IMPROVED LIGHTWEIGHT MULTISCALE FINGER VEIN RECOGNITION FOR VISION TRANSFORMER

Zhiyong TAO[1], Yajing GAO[1*], Sen LIN[2]

Finger vein recognition methods suffer from ignoring local information, complexity and slow recognition speed when applying Transformer architecture. In this paper, an improved visual transformer multi-scale finger vein recognition method is proposed. Specifically, the network backbone adopts the improved visual transformer architecture and grouped convolutional structure. The improved vision transformer architecture can extend the global features of an image while simultaneously reducing the computational cost-effectively. Group convolution realizes low-cost multi-scale image feature extraction. The experiment showed that the method proposed in this paper has a recognition accuracy of 99.86%, which is more suitable for industrial deployment than other state-of-the-art works.

Keywords: Convolutional neural network, Finger vein recognition, Near-infrared image, Light-weight networking, Feature extraction

1. Introduction

Biometrics is one of the most common identification methods, including: human face, fingerprints, palm print, iris, finger vein, etc. Because each person's vein texture is hidden in the body, it is not easy to steal and has uniqueness, so biometrics of finger veins has great advantages in vivo recognition. As far as current research is concerned, light and angle of finger placement all affect the recognition performance of finger veins. It is very important to study more accurate and robust recognition algorithms. The common finger vein recognition process usually includes image acquisition, preprocessing, feature extraction, and comparison. The acquisition of finger vein images necessitates the use of a device that combines an image sensor and an infrared light source. The acquired finger vein image is preprocessed to facilitate the subsequent feature extraction process.

As deep learning technology advances, deep learning-based recognition techniques have shown greater advantages over traditional methods, which is due to the fact that deep learning-based methods can obtain deeper image features through Convolutional neural networks (CNN) and can show more stable recognition results. Therefore, some researchers proposed CNN based finger vein

[1] School of Electronic and Information Engineering, Liaoning Technical University, Huludao, Liaoning, 125105, China, e-mail: xyzmail@126.com
[2] School of Automation and Electrical Engineering, Shenyang Ligong University, Shenyang, Liaoning, 110159, China, e-mail: lin_sen6@126.com
[*] Corresponding author:Yajing Gao, e-mail: mailgyj@163.com

recognition methods, for example, Fang et al. [1] proposed a lightweight dual channel network to improve finger vein verification by extracting mini region of interest (ROI). The Vision Transformer (ViT) [2] method, proposed by researchers recently, has attracted a lot of attention in the field of deep learning. Compared with CNN, ViT focuses more on global features and has shown excellent performance in several domains. In addition, researchers have proposed some improved methods, such as Liu et al. [3] proposed Swin Transformer, which obtains global and local features by constructing hierarchical feature maps and sliding windows, with better experimental results but high model complexity; Peng et al. [4] proposed Parallel Network Architecture, which utilizes convolution and the mechanism of multiple self-attention[5] (MHSA) for parallel extracts local and global features, which improves the network performance but is ineffective for small datasets. Based on the advantages shown by Transformer, researchers started to apply it to finger vein recognition. Huang [6] proposed the Finger Vein Transformer (FVT) model for recognition, which realizes multi-scale feature extraction by reducing the number of Token layer by layer but increases the complexity and computation.

From the above analysis, it can be seen that the existing approaches have achieved better results in terms of recognition performance, but there are still some shortcomings in terms of recognition time and model complexity. In this paper, we conduct an in-depth study on the problems of low accuracy and high model complexity arising from applying Transformer architecture in finger vein recognition. Lightweight multi-scale finger vein recognition with improved Vision Transformer is designed to extract local and global features in finger vein images by improved E-Transformer Block and grouped Group-Conv Block together, avoiding the problem of low accuracy caused by insufficient feature extraction. MHSA is utilized in the E-Transformer Block to maximize the acquisition of global features, while the recognition accuracy is improved by improving MLP. The group convolution in the Group-Conv Block effectively reduces the computational cost in the feature extraction process and realises the lightweight and multi-scale extraction of image features. Finally, comprehensive experiments on self-constructed datasets and three public datasets show that our proposed methods achieve better recognition results with lower model parameters and computational complexity, as well as shorter recognition time and lower equal error rate (EER).

