IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 12 April 2024, accepted 27 May 2024, date of publication 30 May 2024, date of current version 6 June 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3407155

RESEARCH ARTICLE

FV-DMHN: Dual Multi-Head Network for Finger Vein Recognition

ZHIJIN AN, XIAOKUI REN, AND ZHIYONG TAO
School of Electronics and Information Engineering, Liaoning Technical University, Huludao 125105, China

Corresponding author: Zhijin An (13081244451@163.com)

This work was supported in part by the Applied Basic Research Project of the Department of Science and Technology of Liaoning Province under Grant 2022JH2/101300274, in part by the Exploration of an Integrated Practice and Innovation Capability Cultivation Model for Professional Master's Students with a Six-in-One and Industry-Education Integration Approach: the Educational Department of Liaoning Province under Grant LNYJC2023117, and in part by the Educational Department of Liaoning Province under Grant LJKMZ20220679.

ABSTRACT Deep learning-based finger vein image recognition methods usually suffer from high complexity, insufficient global information extraction, and overfitting. The use of lightweight networks can significantly reduce the accuracy owing to the reduction in model parameters. For this reason, this paper proposes a Dual Multi-Head neural Network for Finger Vein Recognition (FV-DMHN), which combines the Multi-Head Self-Attention (MHSA) mechanism with the Multi-Head Convolutional Network(MHCN) to increase the training efficiency of the network while expanding the CNN horizon. The inverted residual structure is also used in the network to enhance the expressive power of the network. The algorithm achieves recognition accuracies of 99.81%, 99.67%, 99.69%, and 99.83% on three publicly available datasets, FV-USM, SDUMLA-HMT, THU-FVFDT2, and self-built datasets FV-SIPL, respectively, with an average equal-error rate of 0.339%, and the recognition time of a single image is only 2.63 ms. The experimental results show that the algorithm is superior to other methods in terms of accuracy and average equal error rate, at the same time, it not only reduces the number of network parameters and computational complexity but also achieves excellent recognition speed.

INDEX TERMS Finger vein recognition, multi-head convolutional network, inverted residual module, multi-head self-attention.

I. INTRODUCTION

With the increasing demand for information security, biometrics is becoming more prevalent in user authentication systems. Biometric methods include fingerprints [1], iris [2], palm prints [3], faces [4], and finger veins [5], and others. Unlike other biometric technologies, finger vein recognition technology is highly secure, offers live recognition, and is non-contact. The equipment for finger vein recognition is simple to operate, smaller in size, and cheaper in cost, making it conducive to wide-scale promotion and application.

The finger vein recognition process typically involves four steps: image acquisition, pre-processing, feature extraction, and matching. Efficiently extraction of finger vein features remains the main challenge in finger vein recognition owing to the impact of finger thickness, near-infrared light angle, and finger position during acquisition. These factors can cause shadow or highlight phenomena and blur the finger vein outline. Currently, there are two main types of finger vein feature extraction algorithms: traditional and deep learning.

Before the wide application of deep learning technology, the research on finger vein recognition is mainly based on traditional image processing algorithms. The traditional recognition methods generally include methods based on subspace learning [6], methods based on vein patterns [7], methods based on detail point matching [8], and methods based on local features [9]. Among them, the methods based on subspace learning can filter the noise while reducing the global feature dimension, but they all extract features from a global perspective and lack the description of local feature information; however, the method based on vein pattern needs to segment the finger vein image, which is

The associate editor coordinating the review of this manuscript and approving it for publication was Nuno M. Garcia.

susceptible to the quality of the finger image; the method based on detail point matching can extract the structure of the blood vessel intersections, endpoints, etc. as the feature points very well, but usually performs poorly on low-quality images owing to the presence of spurious detail features; However, local feature-based methods are robust to image contrast and illumination changes, and only consider the relationship between the target pixel and its surrounding pixels, while ignoring the hidden relationships between surrounding neighboring pixels. Traditional finger vein recognition mainly uses manually designed feature-based methods for recognition, which is a complicated process, that makes it difficult to characterize the various gesture changes of the finger, and the algorithms are less robust and migratory.

Currently, with the development of deep learning, Convolutional Neural Networks (CNN) [10], Deep Belief Networks (DBNS) [11], and Generative Adversarial Networks(GAN), have been used to learn robust features from raw pixel images, and have shown good performance. Yang et al. [12] developed a finger vein recognition system based on a CNN, which is capable of processing finger vein images of varying quality while obtaining stable and highly accurate recognition performance. Das et al. [13] proposed a merged convolutional neural network, which merged multiple short-path CNN structures, to extract the features from images of varying quality and fuse them. Boucherit et al. [14] proposed an iterative Deep Belief Network (DBNS) to extract vein features based on initial labeled data which is automatically generated using very limited a priori knowledge and iteratively corrected by the DBNS. Hu et al. [15] proposed a deep convolutional neural network model called Finger Vein Network (FV-Net) to design the FV-Net architecture and proposed A template-like matching strategy was proposed to extract features with spatial information. Huang et al. [16] proposed a method called ''DeepVein'' for finger vein validation based on deep convolutional neural networks with good results. Yang et al. [17] used a generative adversarial network to learn the feature representation of finger veins. vein feature representation, Shahreza and Marcel [18] used an autoencoder to learn the feature labeling of finger veins. Qin and El-Yacoubi [19]introduced deep learning models for finger vein quality assessment. In these studys, deep learning models have shown excellent results. Most of the above research methods are devoted to increasing the depth or width of the neural network to improve the results, although the recognition accuracy has been improved to a certain extent, the network is too deep or too wide, which is a great test of the computer computing power, and at the same time, it is easy to cause the phenomenon of overfitting.

