A Finger Vein Liveness Detection Method Based on Multi-Scale Spatio-Temporal Image and Light-ViT Model

Liukui Chen , Tengwen Guo , Li Li , Haiyang Jiang , Wenfu Luo , Zuojin Li *

Posted Date: 27 October 2023

doi: 10.20944/preprints202310.1764.v1

Keywords: finger vein living detection; MSTmap; light-ViT

Article

A Finger Vein Liveness Detection Method Based on Multi-Scale Spatio-Temporal Image and Light-ViT Model

Liukui Chen [1], Tengwen Guo [1], Li Li [2], Haiyang Jiang [1], Wenfu Luo [1] and Zuojin Li [1,*]

[1] Chongqing University of Science & Technology, Chongqing, 401331, China
[2] Wuhan Maritime Communication Research Institute, Wuhan, 430202，China
* Correspondence: cqustlzj@sina.cn

Abstract: Prosthetic attack is a problem that must be prevented in the current finger vein recognition application. To solve this problem, a finger vein living detection system was established in this article. The system first captures short-term static finger vein video by uniform near-infrared lighting, segments the veins by Gabor filters with current removing, calculates the multi-Scale spatial-Temporal maps(MSTmap) from the selected vein blocks, and trains the MSTmaps in the proposed Light-ViT network for the liveness detection. The MST maps are used to extract the coarse feature and Light-ViT is used to refine the liveness feature and predict the liveness result. Light-ViT, featuring an enhanced L-ViT backbone as its core, is constructed by interleaving multiple MN blocks and L-ViT blocks. This architecture effectively balances the learning of local image features, controls network parameter complexity, and substantially improves the accuracy of liveness detection. The accuracy of the Light-ViT network is verified to be 99.63% on the self-made living/prosthetic finger vein video dataset. This proposed system can also be directly applied to the finger vein recognition terminal after the model lightweighting.

Keywords: finger vein living detection; MSTmap; Light-ViT

1. Introduction

Biometric-based identification technology [1,2] has gained widespread adoption; however, information security is of paramount importance, with a particularly notable challenge being prosthetic attacks. Instances of prosthetic attacks targeting human faces and fingerprints have been well-documented [3]. Despite the inherent vitality of finger vein recognition, the risk of prosthetic attacks persists when the original vein line information is compromised. Although there have been no reported cases of prosthetic attacks resulting from the exposure of the original vein pattern information, foreign researchers have conducted simulations of prosthetic vein attacks with a notably high success rate, capable of deceiving existing hand vein recognition systems [4]. The detection of vitality in a biometric recognition system has emerged as a pivotal aspect for its secure deployment, and the assessment of vitality in finger vein recognition has consequently become a fundamental prerequisite and a focal point of research [5].

Compared to other body surface features, finger veins are an in vivo characteristic that does not leave marks during daily activities. While theft of the original vein pattern information is more challenging, there remains a risk of leakage of this information. Firstly, the existing company's hand vein recognition technology has demonstrated the collection and extraction of hand vein pattern information under visible light conditions [6]. Secondly, there are risks associated with the unauthorized collection of hand vein patterns using self-made equipment. To address these risks, further research of in vivo detection technology is necessary for the application of this technology [7]. In this paper, we propose a short-frame video of the finger vein to detect the living body of the vein recognition system. The vein video provides more spatio-temporal information compared to the vein image and can offer more robust living body features for detection.

 doi:10.20944/preprints202310.1764.v1

The key to liveness detection technology lies in the extraction of liveness features. Traditional image processing and deep learning have methods and models in vein liveness detection. In particular, neural networks have better performance than artificially designed features in static image texture features and dynamic video spatio-temporal feature extraction.[8,9], deep learning methods have become a hot research field of biometric liveness detection. Inspired by rPPG [10] and multi-scale spatio-temporal graph [11], combined with ViT architecture [12], this paper proposes a Light-VIT network for finger vein short video liveness detection.

2. Related Research

In recent years, attacks against biometric identification systems have primarily focused on prosthesis counterfeiting attacks. This attack, known as a presentation attack, involves using printed pictures, prostheses, videos, and other means to deceive the biometric identification system and authenticate an unauthorized person as a real user. Attackers use tools such as pictures, prostheses, and videos as presentation attack tools [13]. Living detection technology of general biometrics utilises individuals' physiological characteristics. For example, living fingerprint detection can be based on finger temperature, sweating, electrical conductivity and other information. Living face detection can be based on head movement, breathing, red-eye effect, and other information. Live iris detection can be based on the characteristics of iris vibration, the motion information of eyelashes and eyelids, and the contraction and expansion response characteristics of the pupil to the intensity of the visible light source. The corresponding liveness detection procedure can be divided into traditional liveness feature extraction and deep learning liveness feature extraction methods for single static images and multi-frame videos. The primary attack method for finger vein recognition technology is print attack. The attacker can use a stolen finger vein image to print it onto materials such as paper or film and attach it to a real finger to complete the biometric forgery attack.