2. Finger Vein Image Recognition Network

How to take into account the simplicity and light weight of the CNN model while utilizing the Transformer architecture for finger vein recognition is a major challenge in current research. Therefore, this paper designs a multi-scale finger vein recognition method with an improved Vision transformer, as shown in Fig. 1. The network is composed of two components: the E-Transformer Block, which

enhances the ViT, and the Group-Convolution Group-Conv Block. One of the E-Transformer Block is responsible for obtaining the global features in the image. By making the multi-head self-attention mechanism more efficient, the computational cost can be decreased, and the complexity can be minimized, resulting in a lightweight design. Another Group-Conv Block is responsible for obtaining some subtle features in the image, and group convolution can realize feature amplification and multi-scale acquisition of effective information in the image. After feature extraction by the above two blocks, the learning process of the classifier is supervised using the cross-entropy loss function to output more accurate recognition results.

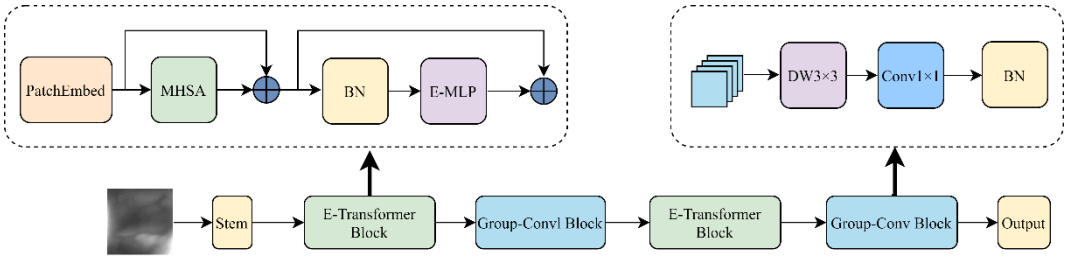


Fig. 1 Network structure diagram

2.1 Group-Conv Block

To further improve the network's effectiveness in recognizing finger veins, Group-Conv Block utilizes group convolution to achieve feature amplification. The details are shown in Fig. 2. Group-Conv Block contains one $\mathrm{DW3\times3}$, one $\mathrm{Conv1\times1}$ and one batch normalization (BN) layer. A DW convolution operation for the input features allows the number of channels that would otherwise be C to be increased to C'. The number of parameters required for a regular convolution operation on a given input feature map is:

$$F_{conv} = H \times W \times C \times C' \tag{1}$$

Nevertheless, when employing DW convolution, the operation necessitates a certain number of parameters:

$$F_{dw} = H \times W \times (C / \mathrm{group}) \times (C' / \mathrm{group}) \times \mathrm{group} \tag{2}$$

The above formula shows that grouped convolution can obtain more finger vein features with a smaller number of parameters, the network parameters and computational workload are decreased while still guaranteeing recognition accuracy.

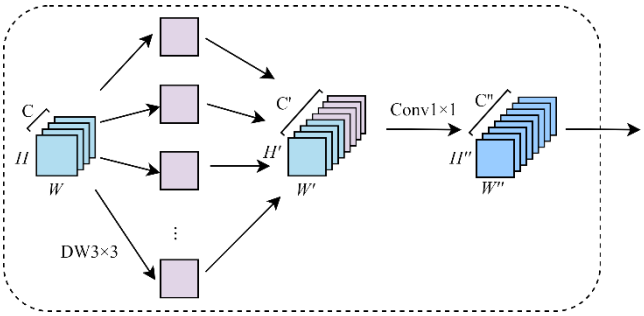


Fig. 2 Grouped convolutional structure

2.2 E-MLP Block

In E-Transformer, this paper focuses on improving the MLP structure in the vision transformer. To show the superiority of the improved MLP in this paper, we compare it with the common converter structure papers such as ResT V2 [7], ViT, Next-ViT [8], and EfficientFormer [9], as shown in Fig. 2. Next-ViT uses ReLU as the activation function, but GELU performs better and converges faster than ReLU. In addition, we find that the BN layer prevents overfitting and speeds up training. For these reasons, in this paper, we use GELU as the activation function and add BN layer and Dropout layer to improve the performance. We demonstrated the effectiveness of E-MLP in the ablation experiment section.