With the continuous development of convolutional neural networks, it is realized that constantly stacking the networks will not always result in performance improvement, but will instead make the performance decrease. In recent years, lightweight networks have gradually become a focus of research. lightweight networks not only run fast and with low complexity, but also facilitate the embedding of mobile devices, which can advance the further development of image recognition. Radzi et al. [20] used a four-layer CNN to recognize finger veins and achieved a high recognition accuracy on a self-built dataset. Xie and Kumar [21] proposed a new method for finger vein authentication using CNN and supervised discrete sequences. The proposed method not only achieves excellent results but also significantly reduces the model size. Zhao et al. [22] proposed a lightweight convolutional neural network based on a central loss function and dynamic regularization, which not only reduced the false recognition rate but also accelerated the convergence speed. Lu et al. [23] proposed a lightweight model based on Vision Transformer (ViT) T2T- ViT, where tokens input into ViT are subjected to Reshape and Soft Split operations to reduce the dimensionality of tokens, making the overall model more lightweight, but the model loses a certain amount of accuracy. Although all of the above lightweight networks have achieved good recognition results, the reduction in the number of network layers also represents the acquisition of less global information, which can only be based on the shallow network extracted edges, corners, points, and some vein texture change information to discriminate, the global information is not enough to grasp, and it is easy to ignore some important details of the information is not conducive to improving the recognition performance.

To address the above problems, this study proposes a Dual Multi Head Net for Finger Vein Recognition (FV-DMHN), which combines the Multi-Head Self-Attention (MHSA) mechanism with a Multi-Head Convolutional Neural network (MHCN) to increase the training efficiency of the network while expanding the CNN field of view. In this work, we use a Multi-Head Convolutional Neural Network to extract the local information of finger vein images, and then add a Dropout layer to suppress the overfitting phenomenon. The introduction of the inverted residual structure can further optimize the model, accelerate its convergence speed, and improve the accuracy and robustness of the model. Finally, the multi-head self-attention mechanism module is utilized to extract the global information and enhance the information flow in the space. The proposed method in this paper is tested on finger vein datasets (FV-SIPL) collected by Universiti Teknologi Malaysia (FV-USM), Shandong University Machine Learning and Applications Group (SDUMLAHMT), Tsinghua University (THU-FVFDT2) and Signal and Information Processing Laboratory of Liaoning Technical University. The experimental results show that this method can effectively extract finger vein features and exhibits excellent recognition performance.

II. DATABASES AND SAMPLE AUGMENTATION

A. SELF-CONSTRUCTED DATASETS

The FV-SIPL dataset was captured using a low-cost near-infrared finger vein acquisition sensor designed in the laboratory. Images were acquired from 27 volunteers, all of whom were students and faculty members of Liaoning

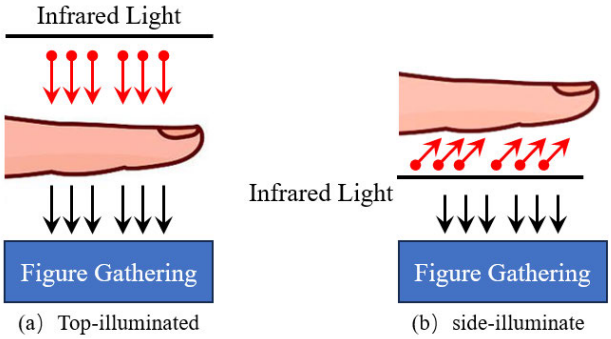


FIGURE 1. Principles of different light transmittance collection devices.

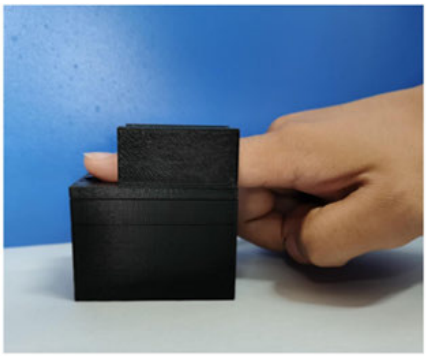


FIGURE 2. FV-SIPL collector.

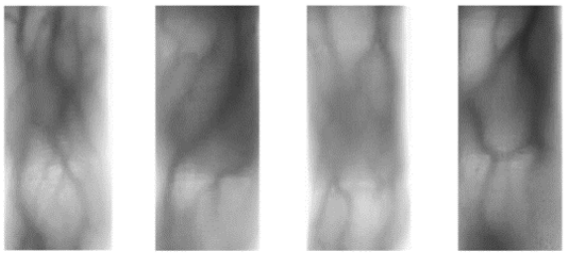


FIGURE 3. Sample image of finger vein.

Technical University. Each volunteer provided 4 fingers (index and middle fingers of the right and left hands), and each finger was acquired 12 times, generating a total of 1296 finger vein images with a finger vein size of 176 × 415 pixels. The acquisition details of the FV-SIPL dataset are as follows:

(1) Light transmission acquisition: Finger vein image acquisition devices are generally divided into light transmission acquisition and light reflection acquisition. Compared with the reflective acquisition, the infrared light of the transmissive acquisition is in a relatively closed environment, which is more conducive to the acquisition of high-quality vein images. Light transmission acquisition can be subdivided into apical illumination and parietal illumination (left and right sides), as shown in Figure. 1. If the light illumination is uniform, the contrast between the vein region and the non-vein region may be greatly reduced, resulting in poor-quality finger vein pictures being collected. The finger vein collection device constructed in the laboratory was adjusted in angle several times, and finally, a certain angle of side light transmission method was adopted, as shown in Fig. 1(b).

(2) Selection of sensors: The acquisition device adopts 850nm near-infrared LEDs and embeds infrared high transmittance filters to eliminate background noise and visible light interference. First, a 1080p resolution near-infrared camera is used to take pictures; subsequently, the LED light intensity is automatically adjusted according to the brightness of the pictures through pulse width modulation (PWM); then the pictures were taken again to obtain clear finger vein images; finally, the images were transmitted to the microcontroller for storage, calculation, and matching. The specific finger vein acquisition image samples and acquisition devices are shown in Figure. 2. and Figure. 3.

B. PUBLIC DATASETS

(1) FV-USM: This dataset was collected from 123 volunteers, each of whom provided 4 fingers (index and middle fingers of the right and left hands), and a total of 492 finger categories were obtained. The acquisition was divided into two phases, and each finger was acquired six times in each phase, and a total of 5904 finger vein images were collected.

(2) SDUMLA-HMT: This dataset acquisition device was made by the Joint Laboratory of Intelligent Computing and Intelligent Systems of Shandong University, which extracted the index, middle, and ring fingers of both hands of 106 volunteers and repeated the collection 6 times for each finger, and a total of 3816 finger vein images were collected.

