracy in the early stages between various multimodal features. Since only one classifier is required, feature-level fusion is generally faster than decision-level fusion, which usually uses multiple classifiers [11]. As a result of the advantages of feature-level fusion, researchers have proposed a large number of fusion models for multimodal biometric recognition.

A fundamental challenge in feature fusion research is learning how to represent and summarize multimodal data in a way that exploits the complementary and redundancy of multimodality [12]. Deep learning has shown attractive capabilities to learn more representative unimodal features and more flexible fusion strategies, so that multimodal recognition based on deep neural networks is a commonly accepted method among researchers today. Redundant features captured by multiple modalities, give considerable scope for resolving incorrect classifications. However, a part of the research is devoted to establishing correlations between different modalities, neglecting the redundancy of multimodal information. Another part of the convolutional neural network-based fusion strategy ignores the condition that the feature extraction models vary widely across modalities.

To address the information redundancy and feature extraction separation problems of the above fusion methods, we are inspired to propose an asymmetric network with fused similarity that is trained in two stages. Considering the correlation and redundancy of multimodal information simultaneously, the aim is to reconstruct a more discriminative common representation that achieves state-of-the-art results in terms of recognition accuracy and recognition time. The key contributions of the presented work are:

- To address the performance degradation due to multimodal information redundancy, we propose a novel attention-based encoder fusion network (AEF-Net) to model a compact representation that makes recognition easier and faster. The proposed AEF-Net fully considers the correlation and redundancy of multimodal features, generating the interdependence weights between different modalities by channel attention module, and simultaneously maximizing the discriminative deep information via a similarity-aware encoder.
- To alleviate the inconsistent distribution between fingerprint and finger vein images, a finger asymmetric backbone network (FAB-Net) is proposed, in which

the asymmetric structure overcomes the gaps in image attributes, thus improving recognition accuracy using modality-specific information. Additionally, intra-modal feature maps of different depths are fused to reduce the network width by making efficient use of the representations.

- In order to train more adequately, an assisted training method consisting of multiple training stages is proposed to shorten the training time while avoiding the low recognition rate caused by unbalanced training in multi-stream networks.

The rest of the paper is organized as follows: the second section briefly describes the related work on multimodal feature-level fusion and notable multi-biometric recognition carried out recently. Section III demonstrates the proposed attention-based encoder fusion network as well as the assisting training method. The fourth presents our experiments on multi-biometric fusion recognition, which focus on the recognition performed on three fingerprint-finger vein databases and the analysis of the results. Finally, the last section concludes the paper.

## II. RELATED WORK

### A. MULTIMODAL BIOMETRIC RECOGNITION

Compared with unimodal systems, multimodal biometric systems can effectively improve the recognition performance in accuracy, and security [13], [14]. Multimodal feature fusion methods re-model the redundant and complementary unimodal features as a common representation to achieve cross-modal feature association. Naturally, traditional multimodal fusion studies can be divided into roughly two categories. The first one is based on matrix transformation strategies, such as Principal Component Analysis (PCA) [15], enhanced partial discrete Fourier transform [16], and weighted joint sparse representation-based classification (WJSRC) [17]. As a further improvement of the matrix fusion strategy, Canonical Correlation Analysis (CCA) [18], [19] maximizes the correlation of different modal information by iteration to derive a linear mapping matrix, which further maps the separated different modal features into the same common space. Discriminant Correlation Analysis (DCA) [7] and Adversarial Canonical Correlation Analysis (ACCA) [20] have been applied to biometrics as variants of the above methods. During the fusion methods based on matrix transformation, the projection transformations are often linear or consider only interrelationships, leading to the existence of constraints in the common potential space. Another traditional fusion method is based on spatial models. Graph model [21], [22] and rotation invariant hierarchical model [23] have been proposed for feature construction of multimodal data, modeling cross-modal public representations by splicing tandem. An improved graph fusion model [22] extracts graph structure features of multimodal finger bio images and connects nodes with similar structures to fuse the graphs. So far, a large number of hand-crafted

**TABLE 1. Data sources and methodology on finger-based multimodal research.**

Reference	Methodology	Biometric Traits	Data sources	Database Type
Kamlaskar et al. (2021) [18]	PCA, CCA	FP + Iris	SDUMLA	Public
Wang et al. (2019) [15]	CNN, PCA	FP + FV + FKP	Finger trimodal database	Own
Qu et al. (2021) [22]	Competitive Fusion, Graph CNN	FP + FV + FKP	Homemade finger trimodal database	
Yang et al. (2018) [16]	Discrete Fourier transform	FP + FV	FVC2002 + FV-HMTD	Combined
Su et al. (2019) [27]	LBP, DCA	FV + EEG	HKPU + ECG-ID	
Fang et al. (2021) [17]	Weighted approach, Sparse representation	FP + FV	FVC2006 + HKPU	
Li et al. (2021) [13]	LC-CNN	FP + FV+FKP	Our-tri / SDUMLA + Our-tri	Own and Combined
Ren et al. (2022) [14]	CNN, Attention mechanism	FP + FV	NUPT-FPV / FVC2002 + MMCBNU	

feature representations have been widely proposed and have been studied and applied in biometric recognition.

After the emergence of traditional methods, deep learning-based methods are more flexible and versatile when dealing with heterogeneous cross-modal data. An intuitive approach is to use summation or collocation strategies to sum [24] or splice feature maps [14], [15] of specified dimensions to accomplish combined prediction. Leghari et al. [25] compared feature-level fusion strategies on multimodal biometric data and showed that higher accuracy was obtained by performing fusion at the convolutional layer than at the linear layer. Selective fusion networks [26] weight the estimated high-quality information with the original depth information as a whole, but ignore the intra-modal inter-channel variability. Table 1 shows the mainstream data sources in multimodal biometrics research. There are currently rather scarce databases on multimodal finger biometrics. Several studies have collected their proprietary multimodal finger data for fusion studies, such as in the literature [14], [15], [22]. Furthermore, combining biometric traits from different individuals is a widely used solution by researchers, as demonstrated by the literature [16], [17], [27], which integrated two publicly available unimodal databases.

### B. AUTOENCODER-BASED FUSION

Researchers have started to consider breaking the limitations imposed by manual collocation and using nonlinear methods to obtain more representational fused features. Since the success of the autoencoder reduced the dimensionality of the data, the encoder architecture has become the benchmark approach for coding problems in representation learning. It contains two parts, encoder and decoder, where the encoder models smaller-scale feature vectors and then the output of the encoder is fed to the decoder for learning of reconstructed features. The decoder learns reconstructed features of the same scale as the original features, and then solves the two-part mapping relationship that minimizes the reconstructed error between the input and output features. Kuzu et al. [28] introduced an autoencoder for finger biometric recognition with a backbone network consisting of dense links.

Not only in the field of representation learning [29] but there is also a trend for researchers to use autoencoders to handle multimodal data modeling common representations.