In 2011, J. Maatta et al. [14] proposed a texture-based analysis method, which uses color texture as a PAD clue. By extracting complementary low-level features from different color spaces to describe the joint texture information of brightness and chrominance channels, combined with discrete Fourier analysis, the characteristics of energy distribution in the spectrum of prosthesis images are used as PAD clues, which has a high accuracy for printed prosthesis images.

In 2012, I. Chingovska et al. [15] proposed a PAD method for texture analysis of photo attacks based on weighted neighborhood difference quantization local binary pattern, which is used to combat the rendering attacks of printed images and video replays. It has achieved good results in three face anti-fraud databases.

In 2015, the GRIP-PRIAMUS team [16] used the characteristics that the printer always prints in a certain direction to find a larger energy relative to the vertical direction from the frequency domain transformation, so as to detect whether it is a vein living body. In the first finger vein deception attack prevention competition, a method based on the use of local descriptors was proposed. Firstly, the local phase quantization (LPQ) was used to encode the spatial domain information for the complete image, and then LBP was used to extract the texture features of the segmented vein region image. After fusing the features of the two, they are sent to the support vector machine (SVM) for classification.

In 2018, Fang et al. [17] proposed a novel finger vein recognition method. This method initially utilizes total variation regularization techniques to decompose the raw finger vein image into two components: structure and noise. Subsequently, block local binary pattern descriptors are employed to encode the structural and noise information in the decomposed image components. Finally, a cascaded support vector machine model is used for classification to achieve precise finger vein recognition. This approach not only reduces the impact of image noise and other interfering factors on recognition but also enhances recognition accuracy and robustness.

In 2019, J. Lee and their research team in South Korea [18] utilized the rPPG principle to effectively extract live features from finger veins via Fourier analysis and wavelet transform. These features were then classified using SVM to achieve vein live detection. This technique extracts frequency data by enhancing the vein video in multiple directions and widths. The resulting

 doi:10.20944/preprints202310.1764.v1

frequency signal is obtained through wavelet transformation and allows for non-contact, precise, and robust biometric attack detection.

The aforementioned frequency-domain-based approach, while advantageous in terms of ease of implementation, suffers from several drawbacks, including high computational complexity, sensitivity to noise, and poor robustness. Moreover, it relies on pre-defined feature parameters, limiting its generalization capabilities. Methods based on deep learning have been successfully applied across various domains, especially Convolutional Neural Networks (CNNs), which have witnessed significant advancements in the field of computer vision, notably in biometric recognition [19]. Deep learning and its inherent feature learning capabilities have paved a new path for the development of biometric presentation attack detection (PAD) systems.

In 2017,Assim O et al. [20] proposed a method called GA-CNN, which combines genetic algorithms and Convolutional Neural Networks (CNNs), resulting in significant improvements in both accuracy and sensitivity. The core idea behind this method is to optimize the hyperparameters of the neural network using genetic algorithms, thereby enhancing the model's performance and robustness. GA-CNN excels not only in vein liveness detection but also in its ability to handle vein images at various scales, making it more practical and versatile for real-world applications.

In 2020,Baweja et al. [21] introduced a novel approach for anomaly detection training. The absence of negative samples renders end-to-end learning impractical in one-class classification. Therefore, the authors proposed a "pseudo-negative" sample feature space, which aids the model in better understanding the decision boundaries between genuine and counterfeit samples. The pseudo-negative class is modeled using a Gaussian distribution. In contrast to other existing One-Class Classification (OCC) models, both the classifier and feature representation undergo end-to-end learning in this approach.

In 2020,Zeng et al. [22] devised an end-to-end model that combines Fully Convolutional Neural Networks (FCNN) with Conditional Random Fields (CRF) to extract vein pattern networks for finger vein verification. In 2021, Tao et al. [23] employed Mask Region-based Convolutional Neural Networks (Mask-RCNN) to obtain precise Region of Interest (ROI) images and developed a softmax-based decision algorithm for verification. Gionfrida [24] Proposed a model for recognizing gesture sequences from video using appearance and spatio-temporal parameters of consecutive frames and combine a convolutional network with a long and short-term memory unit, which is experimentally shown to outperform previous temporal segmentation models. In 2022, Zhou [25] proposed a virtual sample generation method called SIFT flow-based Virtual Sample Generation (SVSG). This method uses the Scale-Invariant Feature Transform flow (SIFT-flow) algorithm to obtain displacement matrices for each category with multiple registered samples. The process then involves extracting key displacements from each matrix, removing noisy and redundant displacements, and ultimately producing a final global variation matrix. The effectiveness of this method in improving the performance of single-sample finger vein recognition is demonstrated by experimental results.