(3) THU-FVFDT2: This database provides ROI of finger vein images from 610 subjects. finger vein ROI images were captured in two sessions at intervals of 3 days to 1 week, one image per finger per session, for a total of 1220 finger vein images. The database provides finger vein images normalized to 200 × 100 pixels.

C. DATA PROCESSING

The range and resolution of finger vein images captured by different devices are inconsistent, and the captured samples contain backgrounds unrelated to finger vein recognition, which can cause further interference in finger vein recognition. To extract more meaningful information in the subsequent finger vein recognition methods, the images are preprocessed, taking FV-USM as an example, as shown in Fig.4. Firstly, the contour information of the finger vein is obtained by the Sobel edge detection operator [24]. Then the median filter is used to smooth the image, suppress the noise, and eliminate the sharpness phenomenon, the ROI image is obtained by cutting the boundary of the image. Finally, the input image is normalized by the bicubic interpolation algorithm [25]. In this study, to prevent the

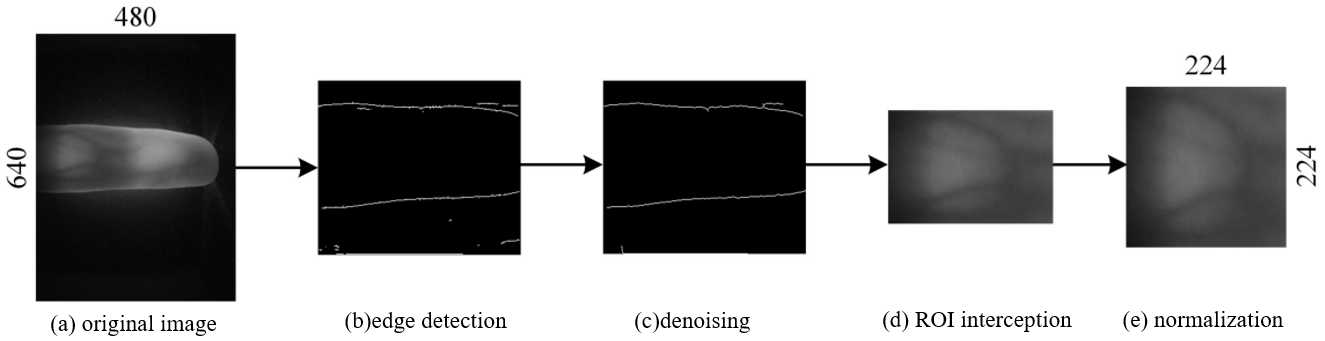


FIGURE 4. Finger vein image preprocessing process.

overfitting phenomenon during the training process owing to the small amount of data, image enhancement was performed on the training set, and data expansion was performed by rotating all the images in the training set by 90 degrees, flipping them, and increasing the contrast and brightness. Because the self-constructed dataset, FV-SIPL contains fewer non-finger regions and no obvious flipping, no ROI extraction operation is needed, and only image normalization and image enhancement are performed on it, as in the case of the THU-FVFDT2 dataset.

III. MODEL ARCHITECTURE

As shown in Fig.5, to improve the recognition performance of finger vein images, we proposed a Dual-Multi Head neural Network (FV-DMHN)for finger vein image recognition. Where Stage 1 is a standard 3×3 convolution and the output layers include the Average Pooling Layer, Dropout Layer, and Fully Connected Layer. Since the use of Skip connection requires that the input and output feature maps be the same size, and feature map down sampling and channel counting are required between each Stage, none of the first modules of Stages 2-5 contain Skip connection. The FV-DMHN first performs on-channel augmentation by a standard 3×3 convolution. The DW convolution kernel of 3×3 in the inverted residual structure is then utilized to improve the extraction of detailed features. Then the local feature information of the image is extracted by the Multi-Head Convolutional Network (MHCN) while constant mapping of the features is performed using jump connections. Aggregation of global information is performed by inserting the Multi-Head Self-Attention Module (MHSA) to achieve distant dependency modeling in the last part of the feature extraction. Finally, the N-class probabilities of all samples are generated through the output layer to produce the recognition results.

A. MULTI-HEAD CONVOLUTIONAL NETWORK

Inspired by the multi-head self-attention mechanism in the Transformer architecture, a Multi-Head Convolutional Network(MHCN), is constructed using group convolution and 1×1 dot convolution. As shown in Fig.6, the multi-head paradigm is used to construct convolutional attention that jointly attends to information from different representation subspaces at different locations for effective local representation learning. Group convolution can divide the input features into multiple groups and then perform convolution operations within each group, which can better preserve the local information of the input features and thus enhance the ability of the model to represent detailed information. Combining group convolution and 1×1 pointwise convolution allows for the construction of an efficient, lightweight, and accurate multi-head convolution structure, and has yielded good results in a variety of network computational complexities and performance. The use of a multi-head convolutional structure improves the generalization ability and learning efficiency of the model. In this section, the number of groups of all group convolutions is set to 1/4 of the number of input channels, and efficient BN and ReLU activation functions are added behind the group convolution. The BN can make the inputs of the network have a certain property of normal distribution, which enhances the stability of the network learning and the convergence speed. The ReLU activation function can add nonlinearities and enhance the ability of feature expression. It accelerates the inference speed based on maintaining the effective extraction of local features. In addition, the 1×1 convolution in the figure performs dimensionality adjustment to ensure the smoothness of jump connections and subsequent feature fusion.

B. INVERTED RESIDUAL MODULE

The inverted residual module was proposed by Sandler et al. in MobileNetV2 [26]. To inhibit the overfitting phenomenon during the training process, a Dropout layer is added on top of it, as shown in Fig.7. The Dropout layer is utilized to randomly discard some neurons to effectively suppress the overfitting phenomenon. The module consists of two PW convolutions, a DW convolution with a convolution kernel size of 3×3, a Dropout layer, and a jump connection.