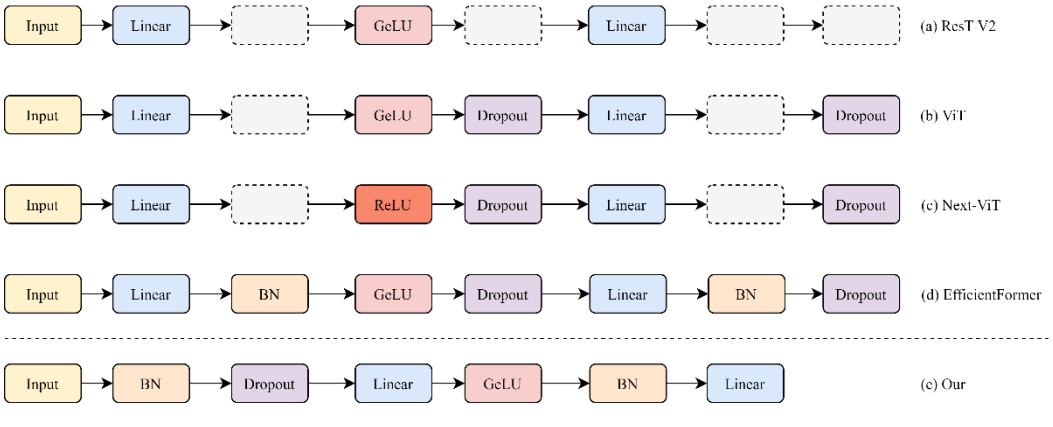


Fig. 3 Comparison of MLP in different methods

2.3 MHSA Block

MHSA can adaptively learn the relationship between different regions in an image to extract more comprehensive feature information. Therefore, this paper uses MHSA from ViT structure for global feature extraction. In ViT, the input image is first divided into multiple subgraphs as input vectors. Then, these vectors are encoded using the multi-head attention mechanism, the correlation between

them is calculated, and the attention weights are obtained by the Softmax function. Finally, these vectors are weighted and summed with corresponding attention weights to obtain the final feature representation. We introduce positional coding to obtain the positional information in the image features. To reduce the computational cost and enable more efficient and lightweight deployment. As shown in the following figure, the low-frequency signal is first captured by MHSA with the following equation:

$$\mathrm{MHSA}(x) = \mathrm{Concat}(SA(x_1), SA(x_2), ..., SA(x_h))W^0 \tag{3}$$

Where denotes the division of input features into multiple heads in the channel dimension and is the number of divided heads. In this paper, 8 is taken as the number of heads for the attention mechanism. Is the attention mechanism computational formula, and the formula is as follows:

$$SA(x) = \mathrm{Attention}(Q \cdot W^Q, K \cdot W^K, V \cdot W^V) \tag{4}$$

Where Attention denotes the standard attention, Q , K , V denotes query vector, key vector, and value vector, respectively, W^Q , W^K , W^V is the linear layer used for context encoding.

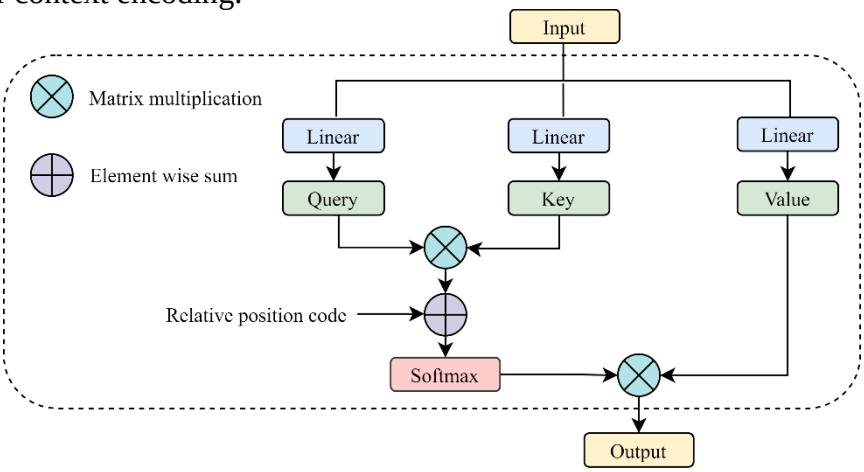
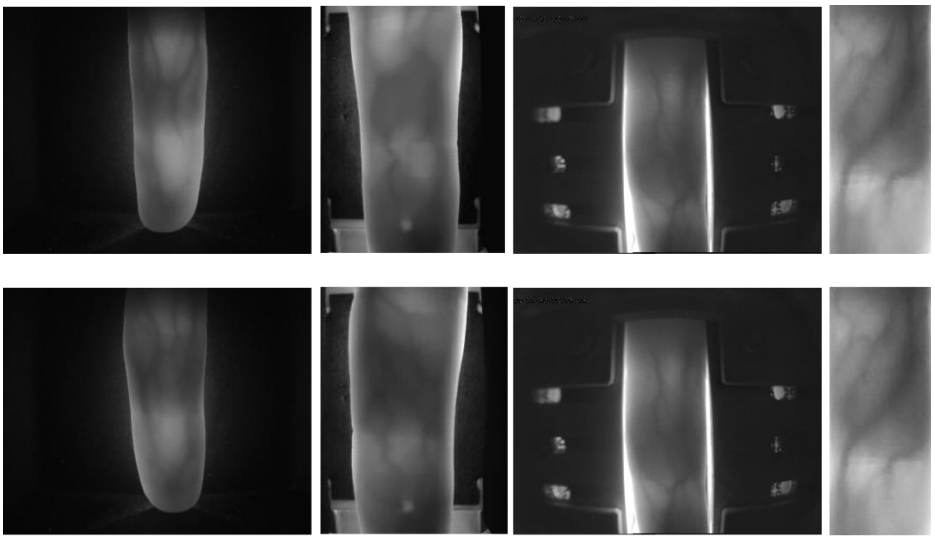