Deep learning-based approaches leverage neural networks to learn data distributions without the need for manually designed feature extraction algorithms. Networks trained on extensive data exhibit high accuracy. However, this method has drawbacks, such as the difficulty in obtaining training data and the need to validate network generalization for specific problems. Additionally, some networks can be structurally complex, have a large number of parameters, lack real-time capabilities, and pose challenges in terms of portability to mobile platforms.

3. Our Method and System

In this paper, the system uses the finger vein static short-term video (Finger vein frame) as the data source for in vivo detection. There is blood flow in the superficial subcutaneous vein, and there will be slight expansion and contraction, and there will be slight changes in the absorption rate. When the near-infrared light penetrates the vein, there will be a slight gray change in the vein area during angiography imaging. This change can be detected by video image processing methods. There is no liquid flow in the ' blood vessel in the prosthesis, and there will be no gray change in the vein area. Such as references [26,27]. However, the slight gray-scale transformation will be submerged in the

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

speckle noise of near-infrared angiography. In order to solve this problem, this paper proposes an artificial shallow feature extraction + deep learning model for finger living detection. The flow chart of this method is as follows. To solve this problem, this paper proposes an artificial shallow feature extraction + deep learning model for finger living detection system. The flow chart of this system is shown in Figure 1. The first step is to obtain short-term static finger vein video, Next step is vein area segmentation, which belongs to the preprocessing. The third step is to select and cut out small blocks on the edge of the vein. The fourth step is to construct multi-scale space-time maps for these sorted small blocks. The fifth step is train the proposed the Light-ViT network with multi-scale space-time map. The last step is to output the result of liveness detection.

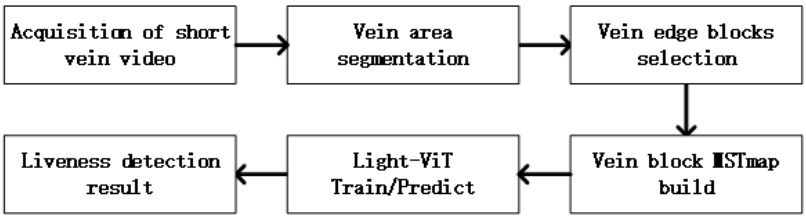

Figure 1. Flowchart of finger vein static short-term video live detection system.

Our self-made finger vein short-term video acquisition platform is shown in the Figure 2. The housing is manufactured using 3D printing technology and the upper light source bracket is designed to hold the light source group securely in place. This arrangement ensures stability and uniformity of the light source. In contrast, the main body of the case is made of black opaque material and the driver circuit board adjusts the light intensity to minimize the effects of ambient light and improve the accuracy of image acquisition. An opening in the top of the housing serves as a viewing window where a finger can be placed for image acquisition. This location facilitates the extraction of the region of interest (ROI) during post-processing. A groove designed on the inside of the bottom stabilizes the position of the camera and the IR filter.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