The module firstly enhances the number of channels by the first PW convolution according to the expansion factor, then reduces the number of model parameters and computation amount while extracting the features by DW convolution,

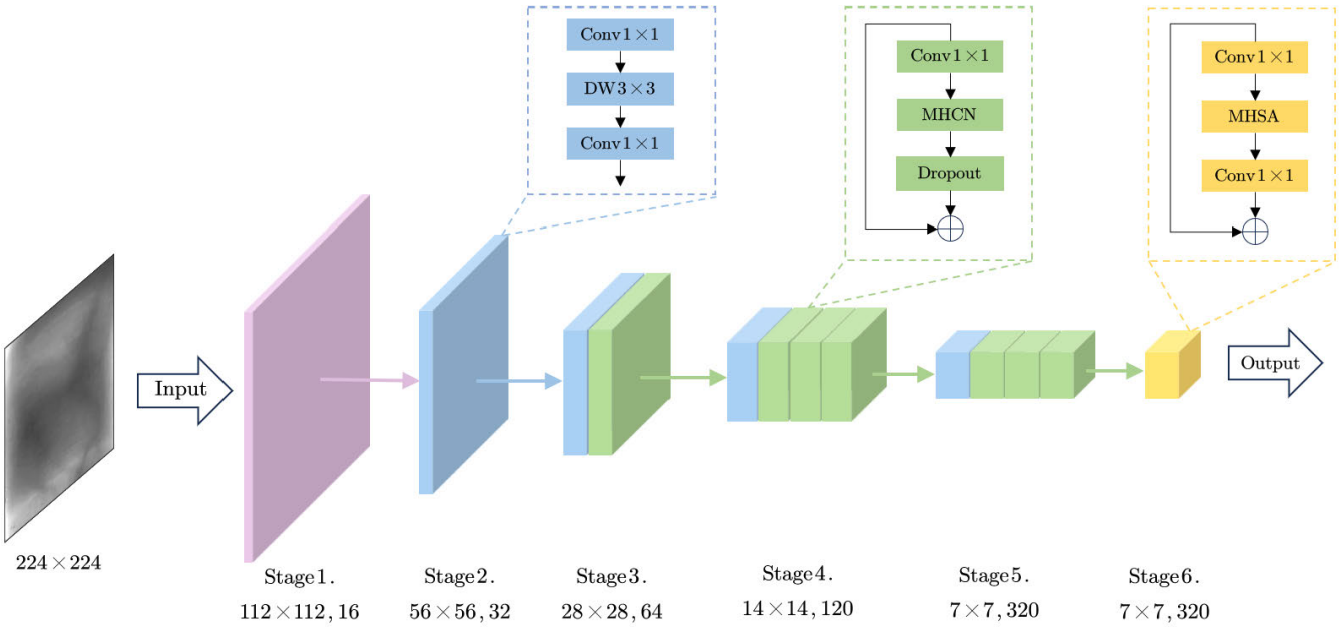


FIGURE 5. The overall architecture of the FV-DMHN model.

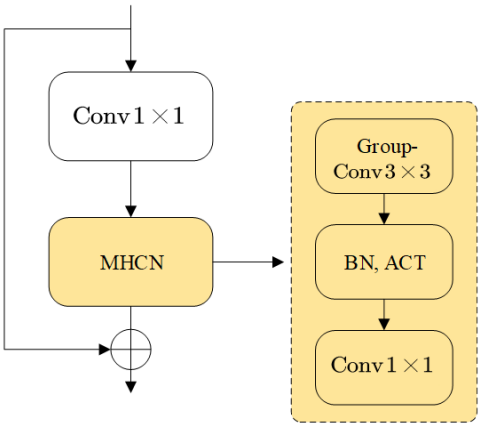


FIGURE 6. Multi-Head Convolutional Network Structure.

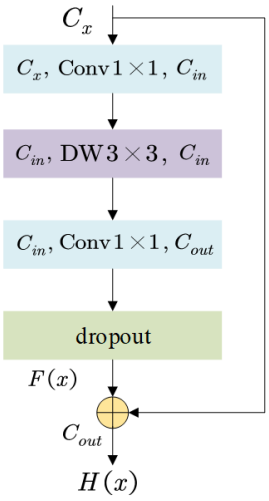


FIGURE 7. Inverted Residual module.

then inhibits the overfitting phenomenon by using the Dropout layer, and finally reduces the information loss of the activation function in the process of channel combining by utilizing the linear activation function according to the characteristics of the inverted residual structure. Where C_x is the number of input channels of the inverted residual module, and C_{in} in is equal to the product of the number of input channels and the expansion factor. The inverted residual structure maps the low-level features to the high-level network through jump connections, so that the input information bypasses the output, which ensures the integrity of the information. The Skip connection is mathematically defined as:

$$H(x) = F(x) + x \tag{1}$$

where $F(\cdot)$ is a function containing convolution, pooling, and correction linear cell operations, x is the input feature map, and $H(x)$ is the output of the inverted residual structure. The use of DW convolutional layers and skip connections in CNN makes network computation easy and helps to learn effective features from images and improve model performance.

C. MULTI-HEAD SELF-ATTENTION MODULE

The proposed model not only wants to be able to extract global information, but also wants to be able to focus on the focused information, so it incorporates a multi-head self-attention mechanism module [27] in the network framework, as shown in Fig.8. The highlighted blue box in the figure is

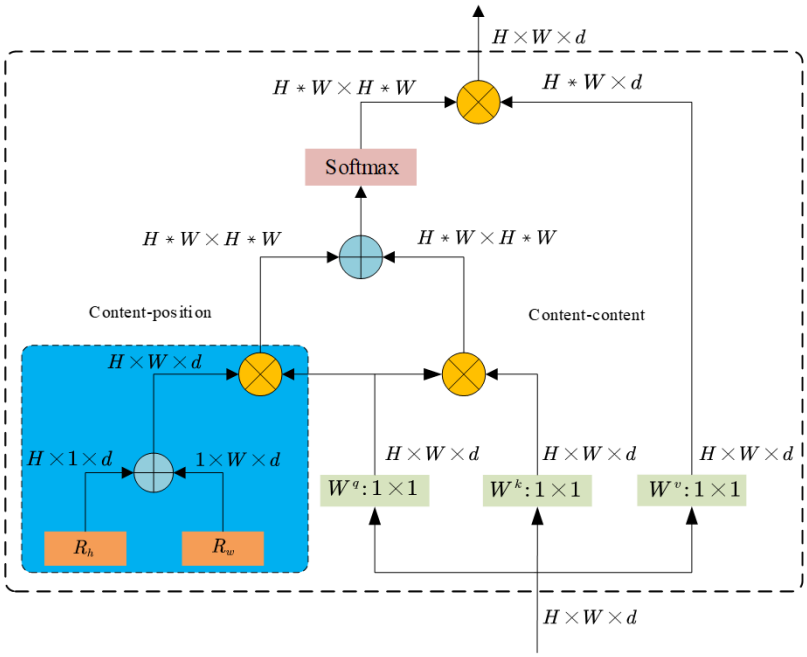


FIGURE 8. Multi-Head Self-Attention module.

the relative position encoding, where R_h and R_w represent the height and width of having segmented relative position encoding on the 2D feature map, respectively.