Fig. 4 Structure of MHSA

3 Experimental results and analyses

3.1 Data set profiles

We conducted experiments on a total of four datasets, namely, FV-USM [10], SDUMLA-HMT [11], THU-FVFDT2 [12], and the self-constructed dataset FV-SIPL, with the exception of the THU-FVFDT2 dataset where the training and test sets are equally distributed in a 2:1 ratio. The data information is shown in the following table.Figure 5 illustrates the sample finger vein maps in the four datasets.



(a) FV-USM (b) SDUMLA-HMT (c) THU-FVFDT2 (d) FV-SIPL

Fig. 5 Sample finger vein plots from the four datasets

Table 1

Image categories and totals in the four datasets

Dataset	Total number of categories	Total image count	Total training sets	Total test sets
FV-USM	492	5904	3936	1968
SDUMLA-HMT	636	3816	2544	1272
THU-FVFDT2	610	1220	610	610
FV-SIPL	108	1296	864	432

3.1.1 FV-USM dataset

Universiti Teknologi Malaysia supplied this dataset, which includes finger vein images taken by 123 volunteers, each volunteer's four fingers were used to capture 12 images. Therefore, the whole dataset covers a total of 492 finger categories and 5904 images. The size of each of these images is 640×480pixels.

3.1.2 SDUMLA-HMT dataset

This dataset is provided by Shandong University, which contains the finger vein images of 106 volunteers, and 6 images are collected for each index, middle and ring finger of each volunteer's hands. The whole dataset covers 636 finger categories and 3816 images, where each image size is 320×240pixels.

3.1.3 THU-FVFDT2 dataset

The dataset was provided by Tsinghua University and contained finger vein images of 610 volunteers. Finger vein images were collected twice for each volunteer, with a total of 1220 images, each with a size of 200×100pixels.

3.1.4 FV-SIPL dataset

This dataset was made by the Signal and Information Processing Laboratory of Liaoning University of Engineering and Technology by using infrared finger vein acquisition sensors to collect finger vein images from 27 volunteers. Among them, 12 images were collected for each of the four fingers of each volunteer, and the whole dataset covered 108 finger categories and 1296 images. The size of each image is 176×415pixels.

3.2 Experimental environment and parameter settings

The experiments were conducted under the Linux operating system using the PyTorch 1.7 framework. The graphics card used for training and testing was GeForce RTX 3090 GPU. The learning rate is set to 0.001, the batch size is set to 16, and Stochastic Gradient Descent (SGD) is chosen as the optimizer, with a momentum of 0.9. The the input size of the finger veins was pre-processed with operations and finally adjusted to 224×224 pixels uniformly, and the final experimental results are obtained through 300 iterations of training.