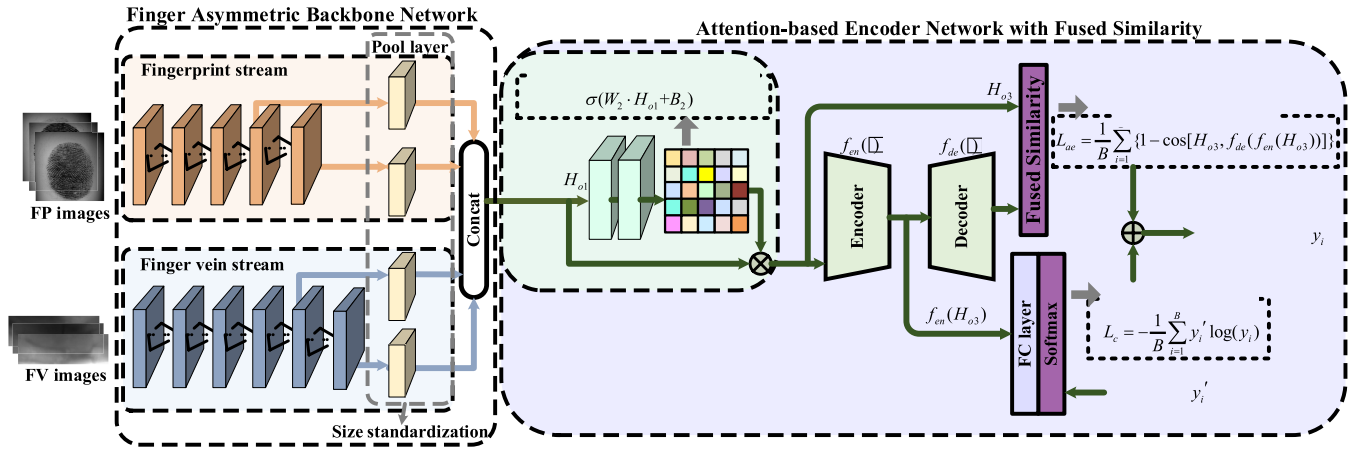
Abavisani et al. [30] provided an autoencoder-based approach for multimodal clustering to project data into the latent space representation. The automatic fusion method [31] alleviates the static nature of existing fusion methods, effectively combining multimodal inputs through autoencoders. Autoencoder-based fusion method is irregular and successfully escapes the limitations of linear projection, thus enabling a more accurate fit to the common feature space.

### C. ATTENTION-BASED FUSION

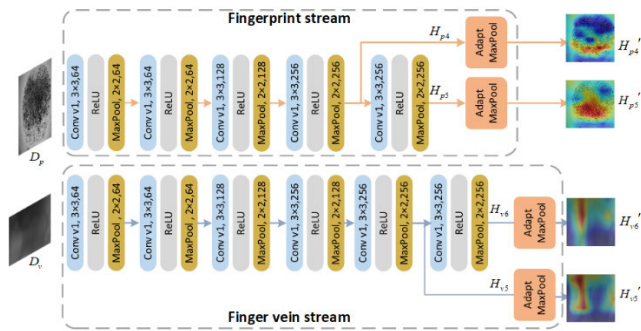
Mimicking the human tendency to focus more on the discriminative part of the image, the information output by the attention mechanism is more helpful for recognition. Squeeze-and-Excitation Network (SENet) [32], Efficient Channel Attention (ECA) [33] recalibrates channel weights based on the large primary school of feature contribution to the classification task. With the superior transferability of the attention mechanism, it has also been widely extended in cross-modal fusion algorithms.

Cross-modal attention [14], [34] is an extension of the previous combination of different modal feature vectors that enhances the representation of multimodal features by assigning different attention weights for modeling information interactions between modalities. Lv et al. [35] input multi-level features from a three-stream network model into the channel attention mechanism to model the common feature space and focus the network on salient targets. In addition, Attn-Hybrid Net [36] was used to alleviate the redundancy between hybrid features. Inspired by [34], Ren et al. [14] proposed an attention-based CNN to fuse the local and global information from fingerprint and finger vein. Wang [24] compared summation and concatenation for feature maps extracted by CNN, and subsequently proposed a fusion network with self-attention that achieved better results. The attention-based fusion method reduces the interference of external factors on the features, thus improving the stability and effectiveness of the network.

Different from position attention that focuses on the representation region of images, we introduce channel attention that fully considers the correlation of inter-modal representational ability differences and improves recognition performance. Furthermore, inspired by the successful application of autoencoders, we exploit both the complementarity and



**FIGURE 1.** An overview of the proposed multimodal recognition framework with two main components, FAB-Net (Section III-A) and AEF-Net (Section III-B). In particular, we introduce channel attention to learn multimodal interdependencies. Subsequently, the encoder nonlinearly fits low-dimensional common features while the decoder performs feature reconstruction. The multimodal recognition model is trained by minimizing a weighted combination of cosine dissimilarity and cross-entropy loss.



**FIGURE 2.** The architecture of finger asymmetric backbone network.

redundancy of multimodality to jointly learn fusion features with greater similarity from fingerprint and finger vein representations.

**III. METHODOLOGY**

The proposed multimodal recognition model based on fused similarity consists of the finger asymmetric backbone network (FAB-Net), extracting intra-modal features, and the attention-based encoder fusion network (AEF-Net), fusing cross-modal features. The general framework of our method is shown in Figure 1. Specifically, each branch of the FAB-Net is trained by adjusting the network depth to the image properties of different modalities, while the AEF-Net learns and reconstructs the common representation via convolutional autoencoder. Combining cross-entropy loss and fusion similarity alleviates the local optimization problem caused by the separation of feature extraction and feature fusion, generating a more discriminative common representation to further improve recognition accuracy.

**A. FAB-NET FOR FEATURE EXTRACTION**

Motivated by the VGG-16 [37] model, we propose a finger asymmetric backbone network (FAB-Net) with dual-stream

to extract feature maps of both fingerprint modal and finger vein modal. In order to efficiently capture model-specific features, two streams of the FAB-Net have different depths.

Compared to fingerprint images, finger vein images contain more global features and require a larger field of perception, so a deeper network than fingerprint modalities is used to represent the intra-modal differences. Although the semantics information tends to be richer as the network deepens, the fingerprint images have simple backgrounds and fewer global features, mostly point features. Therefore, local features extracted by the shallow network, such as edges, points, and textures, are sufficiently discriminative, where the differences between fingerprint images can be clearly distinguished. To better capture the modality-specific features of fingerprint and finger vein images in the neural networks to describe the subject to be authenticated, the FAB-Net contains two streams in different depths.

The FAB-Net consists of two streams: the fingerprint stream contains five convolutional layers and adaptive pooling layers, and the finger vein stream contains six convolutional layers, as shown in Figure 2. The relevant representation of multimodal data is defined, which consists of fingerprint modal and finger vein modal, and the multimodal database is denoted as  $D = \{D_p, D_v\}$ . Specifically,  $p$  denotes the fingerprint modal and the vein modal is  $v$ . Moreover, the output feature map of each pooling layer is denoted by  $H$ , where  $H_{pi}$  denotes the  $i$ th layer feature map of the fingerprint modal and the output feature of vein modal is  $H_{vi}$ . To take full advantage of the powerful representation of deep features, FAB-Net effectively reduces the number of parameter definitions by collocating multiple layers of intra-modal feature maps and outputs features with standardized sizes.