Combining the self-attention mechanism with position coding in the network architecture allows MHSA to take into account both the content information of image features and the relative distances between the features, thus effectively correlating the information between the objects with the location awareness. The module adopts the 2D relative position coding from the literature [28], which treats position coding as spatial attention. The position encoding vector r and the key value vector k are multiplied with the query vector q, respectively, and summed to obtain spatially sensitive features so that the MHSA focuses on the appropriate region and converges more easily. In this paper, we use a self-attention mechanism with 8 heads to focus on different parts of the features and finally cascade them. The formula is as follows:

$$Multi - Head\,(q,k,v) = concat\,(head_1,\ldots,head_n)\,W^o \tag{2}$$

$$head_i = self - Attention\left(qw_i^q, kw_i^k, vw_i^v\right) \tag{3}$$

$$self - Attention\,(q,k,v) = softmax\left(qr^{\mathrm{T}} + qk^{\mathrm{T}}\right)v \tag{4}$$

where v is the attention value, W^Y,$Y \in \{q,k,v,o\}$ are the respective parameter matrices, and i denotes the i-th self-attention mechanism, $n = 8$. Considering that $O\left(m^2d\right)$ memory and computation are required to perform self-attention globally across m entities, the MHSA is merged into the lowest-resolution feature maps in the backbone network, the last layer of the entire feature extraction.

TABLE 1. Data information for each dataset.

Dataset	Total categories	Total images	Train images	Test images
FV-USM	492	5904	3936	1968
SDUMLA-HMT	636	3816	2544	1272
THU-FVFDT2	610	1220	610	610
FV-SIPL	108	1296	864	432

IV. EXPERIMENTS AND DISCUSSION

A. EXPERIMENTAL CONFIGURATION

The experimental environment for the methods in this study is the Linux operating system, and the graphics card used for training is NVIDIA GeForce RTX 3090 GPU with Python3.8 and pytorch1.7 framework. The datasets were divided according to the ratio of 2:1, except for the THU-FVFDT2 dataset where the training and test sets were equally divided. The information of which data is shown in Table 1. In addition, a total of 100 epochs were trained during this experiment.

Model hyperparameter tuning plays a crucial role in training network models. After several configurations, the model in this study uses a learning rate of 0.01; the batch_size is set to 16; the optimizer chooses Stochastic Gradient Descent (SGD) with a momentum value of 0.9; the expansion

TABLE 2. Comparison of accuracy with other methods.

	FV-USM	SDUMLA-HMT	THU-FVFDT2
Merge CNN [22]	96.15%	89.99%	/
DS-CNN [29]	/	98.00%	89.00%
Coding SA [30]	99.48%	95.91%	/
L-MRFB [31]	99.59%	98.90%	/
FV-ViT [32]	99.73%	94.51%	/
LFVRN-CE [33]	98.58%	97.75%	/
T2T-ViT [23]	99.48%	94.85%	/
FV-AF [34]	/	97.28%	/
FV-DMHN	**99.81%**	**99.67%**	**99.69%**

TABLE 3. Comparison of EER with other methods.

	FV-USM	SDUMLA-HMT	THU-FVFDT2
FV-GAN [17]	/	0.940%	1.120%
Siam CNN [35]	0.110%	0.750%	/
Coding SA [30]	0.091%	2.450%	/
FVRAS Net [36]	0.950%	1.170%	/
FV-ViT [32]	**0.068%**	1.003%	/
LFVRN-CE [33]	0.610%	1.420%	/
T2T-ViT [23]	2.460%	0.940%	/
FV-AF [34]	/	**0.030%**	/
FV-DMHN	0.087%	0.378%	**0.643%**

factor is [1, 5, 5, 5, 5, 1], and its value indicates Stage2-6 expansion factors for each stage.

B. EVALUATION DETAIL

The experiments use several evaluation metrics to assess the recognition performance of this paper's method, including the recognition accuracy, EER, and recognition time. Among them, the accuracy rate is the most commonly used evaluation index in image recognition, which represents the ratio of the number of correctly classified samples by the classifier to the total number of samples, and its mathematical expression is as follows:

$$A_{\text{accuracy}} = \frac{N_{\text{TP}} + N_{\text{TN}}}{(N_{\text{TP}} + N_{\text{FP}} + N_{\text{TN}} + N_{\text{FN}})} \tag{5}$$

where N_{TP} denotes the number of correct positive sample predictions, N_{TN} denotes the number of correct negative sample predictions, N_{FN} denotes the number of incorrect positive sample predictions, and N_{FP} denotes the number of incorrect negative sample predictions. ROC is a graphical representation of the model performance under various threshold settings and is a plot of the true positive rate (TPR) and false positive rate (FPR) plots on two parameters, where the formulas are given below:

$$P_{\text{TPR}} = \frac{N_{\text{TP}}}{N_{\text{TP}} + N_{\text{FN}}} \tag{6}$$

$$P_{\text{FPR}} = \frac{N_{\text{FP}}}{N_{\text{TP}} + N_{\text{TN}}} \tag{7}$$

EER is used to predetermine the threshold of the TPR and its f FPR. When both are equal, the common value is called Equal Error Rate (EER). The lower the EER value, the better is the performance of the biometric system.

C. COMPARATIVE EXPERIMENTS WITH DIFFERENT MODELS

To verify the effectiveness of the proposed method, accuracy and EER comparison experiments with novel finger vein recognition methods such as DS-CNN and L-MRFB and lightweight finger vein recognition methods such as FV-ViT and T2T-ViT are conducted on the publicly available datasets, FV-USM, SDUMLA-HMT, and THU-FVFDT2, and the results are shown in Table 2 and Table 3. The experimental results show that the FV-DMHN demonstrates a good recognition performance over other methods on multiple datasets. This indicates that the combination of the multi-head self-attention mechanism and multi-head convolutional network in this study can better highlight the important detailed information of the image and make the extracted features more recognizable.