3.3 Evaluation Metrics

To evaluate the performance and advantages of the model, metrics such as Accuracy, EER, Average Processing Time of a Single Image (Time), Number of Parameters, and Floating Point Operations (FLOPs) are selected for evaluation. The accuracy rate is a frequently employed metric in recognition of finger veins. It can reflect the model's ability to correctly recognize different categories of samples in the entire data set. The formula for accuracy rate is shown in equation (5):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5}$$

TP represents the quantity of accurate positive sample predictions, while TN signifies the quantity of accurate negative sample predictions, The number of false positive sample predictions is denoted by FP, and the number of false negative sample predictions is denoted by FN. The EER value is typically employed in image recognition tasks to gauge the model's effectiveness, the determination is based on the False Acceptance Rate(FAR) and the False Rejection Rate(FRR). The formulas for FAR and FRR are shown below:

$$FAR = \frac{FP}{FP + TN} \tag{6}$$

$$FRR = \frac{FN}{TP + FN} \tag{7}$$

The predetermined threshold determines the quantity of samples for both false acceptance and false rejection when the threshold of the match is greater than the preset threshold, it is determined as wrong acceptance and vice versa as wrong rejection. When FAR and FRR are equivalent, the result is EER, which indicates

the effectiveness of the recognition technique, the recognition method's performance improves as the EER value decreases.

The magnitude and intricacy of the model have a considerable influence on the training and prediction results. The size of the model can be determined by the number of parameters, and the complexity of the model can be measured using FLOPs. The smaller the values of average processing time, number of parameters, and FLOPs for a single image, the lower the complexity of the model and the faster the recognition speed is proved.

3.4 Comparison Experiments

To validate the method in this paper, we compared it with a classical transformer network model: ViT-B, Swin-T, Conformer-B, Next-ViT and the lightweight CNN network model EfficientNetV2. The results of the recognition accuracy of the different methods on the datasets are shown in Table 2. The higher the accuracy rate, the better the recognition effect of the method is proved, and the results in the table show that the proposed method in this paper achieves the best recognition effect on all four datasets. The most favorable outcome is denoted by bold type, whereas underlining signifies the second most favorable outcome.

Table 2

Recognition accuracy of different methods on four data sets(unit: %)

Method	FV-USM	SDUMLA-HMT	THU-FVFDT2	FV-SIPL
ViT-B	84.00	83.00	75.34	93.02
Swin-T	98.33	97.33	90.13	99.52
Conformer-B	97.00	97.00	97.09	97.91
Next-ViT	98.56	<u>99.0</u>	<u>98.87</u>	<u>99.53</u>
EfficientNetV2[13]	<u>99.00</u>	98.00	98.78	97.20
MobileNetV2[14]	98.20	99.00	98.32	99.00
ResNet101[15]	98.33	98.34	98.21	99.00
Our	**99.69**	**99.86**	**99.33**	**99.73**

In addition to the comparison of accuracy, the average processing time, number of parameters and FLOPs of individual images of different methods are also compared, as shown in Table 3. In terms of average processing time for a single image, MobileNetV2 is 2.27ms, which is 1.16ms faster than the method proposed in this paper, due to the multi-attention mechanism used in this paper's method. Other than that, this paper's method outperforms other methods.

Table 3

Analyzing the evaluation index outcomes of various methodologies

Method	Time/ms	Parameters/M	FLOPs/G
ViT-B	11.30	103.03	16.88
Swin-T	7.21	28.27	4.37
Conformer-B	7.15	96.63	21.01
Next-ViT	3.52	31.76	5.79
EfficientNetV2	3.49	21.46	2.90

MobileNetV2	**2.27**	3.50	**0.33**
ResNet101	7.61	44.55	7.84
Our	3.43	**1.97**	0.35

This research utilizes four datasets to assess different recognition methods, including ViT-B, Swin-T, and Conformer-B. Fig. 6 displays the outcomes. On the SDUMLA-HMT dataset, the EER value of the method proposed in this paper is slightly higher than has excellent performance in finger vein recognition and can be used as an effective recognition method. This paper's proposed method is more accurate and robust than other methods, it has the potential to be used in a variety of practical areas.