Without a uniform size, the irregular feature maps cannot be concatenated and fused, so size standardization is necessary. The adaptive pooling layer in the backbone network converts heterogeneous features into standardized sizes. In the



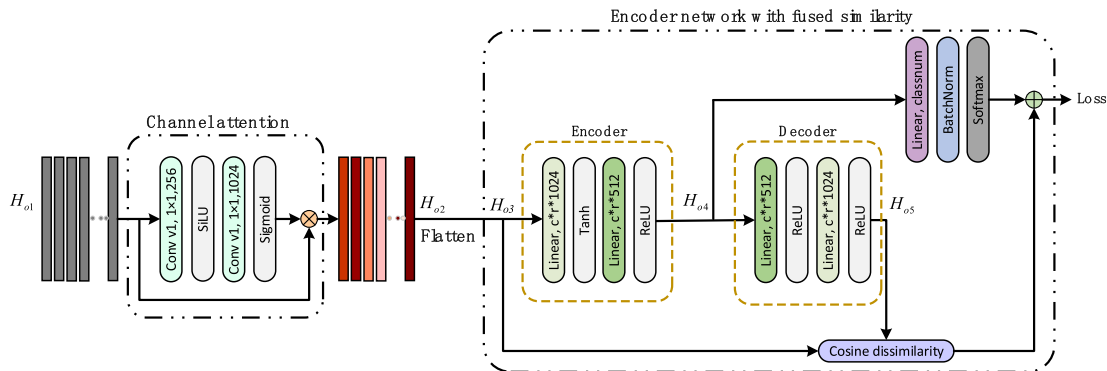


FIGURE 3. The architecture of attention-based encoder fusion network (AEF-Net).

four higher layer features to be fused, the sizes of the higher features under the same modal are smaller than those of the upper layer features, so the standardization size is shown in the following equation.

$$r = \min(\text{row}(H_{p5}), \text{row}(H_{p6})) \tag{1}$$

$$c = \min(\text{column}(H_{p5}), \text{column}(H_{p6})) \tag{2}$$

After obtaining the standardized sizes, the kernel size *pool\_kernel* and stride *pool\_stride* of the adaptive pooling layer are determined based on the known input feature maps and output feature maps as follows:

$$\text{pool\_stride} = \text{floor}(\text{in\_size}/\text{out\_size}) \tag{3}$$

$$\text{pool\_kernel} = \text{in\_size} - (\text{out\_size} - 1) \times \text{pool\_stride} \tag{4}$$

where *floor*(·) is rounded down, *in\_size* denotes the size of the feature input to the adaptive pooling layer, and the standardized size denotes *out\_size*.

**B. AEF-NET FOR FEATURE FUSION**

Strikingly, a novel attention-based encoder fusion network (AEF-Net) consisting of multiple convolutional layers is proposed to model various nonlinear transformations. To address the degradation of recognition performance due to multimodal feature redundancy, the AEF-Net is capable of non-linearly fitting a more accurate low-dimensional common representation. In addition, the attention module is added to the encoder network to highlight the main features, unlike image attention, which is used to learn the interdependencies between different modalities. As shown in Figure 3, the AEF-Net consists of an attention module and a similarity-aware encoder.

Feature maps with the same channels are concatenated together and fed into the AEF-Net. Firstly, the standardized output of the backbone network  $H_{p4}'$ ,  $H_{p5}'$ ,  $H_{v5}'$ , and  $H_{v6}'$  are concatenated to obtain the multimodal feature  $H_{o1}$ .

For the heterogeneity gap caused by inconsistent representation of different modalities in multimodal fusion, the attention model automatically learns the interdependencies between modalities and effectively recalibrates multimodal feature response. Instead of the traditional attention focused

on the representational regions of the image, we innovatively apply channel attention to generate weights representing the global distribution of responses on multimodal channels. The parameters of the attention module are updated as shown in the following equation:

$$H_{o2} = \sigma(W_2 \cdot H_{o1} + B_2) \cdot H_{o1} \tag{5}$$

where  $\sigma$  is the sigmoid function, the weight and bias of the second convolutional layer are denoted as  $W_2$ ,  $B_2$ . Finally, the output tensor is applied to the input feature map  $H_{o1}$  channel-by-channel to generate the original multimodal feature map  $H_{o2}$ , which can be fed into a subsequent encoder network.

However, the multimodal features generated by the attention module still suffer from redundancy. To solve this problem, a similarity-aware encoder is proposed to project the original features into a low-dimensional space to further learn the distributed representation of multimodal.

The output feature  $H_{o2}$  of the attention module is first expanded by rows into original common features, denoted as  $H_{o3}$ , as the input to the encoder network. Theoretically, more than two convolutional layers are available to simulate all nonlinear distributions. The encoder network consists of two convolutional layers, where the first layer downscales the output to half the number of input features and the second layer generates a low-dimensional common feature, which is formulated as:

$$H_{o4} = f_{en}(H_{o3}) \tag{6}$$

where  $f_{en}(\cdot)$  is denoted as the nonlinear transformation function of the encoder network. The generated multimodal features  $H_{o4}$  are more compact and more discriminative compared to the original multimodal feature map  $H_{o3}$ .

The decoder likewise contains multiple convolutional layers, generating the reconstructed feature  $H_{o5}$ , to maintain semantic consistency of the multimodal common feature, represented as:

$$H_{o5} = f_{de}(f_{en}(H_{o3})) \tag{7}$$

where  $f_{de}(\cdot)$  is denoted as the nonlinear transformation function of the decoder. Minimize the reconstruction error

between the reconstructed feature and the original common feature, so that the common feature generated by the encoder fusion network has stronger discriminative power and less redundant information.

### C. LOSS FUNCTION WITH FUSED SIMILARITY

Multimodal recognition model is trained by minimizing the weighted combination of the cosine dissimilarity and cross-entropy loss. Specifically, the original common feature Ho3 is used as the original data and the reconstructed feature Ho5 is used as the generated data for discrimination. We measure the reconstruction error using the cosine dissimilarity:

$$L_{ae} = \frac{1}{B} \sum_{i=0}^B \{1 - \cos[H_{o3}, f_{de}(f_{en}(H_{o3}))]\} \quad (8)$$

where  $\cos[\cdot]$  denotes the cosine similarity and  $B$  denotes batch size. The smaller the cosine dissimilarity  $L_{ae}$  is, the closer the original data and the reconstructed data are.

Aiming to highlight representative features and effectively reduce redundant information, the common features output from our AEF-Net are used for authentication, enabling faster and better multimodal biometric recognition. The classifier includes a linear layer that outputs a  $K$ -dimensional probability vector. Batch normalization (BN) is added to speed up training and improve network generalization. Since the features generated by the encoder are generic but not necessarily optimal for the recognition task, the softmax function performs a cross-entropy operation on the predicted and actual labels of the samples as follows.

$$L_c = -\frac{1}{B} \sum_{i=0}^B y_i' \log(y_i) \quad (9)$$

where  $i$  denotes the serial number of the sample within the mini-batch,  $B$  denotes the batch size.  $y_i'$  is the true class of the  $i$ th sample;  $y_i$  denotes the value of the  $i$ th element in the output vector  $[y_1, y_2, \dots, y_K]$ . The vectors output from the linear layer are cross-entropy operated with the true labels by the softmax loss function.

To avoid the local optimum problem due to the separation of feature fusion and feature extraction, the loss to be minimized is defined as:

$$L = L_c + \beta L_{ae} \quad (10)$$

where  $\beta$  is the sparsity tuning parameter; cosine dissimilarity is used for the reconstruction loss  $L_{ae}$  of the encoder network;  $L_c$  denotes the multimodal recognition model trained on fingerprint-finger vein sample pairs using the cross-entropy function.