TABLE 4. Comparison of the accuracy of classical deep learning methods.

	FV-USM	SDUMLA-HMT	THU-FVFDT2	FV-SIPL
MobileNetV2	95.44%	95.00%	51.57%	99.54%
EfficientNet	99.00%	98.67%	97.53%	99.30%
Next-Vit	98.56%	99.00%	98.78%	99.33%
VIT-B	68.67%	72.67%	53.14%	91.63%
ResNet-101	98.33%	98.43%	98.21%	99.00%
Conformer-B	98.12%	99.00%	96.63%	99.31%
Swim-T	98.00%	97.00%	90.58%	99.30%
FV-DMHN	**99.81%**	**99.67%**	**99.69%**	**99.83%**

To evaluate the effectiveness of the method more rationally, it is compared with popular lightweight network models such as MobileNetV2 [26] and EfficientNet [37], purely attentional mechanism network models such as Next-ViT [38] and Swin-T [39], the classical CNN model ResNet-101 [40], and novel CNN vs. Transformer combined networks such as Conformer-B [41] and Mobile-ViT [42] are compared in terms of recognition accuracy. Table 4 shows the results of the accuracy comparison. From the table, it can be seen that FV-DMNH shows the best recognition accuracy on all four datasets. Due to the inevitable factors in the process of finger vein acquisition, a small number of low-quality finger vein images are generated. Although some difficulties can be solved by preprocessing, it is still impossible to obtain the most perfect recognition rate. Therefore, the collection of better finger vein images is also a focus of future work.

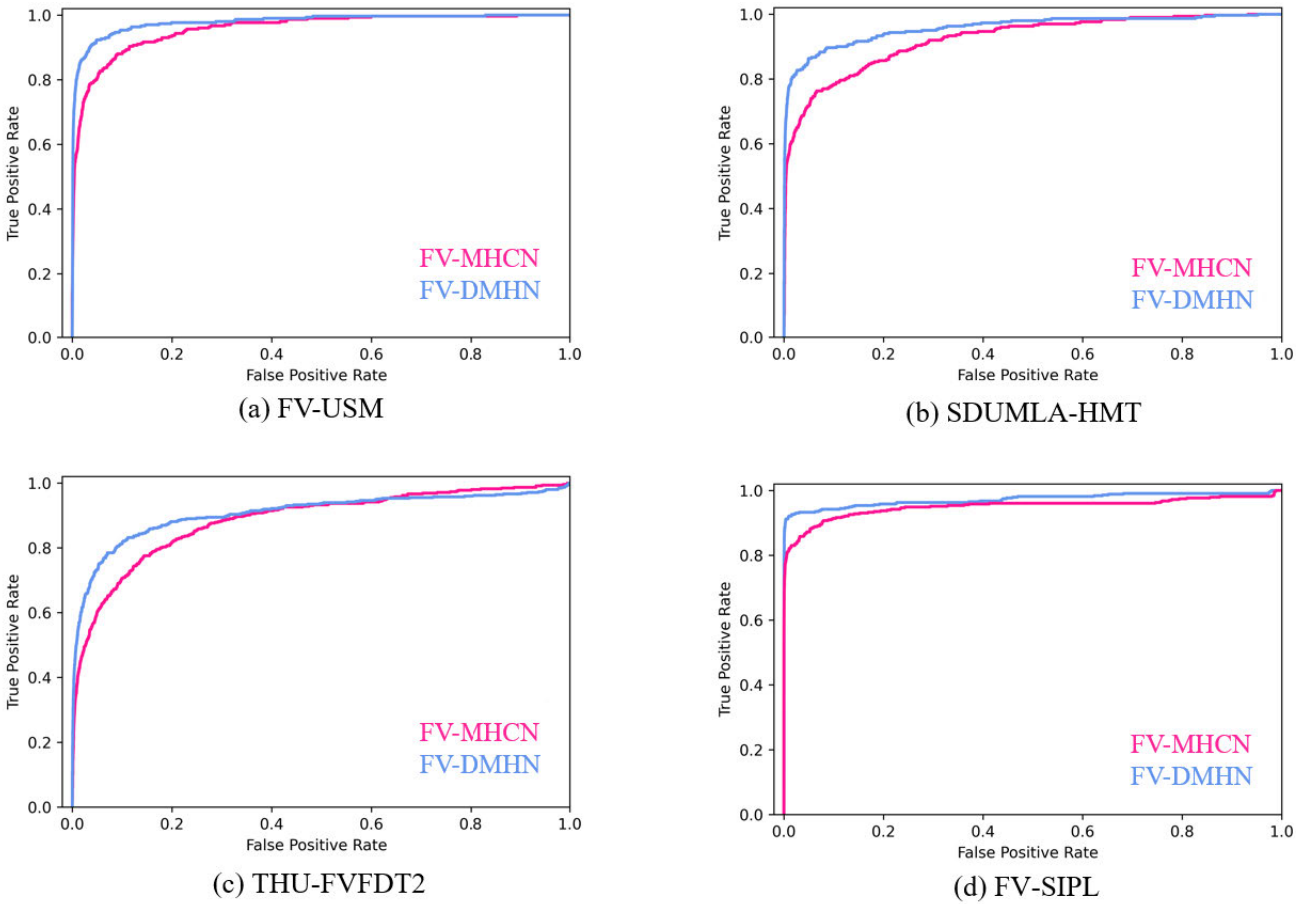


FIGURE 9. ROC comparison on four datasets.

To better discuss the network complexity and running speed of different methods, the number of parameters, FLOPs(floating point of operations), and single image recognition time are used to further illustrate them, and the experimental results are shown in Table 5. Due to the multi-head self-attention mechanism used in this paper, there are more multiplications and multiple attention mechanisms that need to be cascaded, resulting in higher results for FV-DMHN than the lightweight model MobileNetV2, but collectively this paper's algorithm still has a large advantage. From the above table, it can be seen that this paper's method can achieve higher recognition accuracy and faster recognition speed with lower number of parameters and FLOPs, and it also proves that the combination of self-attention mechanism and multi-head convolution can integrate the characteristics of convolution and self-attention in the process of finger vein recognition, and achieve more excellent recognition results.