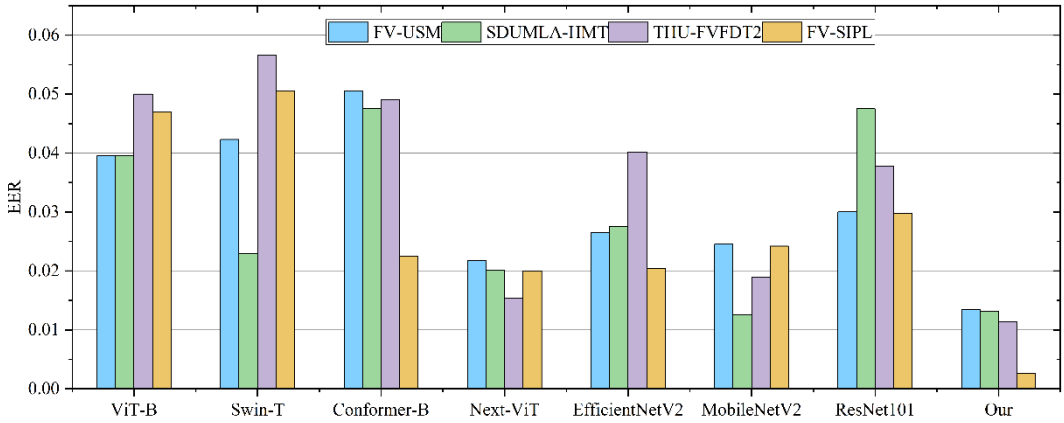


Fig. 6 Comparison of EER of different methods

The results of comparing the method proposed in this paper with novel finger vein models in recent years are shown in Table 4. Out of all the methods, our approach achieves the highest recognition accuracy on both public datasets, SDUMLA-HMT and THU-FVFDT2. Although the recognition accuracy on the FV-USM dataset is lower than that of the FVT method by 0.04%, it is higher than that of FVT by 1.96% and 8.67% on the SDUMLA-HMT and THU-FVFDT2 datasets, respectively. Consequently, this paper's proposed method yields the most favorable recognition outcomes based on the overall results.

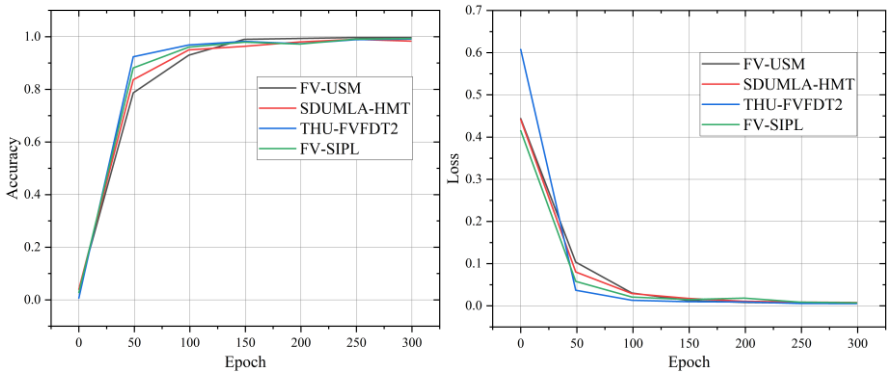
By comparing the novel finger vein recognition algorithms in recent times, this paper's suggested approach yields superior recognition accuracy, recognition time and complexity. We take the data at epoch 0, 50, 100, 150, 200, 250, 300 for image plotting, and the recognition accuracy versus test loss curves on the four datasets are shown below.

Table 4

The precision of various techniques in identifying public datasets(unit: %)

Method	FV-USM	SDUMLA-HMT	THU-FVFDT2
Merge CNN[16]	96.15	89.99	—
DS-CNN[17]	—	98.00	89.00

Semi-PFVN[18]	94.67	96.61	—
LFVRN_CE[19]	98.58	97.75	—
DGLFV[20]	—	99.25	—
CMrFD[21]	98.33	98.92	—
FVT	**99.73**	97.90	90.66
TFHFT-DPFNN[22]	—	98.00	—
CNNs[23]	97.95	—	—
Coding SchemeA[24]	99.59	95.91	—
FV-GAN	—	—	98.52
Triplet-Classifier GAN[25]	99.66	99.53	—
Our	99.69	**99.86**	**99.33**



(a) (b)

Fig. 7 (a) Accuracy curve;(b) Test loss curve

It is clear from the diagram that THU-FVFDT2 has the best recognition effect from 50 to 100 epochs, followed by FV-SIPL, SDUMLA-HMT, and FV-USM, and finally stabilizes after 150 epochs. Similarly, the loss curve shows that THU-FVFDT2 has the smallest loss from 50 to 100 epochs, followed by SDUMLA-HMT, FV-SIPL, and FV-USM, whose losses increase sequentially and eventually converge to zero.