Therefore, the total loss  $L$  of multimodal system updating using stochastic gradient is:

$$L = \frac{1}{B} \sum_{i=0}^B \{\beta - \beta \cos[H_{o3}, f_{de}(f_{en}(H_{o3}))]\} - y_i' \log(y_i) \quad (11)$$

where Eq. (11) is the total loss function of the network training as a weighted sum of the softmax loss function and

**TABLE 2. Training process design for multimodal model.**

<b>Algorithm 1</b> The procedure of the proposed method	
<b>Required:</b>	fingerprint training images $D_p$ , finger vein training images $D_v$ , first stage batch size $B_1$ , first stage learning rate $\alpha_1$ , second stage batch size $B_2$ , second stage learning rate $\alpha_2$ .
1:	Stage 1: Training fingerprint recognition model.
2:	<b>repeat</b>
3:	Fingerprint Image $\{p_j\}_{j=1}^N \subseteq D_p$ is input to fingerprint stream $f_p(p_j)$ , and the unimodal fingerprint recognition model is updated using a stochastic gradient to generate fingerprint stream parameters.
4:	<b>until</b> the fingerprint recognition model converges.
5:	<b>return</b> the optimized fingerprint stream parameters.
6:	Stage 2: Training multimodal recognition model.
7:	Fingerprint image $\{p_j\}_{j=1}^N \subseteq D_p$ and finger vein image $\{v_j\}_{j=1}^N \subseteq D_v$ of the same finger are combined as an image pair, and then the FP stream of the multimodal network is initialized using the parameters returned in the first stage.
8:	<b>repeat</b>
9:	Dual-stream backbone network generation. $f_p(p_j) \rightarrow (H_{p4}, H_{p5}), f_v(v_j) \rightarrow (H_{v5}, H_{v6})$ .
10:	Size standardization for feature maps. $pool(H_{p4}, H_{p5}, H_{v5}, H_{v6}) \rightarrow (H_{p4}', H_{p5}', H_{v5}', H_{v6}')$ .
11:	Channel attention module generates the original common feature. $f_{att}(H_{p4}', H_{p5}', H_{v5}', H_{v6}') \rightarrow H_{o2}$ .
12:	Autoencoder network for encoding and reconstruction. $f_{ae}(H_{o2}) \rightarrow (H_{o4}, H_{o5})$ .
13:	Update the multimodal recognition model with ascending its stochastic gradient by Equation (11).
14:	<b>until</b> the multimodal recognition model converges.
15:	<b>return</b> the output prediction of the optimized model $y'$ .

the cosine phase dissimilarity, and the total loss is adjusted dynamically by  $\beta$ .

Cosine dissimilarity measures the magnitude of the difference in terms of the cosine of the angle between two features in vector space, focusing on the difference between the original and reconstructed features in terms of direction, rather than distance or length. Since Euclidean distance is meaningless for fusion similarity, we employ encoders to make the category differences larger, not just pursuing numerical agreement on the common representations. The combination of cosine dissimilarity and cross-entropy loss function is used to maximize the probability of correct classification as much as possible. Remarkably, the encoder network with fused similarity compensates for the disadvantage that cross-entropy loss only encourages the differentiability of categories and does not optimize the inter-class distance.

### D. DESIGN OF TRAINING APPROACH

To solve the dominant training inadequacy problem in fusion, we propose an assisted training method, divided into multiple stages. The method applies to the environment where the image properties of different streams differ greatly in a multi-stream network, and successfully solves the training imbalance problem in multimodal fusion recognition. Our proposed general training method is shown in Table 2.

For the first stage, unimodal recognition is used to train the stream which is more difficult to train and has a larger number of image features, such as fingerprint stream. It is worth noting that pre-training of unimodal streams with high

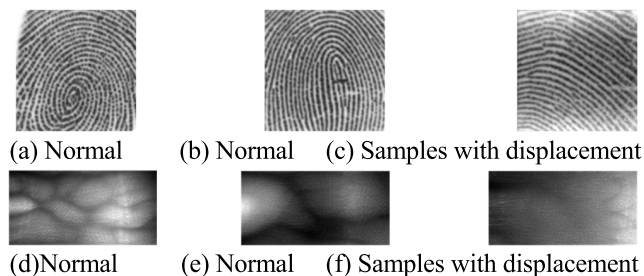


FIGURE 4. Some finger images in HDPR-310 database.

complexity is not only useful for hyperparameter search, but is also very strongly correlated with multimodal fusion recognition results. Then, the unimodal network parameters obtained from the previous stage are used as the initial values of the parameters to the corresponding stream, and then the multimodal fusion model is trained. The general recognition model, including the backbone network and the attention-based encoder network, is trained jointly using the cross-entropy loss and the cosine dissimilarity loss in the encoder network.

#### IV. EXPERIMENTS AND RESULTS

In this section, the effectiveness and advancement of the proposed multimodal finger recognition system are evaluated. We conduct the experiments on PyTorch 1.8 on the computer with NVIDIA 2060 GPU.

##### A. DATABASES

Due to the lack of publicly available multibiometric databases, we conducted recognition experiments on a self-constructed multimodal database. Additionally, since human fingerprint features do not affect finger vein morphology, we combined publicly available fingerprint (FP) and finger vein (FV) databases according to common strategies in the field of multimodal biometric recognition such as [13], [14], [16], and [27].

##### 1) HDPR-310

The HDPR-310 is a multi-biometric database constructed in our laboratory, including 256 individuals, where each individual contains 5 fingerprint images and 10 finger vein images. The multimodal image is captured by the same camera, where an infrared filter is added in front of the camera to capture the finger vein image, and the fingerprint image is acquired by pressing the prism with the finger. The image format is 256 gray levels, where the fingerprint resolution is  $192 \times 192$  pixels and the finger vein resolution is  $240 \times 120$  pixels. Although the finger vein images are preprocessed, image noise exists and the texture is poorly distinguished from the background. The fingerprint images in the self-built library are prone to offset due to the smaller acquisition area, which results in the extracted area not being the region of interest as shown in Figure 4 (c).

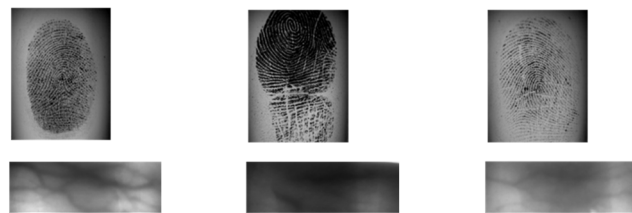
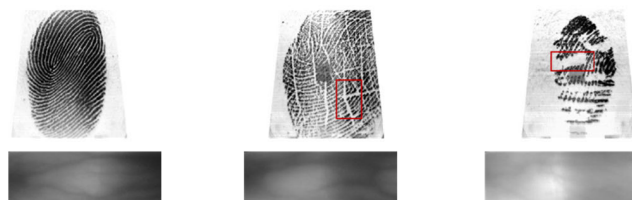


FIGURE 5. Some finger images in FVC-HKP database.

(a) Normal (b) Samples with dry lines (c) Samples with tear



(d) Normal (e) Normal (f) Samples with overexposure

FIGURE 6. Some finger images in CAS-FVU database.

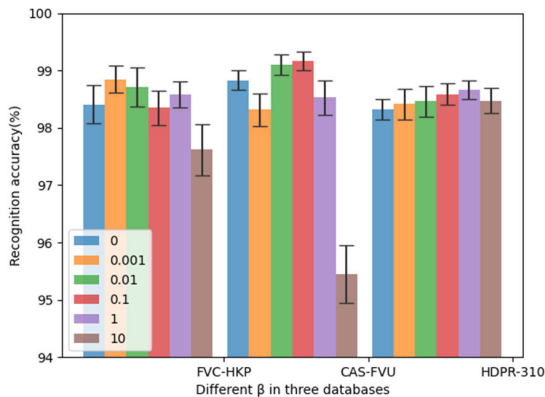
##### 2) FVC-HKP

The combined database with FVC2006 [38] fingerprint database and HKPU [39] finger vein database is denoted as FVC-HKP, as shown in Figure 5. In which, the FVC2006 database contains four databases, and the DB2 database obtained from optical sensor acquisitions is selected for the experiments. The participants include manual workers and elderly people, and each database has 150 finger widths and 12 samples per finger depth.