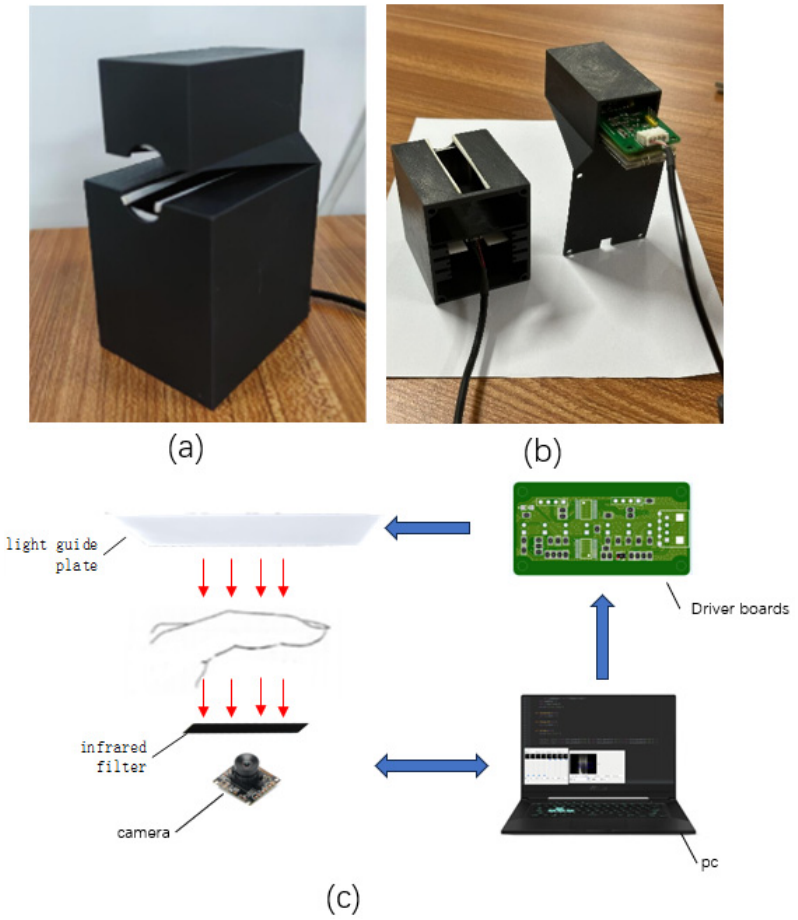

Figure 2. Hardware Equipment (a): Housings image; (b): Internal structure diagram; (c): Acquisition process image.

When the user places his/her finger between the light source of the device and the viewing window, the image grabber starts capturing the image of the finger vein. In this process, the controller constantly monitors and analyzes the quality of the captured image, and quickly adjusts the light intensity of the light source according to the changes in the brightness, contrast and other parameters of the image, so that the brightness and clarity of the captured image are always kept within the preset range. The application of intelligent control can be quickly adjusted according to the actual situation to ensure that the quality of the captured images are similar, thus improving the recognition accuracy and stability of the equipment.

3.1. Obtaining short-term static finger vein images

In this paper, a uniform illumination method is used to obtain a short-term static finger vein image. A light guide plate is used as the light source plate. The back and front panels of the light guide plate have a reflective layer. The near-infrared light emitted by the side near-infrared LED beads is reflected between the back plate and the front panel. After being reflected, it is emitted through the pre-distributed light guide hole of the front panel. The optical simulation software can be used to calculate the aperture and distribution of the light guide hole, and finally form a uniform illumination, as shown in Figure 3. The advantage of using the light guide plate is to avoid overexposure caused by glare. After multiple reflections, the light passes through the small light guide hole with less energy and does not cause glare. When the traditional LED lamp bead array light source collects the vein image, if the finger does not block the lamp bead, the direct light of the lamp bead will cause a large area of overexposure when it enters the camera. The image/video of uniform near-infrared illumination transmission finger venography has the same illumination conditions. The

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

noise distribution of each region of the finger vein image is consistent, and the later modeling is simpler and more convenient.

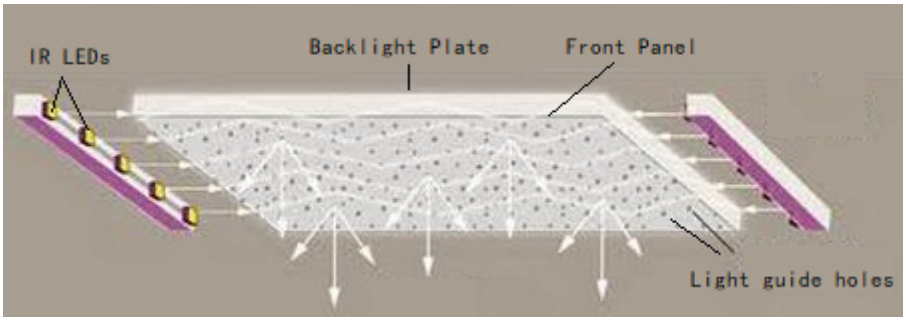

Figure 3. Low chart of finger vein static short-term video live detection system.

3.2. Preprocessing of video frames

In order to meet the real-time application of vein liveness detection and facilitate users, when collecting finger vein video, this paper uses the three-frame difference method to extract the frames in which the fingers remain static in the short-term video, so that it does not need to perform multi-frame pixel-level registration of the finger vein in the later period to avoid excessive calculation. The user 's finger only needs to stay in the acquisition device for 1.5 seconds. The camera uses a high-speed camera, 120 frames per second, and the pixel resolution is 640 * 480. The camera collects RGB three-channel images. Although there are IR filters in front of the lens, the corresponding near-infrared contrast images can still be collected. The noise of each channel is different, which can be used to increase the signal-to-noise ratio of small gray-scale changes in the vein area.