3.5 Ablation Experiments

We conducted ablation experiments to better verify the conjecture. Under the premise that the rest of the conditions remain unchanged, the Group-Conv Block is combined with different models of the MLP structure, and the accuracy is tested on four datasets and the results are shown in Fig. 8.Based on the information depicted in the figure, it can be seen that the improved MLP structure in this paper has obvious advantages in the recognition effect compared with the classic Transformer architecture paper, thus verifying that the previous conjecture is correct.

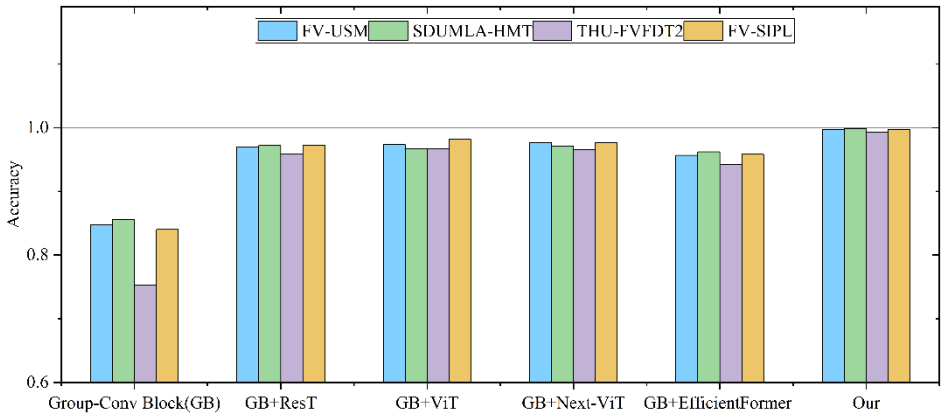


Fig. 8 Results of ablation experiments on four datasets with different methods

4 Conclusion

Aiming at the finger vein recognition process, which does not fully consider the global features of the image and is easy to overfit and other problems, this paper proposes a multi-scale finger vein recognition method with an improved Vision Transformer. The improved E-Transformer and group convolution are utilized to form the backbone network. In the network, the E-Transformer is responsible for extracting global features, where the improved MLP makes the feature extraction capability substantially enhanced. Secondly, the use of low-cost packet convolution allows for feature amplification and multi-scale information acquisition. The method is experimentally tested on multiple datasets, and good experimental results are obtained under several evaluation metrics. At present, the method has good performance and considerable potential in finger vein recognition, but there are still many aspects that need further improvement and refinement, which will be the focus of our future work.

Acknowledgement

This project was funded by the Applied Basic Research Project of Liaoning Provincial Department of Science and Technology (Approval No. 2022JH2/101300274), the Basic Research Project of Higher Education Institutions in Liaoning Province (LJKMZ20220679), and the Teaching Reform Project of Liaoning Provincial Department of Education (LNYJG2023117).

REFERENCES

[1] *Y. Fang, Q. Wu, W. Kang*, "A novel finger vein verification system based on two-stream convolutional network learning", Neurocomputing, **No. 290**, Sept. 2018, pp. 100-107.

[2] *A. DosoVITskiy, L. Beyer*, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale", ArXiv arXiv:2010.11929, 2020.

[3] *Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang and B. Guo,* "Swin transformer: Hierarchical vision transformer using shifted windows", IEEE/CVF international conference on computer vision, pp. 10012-