##### 3) CAS-FVU

The CASIA-Fingerprint [40] database used for fingerprint modal is divided into two time periods, 2009 and 2013, where three different sensors were considered in 2013 and the uru4500 was one of them. Each finger produced 5 fingerprints, for a total of 1960 ( $49 \times 4 \times 5$ ) images. It is noteworthy that the fingerprint images are of below-average quality, as can be seen in Figure 6 (b), where some fingerprint images have a large number of dry lines. Some fingerprint images even show large missing and shifted central areas, as shown in Figure 6 (c). The FV-USM [41] database collected finger vein images of 123 subjects, in which each finger was collected six times.

The data augmentation is performed by rotation and translation strategy, so that the maximum number of samples within the class is the same for different modalities, uniformly 12, in order to combine as image pairs with the same labels. In this case, the fingerprint image is converted to  $224 \times 224$  pixels and randomly flipped, and the region of interest (ROI) of the finger vein image are fed into the backbone network as a set of inputs. Each finger of each person is considered as a class and experiments are performed on multiple fingers of the same person. To better validate the effectiveness of the proposed method, we randomly divide the training set, validation set, and testing set for each database. The number



**FIGURE 7.** Recognition performance using different  $\beta$ . The mean value and standard error of recognition accuracy is given.

of training set, validation set, and test set samples within each category of databases is 6:2:4.

## B. EXPERIMENTS SETTINGS OF KEY PARAMETERS

### 1) PARAMETERS SETTING

The parameters of the CNN are randomly initialized using the “kaiming” uniform distribution method and the batch size is set to 16. In the first stage, the learning rate is first set to 0.01, and updated to 0.001 using cosine decay. In the second stage, the learning rate is set in the same way as the previous stage. The number of iterations (epoch) of the network in the first stage was 300 and in the second stage it was set to 100. The whole model was trained using the SGD algorithm with 0.9 momentum and 0.0001 weight decay.

### 2) LOSS FUNCTION

The encoder fusion network is optimized by the cosine dissimilarity, while the cross-entropy loss is optimized for the overall recognition model. In summary, the loss function is formulated as equation (11). The autoencoder network was structured as part of the overall recognition network, and the sparsity tuning parameter  $\beta$  was chosen in the range  $\{0, 0.001, 0.01, 0.1, 1, 10\}$  for the experiments. We performed experiments on the parameter determination of the loss function by evaluating the achievable performance at different sparsity tuning parameters. The recognition rate tends to be best in a certain range, and the performance decreases when it is greater or less than that range, especially on database CAS-FVU and database HDPR-310, which are approximately normally distributed. As shown in Figure 7.

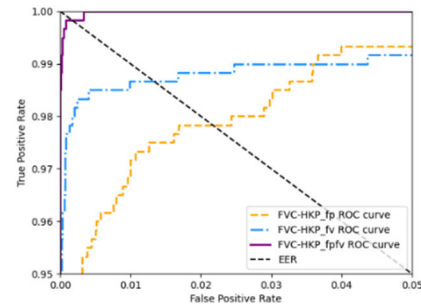
Here the sparsity tuning parameter  $\beta$  is chosen in the range of  $\{0, 0.001, 0.01, 0.1, 1, 10\}$  to obtain a better optimized mode. Therefore, for subsequent experiments we set the sparsity tuning parameter  $\beta$  for the combined database FVC-HKP set to 0.001, the parameter for CAS-FVU set as 0.1, and the parameter  $\beta$  for HDPR-310 is 1.

## C. DISCUSSION OF THE IMPACT OF DIFFERENT ENCODER LAYERS

The fusion network simulates more complex common representations by increasing the number of encoder layers to

**TABLE 3.** Recognition performance of the autoencoder network with different convolutional layers.

Autoencoder layers	Acc (%)			Test time (ms)
	FVC-HKP	CAS-FVU	HDPR-310	
2	$98.42 \pm 0.28$	$98.22 \pm 0.22$	$98.33 \pm 0.20$	5.48
4	$98.84 \pm 0.24$	$99.17 \pm 0.17$	$98.67 \pm 0.16$	5.85
6	$98.65 \pm 0.21$	$99.40 \pm 0.15$	$98.62 \pm 0.17$	5.99



**FIGURE 8.** ROC curves on FVC-HKP database of fingerprint recognition model, finger vein recognition model and multimodal recognition model.

achieve excellent fusion recognition performance. Within a certain range, the recognition time gradually improves as the number of autoencoder layers increases. However, the improvement of recognition performance of the fusion network conforms to the above rules is an open question. Next, we discuss the effect of the number of autoencoder layers on multimodal recognition, designing a set of experiments on three databases. Three variants with different numbers of convolutional layers are constructed, where the number of layers of the autoencoder is set to two, which it means that both the encoder and decoder consist of one convolutional layer. Similarly, when the parameters are set to 4 or 6, the encoder is implemented by 2 and 3 convolutional layers, respectively.

According to the evaluation results in Table 3, the highest recognition rates were obtained for the FVC-HKP database and HDPR-310 database when the number of autoencoder layers was equal to 4. However, for the CAS-FVU database, which has a large variation in image quality, the fingerprint images have a large number of dry lines with some missing, while the vein images have clear and stable textures. Too shallow autoencoder layers would restrict our proposed method to a larger space of encoded common features. The model with six autoencoder layers achieved the best recognition results on this database, but too many parameters and a limited amount of training data would lead to an increase in the average recognition time and overfitting of the model. In the parameter selection experiments, an increase in the number of encoder layers will directly lead to an increase in the recognition time, which is one of the important metrics in biometric recognition tasks. The recognition rate decreases when the encoder is 6 layers, so we choose an encoder network architecture with 4 layers to better trade-off the time and accuracy metrics.



TABLE 4. Comparison of unimodal and multimodal recognition performance.

Modal	FVC-HKP		CAS-FVU		HDPR-310		Test time (ms)
	Acc (%)	EER(%)	Acc (%)	EER(%)	Acc (%)	EER(%)	
FP	89.11 ± 0.80	2.08	92.94 ± 0.44	1.71	91.91 ± 0.77	1.18	2.99
FV	94.98 ± 0.46	1.40	95.85 ± 0.40	0.97	96.63 ± 0.34	0.78	2.49
FP + FV	98.84 ± 0.24	0.17	99.17 ± 0.17	0.11	98.67 ± 0.16	0.14	5.85

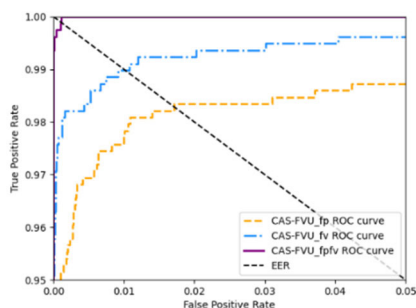


FIGURE 9. ROC curves on CAS-FVU database of fingerprint recognition model, finger vein recognition model and multimodal recognition model.

TABLE 5. Recognition performance of different backbone networks.