The multi-scale and multi-direction Gabor filtering method is used to segment the vein region. This paper presents a fast Gabor filter design method to remove the DC component, which is equivalent to the method in Reference [28]. The real part of the traditional Gabor filter is designed as follows: (1)

$$G(x,y) = \exp\left(-\frac{x_1^2+\gamma^2 y_1^2}{2\sigma^2}\right)\left(\cos\left(2\pi\frac{x_1}{\lambda}+\varphi\right) - exp\left(\frac{\nu^2}{2}+\varphi\right)\right) \quad (1)$$

$$x_1 = xcos\theta + ysin\theta \quad (2)$$

$$y_1 = -xsin\theta + ycos\theta \quad (3)$$

Here, γ represents the aspect ratio of the space, which determines the ellipticity of the shape of the Gaussian kernel function curve. When γ = 1, the shape is a positive circle; σ is the standard deviation of the Gaussian function, and its value cannot be directly set, but is related to the bandwidth. λ is the wavelength of sine function; φ is a sine function phase; θ is the rotation angle. Equations (2) and (3) illustrate that the Gabor function can be elongated in the (x, y) plane in any direction determined by θ.

In order to quickly remove the DC component of the filter template, this paper proposes to directly calculate the mean value of the Gabor filter as the DC component removed. The formula is (4) :

$$G(x,y) = \exp\left(-\frac{x_1^2+\gamma^2 y_1^2}{2\sigma^2}\right)\left(\cos\left(2\pi\frac{x_1}{\lambda}\varphi\right)\right) \quad (4)$$

$$G' = G - mean(G) \quad (5)$$

3.3. Selection of vein edge image block

The segmented binary vein image is subjected to morphological corrosion operation, and the vein edge region is obtained by subtracting the corroded image from the binary original image, as shown in Equation (6):

$$I_{edge} = I_{bw} - erode(I_{bw}, H) \quad (6)$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

In formula (6), I_{bw} is the binary image segmented from the previous vein, erode is the binary image erosion operation function, H is the morphological operator, here take 3 * 3 size.

3.4. Multi-scale spatio-temporal map calculation

In order to highlight the change of vein gray caused by blood flow, it is more noise-resistant to observe the mean change of vein block than the gray change of one pixel. In this paper, the image blocks on the edge of the vein are selected from the root of the finger to the tip of the finger along the main vein, and p blocks are selected. The gray mean value of each frame image block in the video is taken, and the gray change of the image block with time can be obtained. The gray value of the gray mean value of multiple image blocks with time changes to form a time-space map, as shown in Figure 4.

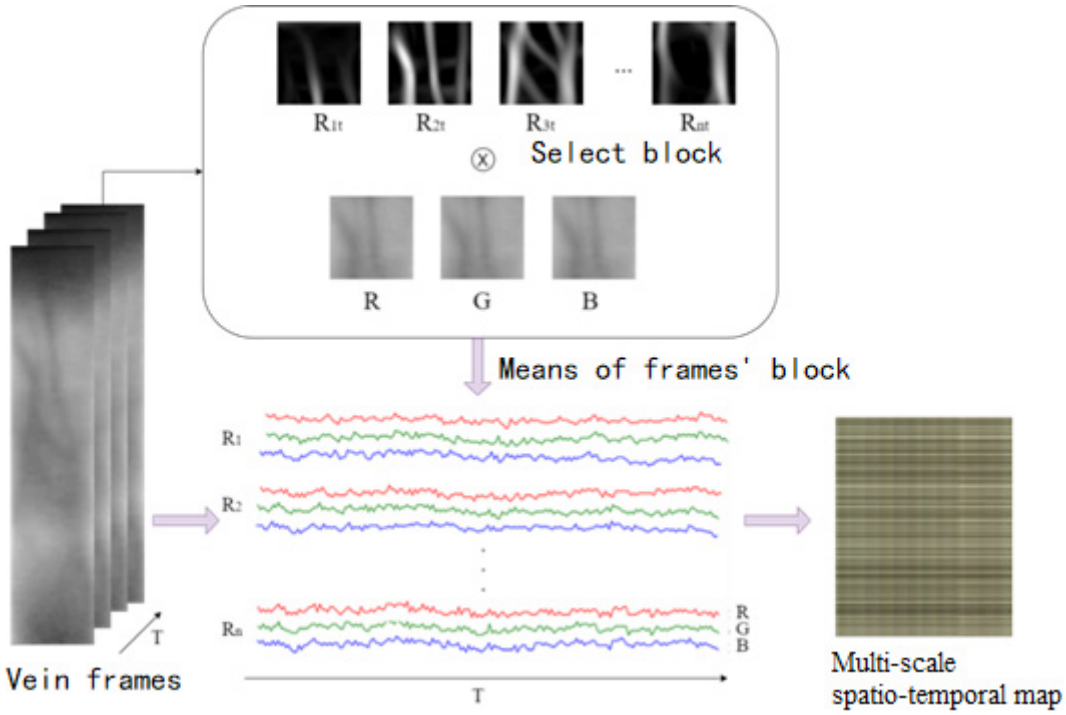

Figure 4. Construction of space-time graph.