10022.
[4] *Z. Peng, W. Huang, S. Gu, L. Xie, Y. Wang, J. Jiao and Q. Ye,* "Local features coupling global representations for visual recognition", ArXiv arXiv:2105.03889, 2021.
[5] *A. Srinivas, T. Y. Lin, N. Parmar, J. Shlens, P. Abbeel and A. Vaswani,* "Bottleneck transformers for visual recognition", IEEE/CVF conference on computer vision and pattern recognition, pp. 16519-16529.
[6] *J. Huang, W. Luo, W. Yang, A. Zheng, F. Lian and W. Kang,* "FVT: finger vein transformer for authentication", IEEE Transactions on Instrumentation and Measurement, **No. 71**, Sept. 2022, pp. 1-13.
[7] *Q. Zhang, Y. B. Yang,* "Rest v2: simpler, faster and stronger", Advances in Neural Information Processing Systems, **No. 35**, Sept. 2022, pp. 36440-36452.
[8] *J. Li, X. Xin, L. W, H. Li, X. Wang, X. Xiao and X. Pan,* "Next-ViT: Next generation vsion transformer for efficient deployment in realistic industrial Scenarios", ArXiv arXiv: 2207.0550, 2022.
[9] *Y. Li, G. Yuan, Y. Wen, J. Hu, G. Evangelidis, S. Tulyakov and J. Ren,* "Efficientformer: Vision transformers at mobilenet speed", Advances in Neural Information Processing Systems, **No. 35**, Sept. 2022, pp. 12934-12949.
[10] *M. S. MAsaari, S. A. Suandi, B. A. Rosdi,* "Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics", Expert Systems with Applications, **vol. 41**, no. 7, Sept. 2014, pp. 3367-3382.
[11] *Y. Yin, L. Liu, X. Sun,* "SDUMLA-HMT: A multimodal biometric database",Biometric Recognition: 6th Chinese Conference, Sept. 2011, pp. 260-268.
[12] *W. Yang, C. Qin, Q. Liao,* "A database with ROI extraction for studying fusion of finger vein and finger dorsal texture", Biometric Recognition: 9th Chinese Conference, Sept. 2014, pp. 266-270.
[13] *M. Tan, Q. Le,* "Efficientnet: Rethinking model scaling for convolutional neural networks", International conference on machine learning, Sept. 2019, pp. 6105-6114.
[14] *M. Sandler, A. Howard, M. Zhu, A. Zhmoginov and L. C. Chen,* "Mobilenetv2: Inverted residuals and linear bottlenecks", IEEE conference on computer vision and pattern recognition, Sept. 2018, pp. 4510-4520.
[15] *K. He, X. Zhang, S. Ren, J. Sun,* "Deep residual learning for image recognition", IEEE conference on computer vision and pattern recognition, Sept. 2016, pp. 770-778.
[16] *D. Zhao, H. Ma, Z. Yang, J. Li and W. Tian,* "Finger vein recognition based on lightweight CNN combining center loss and dynamic regularization", Infrared Physics & Technology, **No. 105**, Sept. 2020, pp.103221.
[17] *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 Systems with Applications, **No. 191**, Sept. 2022, pp. 116288.
[18] *T. Chai, J. Li, S. Prasad, Q. Lu and Z. Zhang,* "Shape-driven lightweight CNN for finger-vein biometrics", Journal of Information Security and Applications, **No. 67**, Sept. 2022, pp. 103211.
[19] *Y. Zhong, J. Li, T. Chai, S. Prasad and Z. Zhang,* "Different Dimension Issues in Deep Feature Space for Finger-Vein Recognition", Chinese Conference on Biometric Recognition, Sept. 2021, pp. 295-303.
[20] *Z. Tao, H. Wang, Y. Hu, Y. Han, S. Lin and Y. Liu,* "DGLFV: Deep Generalized Label Algorithm for Finger-Vein Recognition", IEEE Access, **No. 99**, Sept. 2021, pp. 1-1.
[21] *J. Shen, N. Liu, C. Xu, H. Sun, Y. Xiao, D. Li and Y. Zhang,* "Finger vein recognition algorithm based on lightweight deep convolutional neural network", IEEE Transactions on Instrumentation and Measurement, **No. 71**, Sept. 2021, pp. 1-13.
[22] *D. Muthusamy, P. Rakkimuthu,* "Trilateral Filterative Hermitian feature transformed deep perceptive fuzzy neural network for finger vein verification", Expert Syst. Appl, **No. 196**, Sept. 2022, pp. 116678.
[23] *D. Zhao, H. Ma, Z. Yang, J. Li and W. Tian,* "Finger vein recognition based on lightweight CNN combining center loss and dynamic regularization", Infrared Physics & Technology, **vol. 105**, no. 8, Sept. 2020, pp. 103221.
[24] *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, **No. 227**, Sept. 2021, pp. 107159.
[25] *B. Hou, R. Yan,* "Triplet-classifier GAN for finger-vein verification", IEEE Transactions on Instrumentation and Measurement, **No.71**, Sept. 2022, pp. 1-12.