Backbone	Acc (%)			Test time (ms)
	FVC-HKP	CAS-FVU	HDPR-310	
S	98.36 ± 0.26	98.63 ± 0.25	98.57 ± 0.20	6.02
A	98.84 ± 0.24	99.17 ± 0.17	98.67 ± 0.16	5.85

S denotes the symmetric backbone network architecture with the same depth; A denotes an asymmetric backbone network architecture

D. SYSTEM PERFORMANCE COMPARISON WITH UNIMODAL RECOGNITION

For verifying the effectiveness of multimodal feature fusion, we construct two unimodal variants, a fingerprint recognition model and a finger vein recognition model, respectively. The same architecture as the unimodal stream in the backbone network and the softmax function is used for unimodal recognition.

Clearly, as shown in Figure 8, Figure 9, and Figure 10, the area under the receiver operating characteristic curve (ROC curve) of the multimodal recognition model is much larger than the best unimodal recognition result, demonstrating the effectiveness of our fusion method. Specifically, the Accuracy on the multimodal database is reduced by 3.86%, 3.32%, and 2.04%, respectively, compared with the best performance unimodal recognition model. The evaluation results are reported in Table 4. EER is the location on the ROC curve where the true positive rate and the false positive rate are equal, and the algorithm with the smallest EER performs best. Our general network architecture can effectively utilize the rich information of multimodal data for effective fusion, which greatly improves the security and accuracy of the biometric system within an acceptable recognition time.

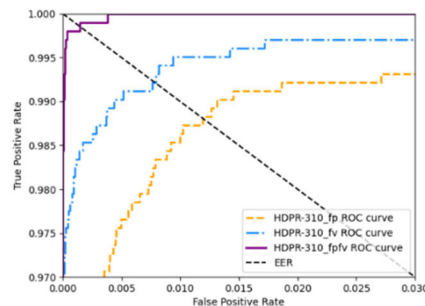


FIGURE 10. ROC curves on HDPR-310 database of fingerprint recognition model, finger vein recognition model and multimodal recognition model.

TABLE 6. Recognition performance with different fusion layers.

Fusion layers	Acc (%)			Test time (ms)
	FVC-HKP	CAS-FVU	HDPR-310	
1	98.33 ± 0.17	99.23 ± 0.30	98.10 ± 0.16	5.48
2	98.84 ± 0.24	99.17 ± 0.17	98.67 ± 0.16	5.85
3	98.89 ± 0.19	98.96 ± 0.18	97.95 ± 0.18	6.62

E. ABLATION STUDY OF FAB-NET

1) EFFECTIVENESS OF ASYMMETRIC ARCHITECTURE

To verify the effectiveness of the asymmetric architecture of the backbone network, we constructed a variant in which the fingerprint stream and the finger vein stream are both composed of 6 convolutional layers. Since the size of the output feature map after standardization of the unimodal stream consisting of 5 convolutional layers is large, which causes a significant increase in parameters and does not meet the advanced requirements of a lightweight fusion network, a control setup was not performed.

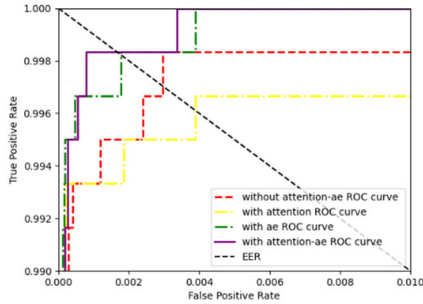
The results are reported in Table 5, where the recognition accuracy using the asymmetric architecture is higher than those using the normal symmetric network on all experimental databases. Although it is common that deeper networks tend to obtain better recognition accuracy, our FAB-Net efficiently learns discriminative intra-modal representations, resulting in higher recognition accuracy. Smaller size feature maps require fewer parameters of the fusion network, so less recognition time is able to meet the real-time requirements of the fusion recognition system.

2) EFFECTIVENESS OF FUSING MULTI-LAYER FEATURE MAPS

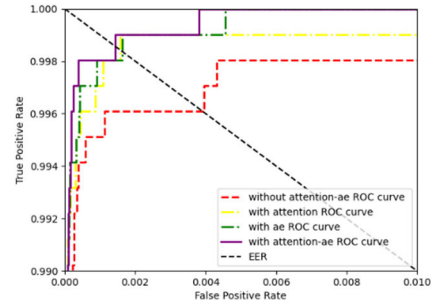
When multimodal features are fused, there may be multiple options for the number of fusion layers. To compare the effect

**TABLE 7. Recognition performance for ablation experiments on the proposed fusion model.**

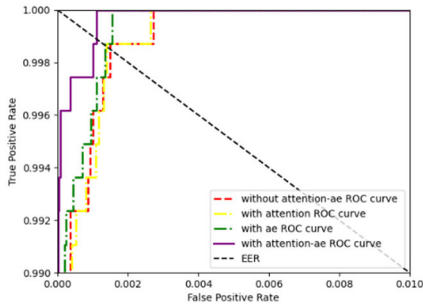
Attention	AE	FVC-HKP		CAS-FVU		HDPR-310		Test time (ms)
		Acc (%)	F1	Acc (%)	F1	Acc(%)	F1	
×	×	98.29 ± 0.46	0.988	97.77 ± 0.28	0.980	98.38 ± 0.23	0.987	4.86
✓	×	98.45 ± 0.30	0.990	98.03 ± 0.27	0.983	98.42 ± 0.17	0.987	5.46
×	✓	98.90 ± 0.28	0.993	98.39 ± 0.28	0.988	98.56 ± 0.19	0.988	5.67
✓	✓	98.84 ± 0.24	0.993	99.17 ± 0.17	0.994	98.67 ± 0.16	0.990	5.85



**FIGURE 11. ROC curves of ablation experiments on FVC-HKP database for the attention-based encoder fusion network.**



**FIGURE 13. ROC curves of ablation experiments on HDPR-310 database for the attention-based encoder fusion network.**



**FIGURE 12. ROC curves of ablation experiments on CAS-FVU database for the attention-based encoder fusion network.**

**TABLE 8. Recognition performance with different training procedures.**

Network trained in first stage	FVC-HKP	CAS-FVU	HDPR-310
None	96.01 ± 0.40	95.40 ± 0.52	97.16 ± 0.18
FVC-HKP_FP	98.84 ± 0.24	97.81 ± 0.27	97.99 ± 0.19
CAS-FVU_FP	98.27 ± 0.29	99.17 ± 0.17	98.04 ± 0.12
HDPR-310_FP	97.59 ± 0.36	97.66 ± 0.21	98.67 ± 0.16

of different fused layers on system performance, we constructed a recognition network without a dense structure and a variant with three dense feature layers for multi-stage learning.

According to the evaluation results in Table 6, when the number of fusion layers is two, all recognition accuracies reach advanced levels, and the best recognition accuracies are achieved on most of the databases, specifically FVC-HKP database and database HDPR-310. For CAS-FVU database, the fingerprint image is of poor quality while the finger vein image is clearer and more stable, only one layer can achieve a better fusion recognition effect. Combining the recognition performance of three databases, it is considered that two-layer fusion layer balances the needs of biometric systems for high accuracy and timeliness. The utilization of multi-layer feature maps effectively reduces the width of the backbone network with fewer parameters and alleviates the overfitting and gradient disappearance phenomena.

**F. ABLATION STUDY OF AFE-NET ON MULTIMODAL REDUNDANT DATA**

We first constructed two variants separately as follows. Removing the attention module and our autoencoder (AE),

so that the backbone network is retained, is the first type of variant we constructed. The variant that retains the attention module but removes the proposed AE is used to compare the impact of the encoder module on system performance.