3.5. Build Light-ViT network

For the purpose of anti-counterfeit detection in this context, the transformation of live finger vein video into spatiotemporal maps necessitates that the network possesses the capability to effectively manage long-range pixel relationships. Simultaneously, the conversion of vein edge position features requires dedicated attention to local characteristics. Convolutional Neural Networks (CNNs) have consistently exhibited exceptional proficiency in feature extraction [29]. However, when dealing with spatiotemporal maps of multiple scales that encompass both global and local features, CNNs still exhibit limitations in comprehensive global feature extraction. In contrast, the ViT network leverages its multi-head attention mechanism to excel in local feature handling, complemented by its capacity for long-range pixel-to-pixel feature extraction through positional encoding. The ViT network has been empirically proven to deliver superior performance, yet it grapples with challenges such as large parameter scales, training complexity, and suboptimal performance on smaller datasets. Furthermore, finger vein Presentation Attack Detection (PAD) serves as a pivotal technology in biometric identification, where precision and real-time responsiveness constitute fundamental requirements. Considering that biometric identification devices are typically compact in size and operate within the constraints of limited computational resources offered by computer chips, it becomes imperative to maintain a compact system

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

architecture. Consequently, we introduce the Light-ViT network, which not only adeptly captures both global and local data features but also facilitates seamless integration into our system, achieving high-precision finger vein counterfeit detection at a significantly reduced cost.

The network proposed in this chapter utilizes an improved L-ViT backbone as its core. This network is composed of multiple MN blocks and L-ViT blocks stacked alternately. Specifically, the MN block employs depthwise separable convolution operations with the aim of learning local image features while controlling the network's parameter count, enabling better adaptation to large-scale datasets. On the other hand, the L-ViT block adopts a Transformer structure to capture global image features and integrates them with locally extracted features obtained through convolution.

MN block is a convolution module within the network used for learning image biases and local features. Its structure is illustrated in the diagram. For input features, it initially undergoes a 1*1 convolution layer, effectively mapping to a higher dimension via pointwise convolution. Subsequently, it passes through a Batch Normalization (BN) layer and the SiLU activation function to obtain, where, and can be adjusted based on network requirements. Following this, it undergoes group convolution, followed by another BN layer and the SiLU activation function to obtain. Here, T represents the stride, and adjusting this parameter controls the output tensor's dimensions. After merging with the input features through a reverse residual structure, the dimensions of are mapped to using pointwise convolution (PW), followed by a BN layer to yield the output Y. The structure of MN block is shown in Figure 5.

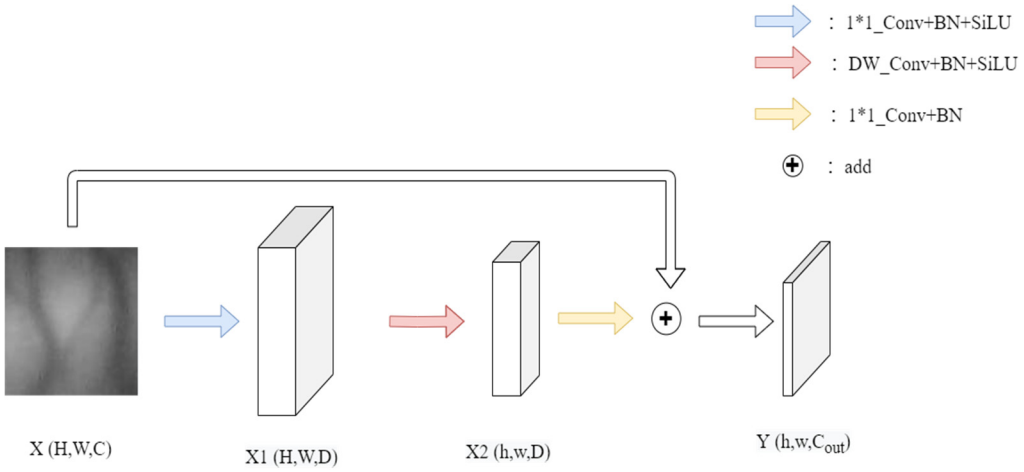

Figure 5. The structure of MN block.