TABLE 5. Comparison of classical deep learning methods.

Method	Time(ms)	Params(M)	FLOPs(G)
VIT-B	11.31	103.3	16.78
Swim-T	7.22	28.28	4.36
Conformer-B	7.16	96.64	21.11
Next-ViT	3.54	31.75	5.77
EfficientNet	3.49	21.47	2.91
MobileNetV2	**2.27**	**3.50**	**0.32**
ResNet-101	7.61	44.56	7.88
FV-DMHN	2.63	10.61	0.92

D. ABLATION EXPERIMENT

To verify that MHSA can effectively improve the recognition performance of the base network, ablation experiments were conducted on four datasets respectively. The comparison results of FV-DMHN and FV-MHCN (without the MHSA module) in terms of recognition accuracy are shown in Table 6 and Table 7, from which it can be seen that the addition of the MHSA module improves the accuracy by 0.50%∼1.59%, and at the same time, the EER value is reduced by 0.071%∼0.177%. The comparison of ROC curves is shown in Fig.9, from which it can be seen that the experimental results of FV-DMNH are significantly better than those of FV-MHCN. The above experimental results

TABLE 6. Accuracy comparison on four datasets.

	FV-MHCN	**FV-DMHN**
FV-USM	99.00%	**99.81%**
SDUMLA-HMT	98.33%	**99.67%**
THU-FVFDT2	98.10%	**99.69%**
FV-SIPL	99.33%	**99.83%**

TABLE 7. EER comparison on four datasets.

	FV-MHCN	**FV-DMHN**
FV-USM	0.158%	**0.087%**
SDUMLA-HMT	0.784%	**0.378%**
THU-FVFDT2	1.196%	**0.643%**
FV-SIPL	0.291%	**0.124%**

demonstrate that the addition of MHSA can effectively improve the recognition performance of the network.

V. CONCLUSION

Considering that the existing finger vein recognition network global information extraction is insufficient, the network model is too complex, and with other problems, this paper proposes a dual-multiple-head network method for finger vein image recognition. Firstly, image preprocessing is applied to convert the collected near-infrared finger vein images into a unified standard format. Then, the inverse residual module is used to effectively learn the abstract and low-resolution feature maps in the image, followed by multi-head convolution to extract the local information of the feature maps at different locations for effective local representation learning, and then the global information contained in the feature maps is processed and summarized using the multi-head auto-attention mechanism, and finally, the recognition classification is carried out through the output layer to obtain the recognition results. Compared with the existing novel finger vein recognition methods and some classical and efficient recognition network models, the method in this study reflects the obvious advantages of FV-DMHN to extract more expressive image features with higher accuracy and lower EER values. Meanwhile, the experiments on multiple datasets also verified that a heavyweight model such as Vision transformer could not achieve the desired recognition effect on a dataset with a small number of samples like finger veins. Therefore, the next work is to collect larger datasets to evaluate the performance of the model further. To make the model more practical, in future work, we will consider designing a more lightweight network with faster recognition speed while maintaining high accuracy.

REFERENCES

[1] D. Noh, W. Lee, B. Son, and J. Kim, "Empirical study on touchless fingerprint recognition using a phone camera," *J. Electron. Imag.*, vol. 27, no. 3, 2018, Art. no. 033038.

[2] X. Liu, Y. Bai, Y. Luo, Z. Yang, and Y. Liu, "Iris recognition in visible spectrum based on multi-layer analogous convolution and collaborative representation," *Pattern Recognit. Lett.*, vol. 117, pp. 66–73, Jan. 2019.

[3] Z. Xie, Z. Guo, and C. Qian, "Palmprint gender classification by convolutional neural network," *IET Comput. Vis.*, vol. 12, no. 4, pp. 476–483, Jun. 2018.

[4] C. Dong, R. Wang, and Y. Hang, "Facial expression recognition based on improved VGG convolutional neural network," *J. Phys., Conf. Ser.*, vol. 2083, Jul. 2021, Art. no. 032030.

[5] J. Choi, K. J. Noh, S. W. Cho, S. H. Nam, M. Owais, and K. R. Park, "Modified conditional generative adversarial network-based optical blur restoration for finger-vein recognition," *IEEE Access*, vol. 8, pp. 16281–16301, 2020.

[6] Y. Xin, Z. Liu, H. Zhang, and H. Zhang, "Finger vein verification system based on sparse representation," *Appl. Opt.*, vol. 51, no. 25, pp. 6252–6258, 2012.

[7] N. Miura, A. Nagasaka, and T. Miyatake, "Extraction of finger-vein patterns using maximum curvature points in image profiles," *IEICE Trans. Inf. Syst.*, vol. 90, no. 8, pp. 1185–1194, Aug. 2007.

[8] F. Liu, G. Yang, Y. Yin, and S. Wang, "Singular value decomposition based minutiae matching method for finger vein recognition," *Neurocomputing*, vol. 145, pp. 75–89, Dec. 2014.

[9] K. R. Park, "Finger vein recognition by combining global and local features based on SVM," *Comput. Informat.*, vol. 30, no. 2, pp. 295–309, 2011.

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, May 2017.

[11] H. Qin, M. A. El Yacoubi, J. Lin, and B. Liu, "An iterative deep neural network for hand-vein verification," *IEEE Access*, vol. 7, pp. 34823–34837, 2019.

[12] S. Yang, H. Qin, X. Liu, and J. Wang, "Finger-vein pattern restoration with generative adversarial network," *IEEE Access*, vol. 8, pp. 141080–141089, 2020.

[13] R. Das, E. Piciucco, E. Maiorana, and P. Campisi, "Convolutional neural network for finger-vein-based biometric identification," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 2, pp. 360–373, Feb. 2019.

[14] I. Boucherit, M. O. Zmirli, H. Hentabli, and B. A. Rosdi, "Finger vein identification using deeply-fused convolutional neural network," *J. King Saud Univ. Comput. Inf. Sci.*, vol. 34, no. 3, pp. 646–656, Mar. 2022.