Table 7 shows that our fusion network outperforms the backbone network on all multimodal databases we tested, with average recognition rates improving by 0.55%, 1.4%, and 0.29%, respectively. A more efficient fusion model is constructed based on the attention-based encoder network with fusion similarity. In addition, the advantage of the method is even more pronounced in the CAS-FVU database. The area under the ROC curve of the attention-based encoder network is larger than other variants as shown in Figure 11, Figure 12, and Figure 13. That is to say, the proposed algorithm not only has a higher recognition rate, but also a lower false rejection rate than the backbone network. These results suggest that the linear representation based on concatenation is not the most effective method for exploring the complementarity of multi-biometric features, and the nonlinear fusion approach using an encoder network can mitigate the redundancy of multimodal information for improved recognition accuracy.

There are two main reasons for the recognition performance improvement obtained by our method. One of the

**TABLE 9.** Comparison with state-of-the-art algorithms on multimodal finger databases.

Method	FVC-HKP		CAS-FVU		HDPR-310		Test time (ms)
	Acc (%)	F1	Acc (%)	F1	Acc (%)	F1	
Multimodal_DL (2015) [42]	97.95 ± 0.33	0.983	94.73 ± 0.30	0.951	98.51 ± 0.09	0.987	73.81
PCA Fusion (2019) [15]	98.65 ± 0.15	0.988	96.08 ± 0.15	0.963	98.73 ± 0.10	0.988	5.06
StructureIII-L5_con (2020) [24]	97.61 ± 0.59	0.983	96.64 ± 0.29	0.973	98.21 ± 0.22	0.986	9.30
LC-CNN (2021) [13]	93.22 ± 0.83	0.948	94.50 ± 0.24	0.948	95.71 ± 1.11	0.968	-
FPV-Net* (2022) [14]	93.75 ± 0.18	0.949	94.76 ± 0.09	0.949	96.52 ± 0.09	0.940	14.35
This paper	98.84 ± 0.24	0.993	99.17 ± 0.17	0.994	98.67 ± 0.16	0.990	5.85

\* Since the loss function is unspecified, we use the cross-entropy loss, and reduce the initial learning rate.

reasons is that the encoder-based fusion network effectively overcomes the performance degradation caused by multimodal information redundancy. Another reason is that the difference in the representational power of features within the same modality and between different modalities is taken into account, and cross-modal information interaction is achieved by reweighting via channel attention.

#### G. DISCUSS OF THE IMPACT OF TRAINING PROCEDURE

To verify the effectiveness of the proposed assisted training method from unimodal to multimodal, two types of variants were constructed separately. The first training method is to train the overall network directly with the same iterations as the multi-stage training method. The second training variant uses fingerprint modalities from the other databases as the first stage of unimodal training, and then performs the overall network training in the second stage.

According to the results in Table 8, it is shown that our proposed assisted training method can produce more accurate results, and pre-training the corresponding unimodal data with complex characteristics can help construct more effective fusion models. In addition, it is experimentally demonstrated that all models using unimodal pretraining outperform the models that are not pretrained, even though their total number of iterations is the same.

#### H. COMPARISON WITH THE STATE-OF-THE-ART ALGORITHMS

In this subsection, we compare the recognition accuracy, F1 Score, and average test time of various multimodal fusion recognition algorithms, as shown in Table 9. PCA fusion methods [15] that combine traditional algorithms with convolutional networks; Structure III-L5\_con [24] that uses one layer of convolutional features for concatenation; Multimodal DL [42] that adopts a bilinear pooling layer, LC-CNN [13] that introduce local operators in the shallow layers of the neural network; FPV-Net [14] that performs fusion based on attention, and many other advanced algorithms provide new ideas for multimodal fusion. The setup of the comparative experiment is slightly different from that in the literature. Due to GPU memory limitations, the Multimodal DL approach used Nvidia 3090 with 24GB RAM. The LC-CNN [13]

method uses a stepwise experiment, so recognition times were not tested.

The recognition accuracy of our proposed method reaches the current state-of-the-art level and also shows an improvement in the average recognition time. Specifically, the accuracy in the multimodal experiments reaches  $98.84 \pm 0.24$ ,  $99.17 \pm 0.17$ , and  $98.67 \pm 0.16$ , respectively, with higher F1 scores than other multimodal recognition methods. By means of GPU, the recognition time of our proposed algorithm is 5.85ms, which can be significantly reduced to the same level as the traditional methods represented by PCA Fusion. We can observe a similar trend where the traditional methods require a shorter average recognition time than most of the deep learning-based methods. Also, some traditional methods were improved from the perspective of CNN features to obtain better recognition rates.

Specifically, both the PCA Fusion and our method aim to avoid the degradation of system performance due to redundant information through dimensionality reduction. However, while the PCA method linearly generates fused feature representations, our approach is able to fit more accurate multimodal features nonlinearly through multilayer convolution with cosine dissimilarity constraint. The literature [13], [15], [24] first analyzes the intra-modal feature representation separately, which can avoid the low recognition rate caused by the undertraining of complex multi-branch networks and the correlation of multimodal information that is not fully observed. FPV-Net learns common features through the attention module, but the direct training strategy it uses leads to insufficient training of a certain network branch, resulting in a lower recognition rate than other deep learning methods. Considering the correlation and redundancy of multimodal information simultaneously, our method is able to learn compact fusion features with identity discriminative power more efficiently, thus achieving advanced recognition accuracy on three databases.

#### V. CONCLUSION

In this paper, we propose an attention-based encoder network with fused similarity. In this way, the lightweight backbone network extracts intra-modal features fast and accurately, and the encoder fusion network enhances the complementarity and reduces the redundancy of multimodal features, which

captures a more discriminative common representation. By the design of the loss function, the autoencoder is trained together with the overall recognition model, encouraging both class differentiability and expanding inter-class distance. In addition, the proposed assisted training approach from unimodal to multimodal enables adequate training so that multimodal data provide as much useful information as possible. The experimental results show that the method proposed in this paper is an effective feature fusion strategy that reaches the current advanced level of multimodal feature recognition.

Most existing deep learning-based biometric recognition methods ignore the relationship between recognition results and feature fusion effectiveness, which fails to achieve the global optimum well. The asymmetric network with fused similarity proposed in this paper can well meet the requirements of accuracy and recognition time. Further, we expect to propose an unsupervised biometric fusion recognition method based on autoencoders toward practical applications. In future work, our approach is expected to jointly model other non-picture forms of multimodal data, such as human voice and gait.