This approach offers the advantage of significantly reducing the number of parameters and computational resources required for convolution, while maintaining the same convolutional kernel sensing field, as previously discussed [30]. Additionally, to imbue the CNN with the capacity to learn global features, we have introduced an L-VIT module, as depicted in Figure 6. Given an input $X(H,W,C)$, we commence by encoding the local spatial information through a 3 × 3 standard convolutional layer. Subsequently, we utilize a 1x1 convolutional layer to project the feature dimensions to a higher space, resulting in $X_1(H,W,D)$. To enable Light-ViT to acquire global representations with spatial inductive bias, we unfold X_1 into N patches and aggregate them to yield $X_2(P,N,D)$, where $P = w * h$, with w and h denoting the width and height of a patch, and $N = H * W/P$.

Subsequent spatial coding of X_2 produces $X_3(P,N,D)$, preserving both the ordering information among each patch and the spatial details of the pixels within each patch. Then, X_3 undergoes a 1x1 convolutional projection to return to a lower-dimensional space. It is then concatenated with X through a residual structure, followed by fusion of features using a standard convolution operation.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

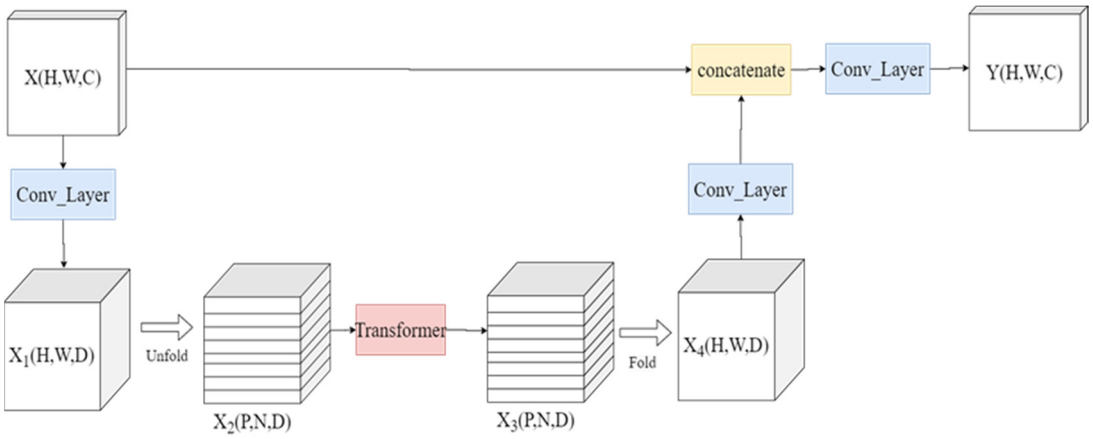

Figure 6. L-ViT block structure diagram.

The L-ViT block is primarily used to learn global features of the feature map. We employ the Unfold operation to process the input tensor, introducing positional information on the feature map while retaining the Encoder part that introduces attention mechanisms. However, we replace the Decoder part with convolution operations to adjust the size of the output feature map. The enhanced L-ViT further strengthens the network's understanding of images and provides more powerful performance through the comprehensive extraction and fusion of local and global features. The structure is depicted in the Figure 7.

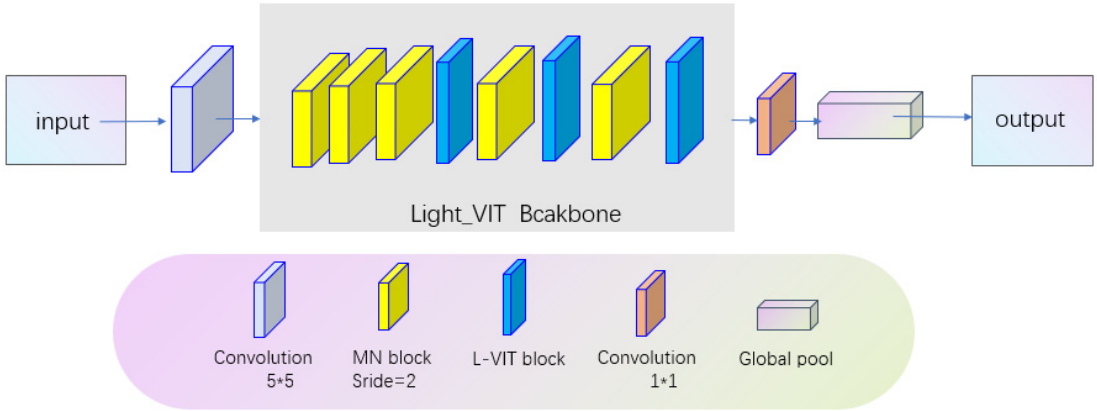

Figure 7. Light -ViT Network Structure.