[15] H. Hu, W. Kang, Y. Lu, Y. Fang, H. Liu, J. Zhao, and F. Deng, "FV-Net: Learning a finger-vein feature representation based on a CNN," in *Proc. 24th Int. Conf. Pattern Recognit. (ICPR)*, Aug. 2018, pp. 3489–3494.

[16] H. Huang, S. Liu, H. Zheng, L. Ni, Y. Zhang, and W. Li, "DeepVein: Novel finger vein verification methods based on deep convolutional neural networks," in *Proc. IEEE Int. Conf. Identity, Secur. Behav. Anal. (ISBA)*, Feb. 2017, pp. 1–8.

[17] W. Yang, C. Hui, Z. Chen, J.-H. Xue, and Q. Liao, "FV-GAN: Finger vein representation using generative adversarial networks," *IEEE Trans. Inf. Forensics Security*, vol. 14, no. 9, pp. 2512–2524, Sep. 2019.

[18] H. O. Shahreza and S. Marcel, "Deep auto-encoding and biohashing for secure finger vein recognition," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 2585–2589.

[19] H. Qin and M. A. El-Yacoubi, "Finger-vein quality assessment by representation learning from binary images," in *Proc. 22nd Int. Conf. Neural Inf. Process.*, Istanbul, Turkey. Cham, Switzerland: Springer, 2015, pp. 421–431.

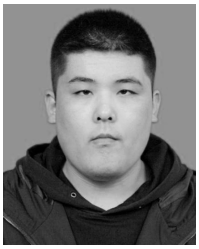
[20] S. A. Radzi, M. K. Hani, and R. Bakhteri, "Finger-vein biometric identification using convolutional neural network," *Turkish J. Electr. Eng. Comput. Sci.*, vol. 24, no. 3, pp. 1863–1878, 2016.

[21] C. Xie and A. Kumar, "Finger vein identification using convolutional neural network and supervised discrete hashing," *Pattern Recognit. Lett.*, vol. 119, pp. 148–156, Mar. 2019.

[22] D. Zhao, H. Ma, Z. Yang, J. Li, and W. Tian, "Finger vein recognition based on lightweight CNN combining center loss and dynamic regularization," *Infr. Phys. Technol.*, vol. 105, Jun. 2020, Art. no. 103221.

[23] Z. Lu, R. Wu, and J. Zhang, "Finger-vein feature extraction method based on vision transformer," *J. Electron. Imag.*, vol. 31, no. 4, 2022, Art. no. 043010.

[24] P. R. Kumar and K. L. Sailaja, "Watermarking algorithm using Sobel edge detection," *Int. J. Adv. Netw. Appl.*, vol. 2, no. 5, pp. 861–867, 2011.
[25] J. Chaki and N. Dey, *A Beginner's Guide to Image Preprocessing Techniques*. Boca Raton, FL, USA: CRC Press, 2018.
[26] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.
[27] A. Srinivas, T.-Y. Lin, N. Parmar, J. Shlens, P. Abbeel, and A. Vaswani, "Bottleneck transformers for visual recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 16514–16524.
[28] I. Bello, B. Zoph, Q. Le, A. Vaswani, and J. Shlens, "Attention augmented convolutional networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3285–3294.
[29] K. Shaheed, A. Mao, I. Qureshi, M. Kumar, S. Hussain, I. Ullah, and X. Zhang, "DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition," *Expert Syst. Appl.*, vol. 191, Oct. 2022, Art. no. 116288.
[30] H. Ren, L. Sun, J. Guo, C. Han, and F. Wu, "Finger vein recognition system with template protection based on convolutional neural network," *Knowl.-Based Syst.*, vol. 227, 2021, Art. no. 107159.
[31] K. Wang, G. Chen, and H. Chu, "Finger vein recognition based on multi-receptive field bilinear convolutional neural network," *IEEE Signal Process. Lett.*, vol. 28, pp. 1590–1594, 2021.
[32] X. Li and B.-B. Zhang, "FV-ViT: Vision transformer for finger vein recognition," *IEEE Access*, vol. 11, pp. 75451–75461, 2023.
[33] Y. Zhong, J. Li, T. Chai, S. Prasad, and Z. Zhang, "Different dimension issues in deep feature space for finger-vein recognition," in *Proc. 15th Chin. Conf. Biometric Recognit.*, Shanghai, China. Cham, Switzerland: Springer, 2021, pp. 295–303.
[34] A. Krishnan and T. Thomas, "Finger vein recognition based on anatomical features of vein patterns," *IEEE Access*, vol. 11, pp. 39373–39384, 2023.
[35] S. Tang, S. Zhou, W. Kang, Q. Wu, and F. Deng, "Finger vein verification using a Siamese CNN," *IET Biometrics*, vol. 8, no. 5, pp. 306–315, Sep. 2019.
[36] W. Yang, W. Luo, W. Kang, Z. Huang, and Q. Wu, "FVRAS-Net: An embedded finger-vein recognition and AntiSpoofing system using a unified CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8690–8701, Nov. 2020.
[37] M. Tan and Q. V. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," 2019, *arXiv:1905.11946*.
[38] J. Li, X. Xia, W. Li, H. Li, X. Wang, X. Xiao, R. Wang, M. Zheng, and X. Pan, "Next-ViT: Next generation vision transformer for efficient deployment in realistic industrial scenarios," 2022, *arXiv:2207.05501*.
[39] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 9992–10002.
[40] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[41] Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao, and Q. Ye, "Conformer: Local features coupling global representations for visual recognition," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 357–366.
[42] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.



ZHIJIN AN received the B.E. degree in electronic and information engineering from Liaoning Technical University, Huludao, China, in 2022, where he is currently pursuing the M.S. degree with the School of Electronic and Information Engineering.

His current research interests include image processing, pattern recognition, and artificial intelligence.

XIAOKUI REN received the B.E. degree in radio from the Department of Physics, Liaoning University, Shenyang, China, in 1989. He is currently an Associate Professor with the School of Electronic and Information Engineering, Liaoning Technical University. His research interests include digital image processing and machine learning.

ZHIYONG TAO received the Ph.D. degree from Liaoning Technical University. He was a Visiting Scholar with Clausthal University of Technology, Germany, in 2006. Since 2008, he has been an Associate Professor and the Master Tutor with Liaoning Technical University. His research interests include the Internet of Things, machine learning, and biometric recognition, with rich project and engineering experience.