## REFERENCES

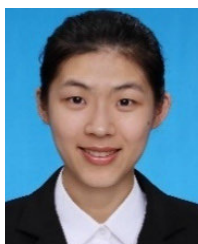
- [1] F. Boutros, N. Damer, F. Kirchbuchner, and A. Kuijper, "ElasticFace: Elastic margin loss for deep face recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1578–1587.
- [2] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, and X. Zhang, "Recent advancements in finger vein recognition technology: Methodology, challenges and opportunities," *Inf. Fusion*, vol. 79, pp. 84–109, Mar. 2022.
- [3] S. Li, R. Ma, L. Fei, and B. Zhang, "Learning compact multirepresentation feature descriptor for finger-vein recognition," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 1946–1958, 2022.
- [4] M. Shahzad, S. Wang, G. Deng, and W. Yang, "Alignment-free cancelable fingerprint templates with dual protection," *Pattern Recognit.*, vol. 111, Mar. 2021, Art. no. 107735.
- [5] A. Sepas-Moghaddam and A. Etemad, "Deep gait recognition: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 264–284, Jan. 2023.
- [6] S. Jia, X. Li, C. Hu, G. Guo, and Z. Xu, "3D face anti-spoofing with factorized bilinear coding," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 10, pp. 4031–4045, Oct. 2021.
- [7] S. K. S. Modak and V. K. Jha, "Multibiometric fusion strategy and its applications: A review," *Inf. Fusion*, vol. 49, pp. 174–204, Sep. 2019.
- [8] R. Brunelli and D. Falavigna, "Person identification using multiple cues," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 17, no. 10, pp. 955–966, Oct. 1995.
- [9] A. Ross and A. Jain, "Information fusion in biometrics," *Pattern Recognit. Lett.*, vol. 24, no. 13, pp. 2115–2125, 2003.
- [10] H. Purohit and P. K. Ajmera, "Optimal feature level fusion for secured human authentication in multimodal biometric system," *Mach. Vis. Appl.*, vol. 32, no. 1, pp. 1–12, Jan. 2021.
- [11] C. H. Chan, M. A. Tahir, J. Kittler, and M. Pietikäinen, "Multiscale local phase quantization for robust component-based face recognition using kernel fusion of multiple descriptors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 5, pp. 1164–1177, May 2013.
- [12] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
- [13] S. Li, B. Zhang, S. Zhao, and J. Yang, "Local discriminant coding based convolutional feature representation for multimodal finger recognition," *Inf. Sci.*, vol. 547, pp. 1170–1181, Feb. 2021.
- [14] H. Ren, L. Sun, J. Guo, and C. Han, "A dataset and benchmark for multimodal biometric recognition based on fingerprint and finger vein," *IEEE Trans. Inf. Forensics Security*, vol. 17, pp. 2030–2043, 2022.
- [15] L. Wang, H. Zhang, and J. Yang, "Finger multimodal features fusion and recognition based on CNN," in *Proc. IEEE Symp. Ser. Comput. Intell. (SSCI)*, Dec. 2019, pp. 3183–3188.
- [16] W. Yang, S. Wang, J. Hu, G. Zheng, and C. Valli, "A fingerprint and finger-vein based cancelable multi-biometric system," *Pattern Recognit.*, vol. 78, pp. 242–251, Jun. 2018.
- [17] C. Fang, H. Ma, and Z. Yang, "A novel dual-modal biometric recognition method based on weighted joint sparse representation classification," in *Proc. Chin. Conf. Biometric Recognit.*, 2021, pp. 3–10.
- [18] C. Kamlaskar and A. Abhyankar, "Iris-fingerprint multimodal biometric system based on optimal feature level fusion model," *AIMS Electron. Electr. Eng.*, vol. 5, no. 4, pp. 229–250, 2021.
- [19] C. Kamlaskar, S. Deshmukh, S. Gosavi, and A. Abhyankar, "Novel canonical correlation analysis based feature level fusion algorithm for multimodal recognition in biometric sensor systems," *Sensor Lett.*, vol. 17, no. 1, pp. 75–86, Jan. 2019.
- [20] Y. Shi, Y. Pan, D. Xu, and I. W. Tsang, "Multiview alignment and generation in CCA via consistent latent encoding," *Neural Comput.*, vol. 32, no. 10, pp. 1936–1979, Oct. 2020.
- [21] J. Li and P. Fang, "FVGNN: A novel GNN to finger vein recognition from limited training data," in *Proc. IEEE 8th Joint Int. Inf. Technol. Artif. Intell. Conf. (ITAIC)*, May 2019, pp. 144–148.
- [22] H. Qu, H. Zhang, J. Yang, Z. Wu, and L. He, "A generalized graph features fusion framework for finger biometric recognition," in *Proc. Chin. Conf. Biometric Recognit.*, 2021, pp. 267–276.
- [23] Z. Zhong, W. Gao, and M. Wang, "A multimodal fusion method based on a rotation invariant hierarchical model for finger-based recognition," *KSII Trans. Internet Inf. Syst.*, vol. 15, no. 1, pp. 131–146, 2021.
- [24] L. Wang, "Research on multimodal fusion recognition method for fingers based on CNN," M.S. thesis, Civil Aviation Univ. China, Tianjin, China, 2020, pp. 27–46.
- [25] M. Leghari, S. Memon, L. Das Dhomeja, A. H. Jalbani, and A. A. Chandio, "Deep feature fusion of fingerprint and online signature for multimodal biometrics," *Computers*, vol. 10, no. 2, p. 21, Feb. 2021.
- [26] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin, "Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion," *IEEE Trans. Image Process.*, vol. 29, pp. 4296–4307, 2020.
- [27] K. Su, G. Yang, B. Wu, L. Yang, D. Li, P. Su, and Y. Yin, "Human identification using finger vein and ECG signals," *Neurocomputing*, vol. 332, pp. 111–118, Mar. 2019.
- [28] R. S. Kuzu, E. Maiorana, and P. Campisi, "Vein-based biometric verification using densely-connected convolutional autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 1869–1873, 2020.
- [29] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.
- [30] M. Abavisani and V. M. Patel, "Deep multimodal subspace clustering networks," *IEEE J. Sel. Topics Signal Process.*, vol. 12, no. 6, pp. 1601–1614, Dec. 2018.
- [31] G. Sahu and O. Vechtomova, "Adaptive fusion techniques for multimodal data," 2019, *arXiv:1911.03821*.
- [32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, "Squeeze-and-excitation networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
- [33] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.
- [34] Y. Dai, F. Gieseke, S. Oehmcke, Y. Wu, and K. Barnard, "Attentional feature fusion," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, Jan. 2021, pp. 3560–3569.
- [35] Y. Lv, W. Zhou, J. Lei, L. Ye, and T. Luo, "Attention-based fusion network for human eye-fixation prediction in 3D images," *Opt. Exp.*, vol. 27, no. 23, pp. 34056–34066, 2019.



- [36] S. Verma, C. Wang, L. Zhu, and W. Liu, "Attn-HybridNet: Improving discriminability of hybrid features with attention fusion," *IEEE Trans. Cybern.*, vol. 52, no. 7, pp. 6567–6578, Jul. 2022.
- [37] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
- [38] R. Cappelli, M. Ferrara, A. Franco, and D. Maltoni, "Fingerprint verification competition 2006," *Biometric Technol. Today*, vol. 15, pp. 7–9, Jul./Aug. 2007.
- [39] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–2244, Apr. 2012.
- [40] *Casia Fingerprint Database*.
- [41] M. S. M. Asaari, S. A. Suandi, and B. A. Rosdi, "Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3367–3382, Jun. 2014.
- [42] A. Eitel, J. T. Springenberg, L. Spinello, M. Riedmiller, and W. Burgard, "Multimodal deep learning for robust RGB-D object recognition," in *Proc. IEEE/RISJ Int. Conf. Intell. Robots Syst. (IROS)*, Sep. 2015, pp. 681–687.



**HUI MA** received the Ph.D. degree in pattern recognition and intelligent system from Harbin Engineering University, China, in 2011. Until 2017, she conducted her postdoctoral research work in pattern recognition at Heilongjiang University, China, where she is currently an Associate Professor. Her current research interests include image processing, pattern recognition, machine learning.



**YIWEI HUANG** received the B.E. degree from Heilongjiang University, Harbin, China, in 2020, where she is currently pursuing the master's degree. Her current research interests include biometrics, pattern recognition, computer vision, and deep learning.



**MINGYANG WANG** received the B.E. degree from the Liaoning University of Science and Technology, Liaoning, China, in 2019. He is currently pursuing the master's degree in control science and engineering with Heilongjiang University. His main research interests include pattern recognition, computer vision, and pedestrian detection.

• • •