The input image first undergoes a standard convolutional layer, primarily serving the purpose of generating shallow-level feature maps and adjusting the input dimensions. Subsequently, the feature maps are fed into the Light-ViT backbone, which consists of multiple MN blocks and Light-ViT blocks stacked alternately. This allows the network to learn new high-level feature maps that integrate both local and global features. These features can capture image details and global information more accurately, thereby providing richer and more comprehensive information for subsequent classification tasks. Then, a 1x1 convolutional layer maps the feature maps into a high-dimensional space. Finally, the network's output is obtained through global pooling and a linear layer.In global pooling, the network can perform statistical operations over the entire feature map to acquire more comprehensive and rich information. Finally, through the transformation of a linear layer, the network converts the feature map into a vector with corresponding class probability distribution.

Our proposed Light-ViT significantly reduces the demand for computational resources and greatly enhances the network's feature learning capabilities by introducing MN blocks and improving L-ViT blocks. When integrated into the system, this lightweight network exhibits high recognition accuracy.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

4. Experiment and Discussion

4.1. Introduction of experimental data.

There is no public finger vein prosthesis and in vivo short video on the Internet for the time being. In this paper, three kinds of finger vein prosthesis are made, and short videos of near-infrared finger veins are collected by using self-made equipment, and simulated forgery attack experiments are carried out. The data set consists of 400 short videos, including 200 live finger vein videos, from 100 different fingers. The prosthesis materials are divided into A4 paper, PVC plastic and laser film, and 100 prosthesis finger vein videos are collected each. These video samples cover different angles and different lighting conditions. In the experiment, the data set is randomly divided into training set and test set according to the ratio of 8:2, and ensure that there are no repeated samples. Therefore, the training set contains 320 videos, and the validation set contains 80 videos. The earlier methods [30] for prosthesis fabrication had suboptimal imaging quality within the data acquisition equipment introduced in this chapter. Common issues included a significant presence of noise, excessive venous contrast, and regions of excessive blurriness within the acquired images,The fabricated prosthesis and the acquired images are depicted in Figure 8.

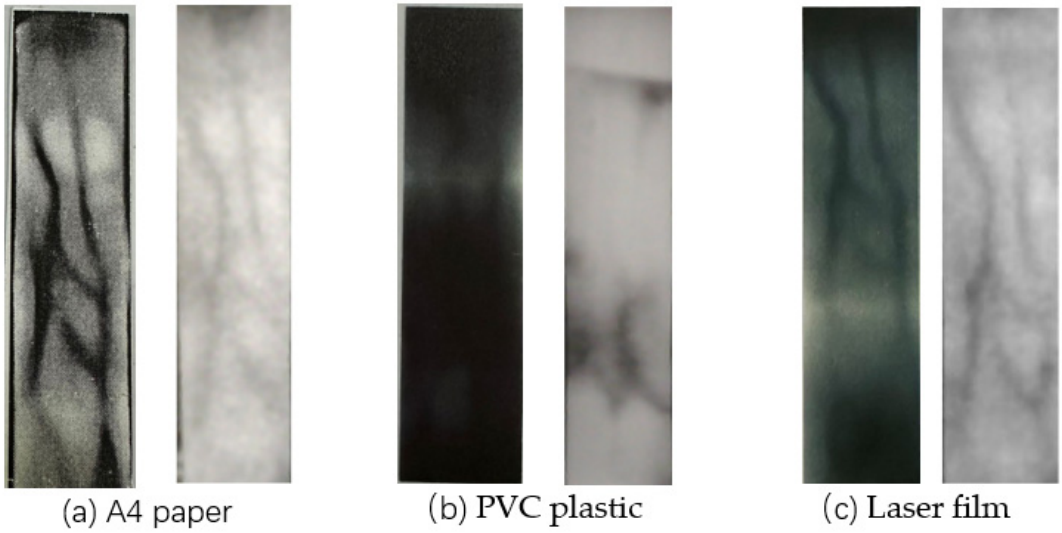

Figure 8. Three prosthesis models and corresponding acquisition diagrams.

Real veins exhibit consistency in color and texture within non-venous areas. To simulate this characteristic in prostheses, we initially applied Gaussian filtering to blur real vein images and eliminate noise. Subsequently, we enhanced vein features through contrast-limited histogram equalization. Following this, we performed adaptive regional threshold segmentation, using the resulting image as a mask for applying local power-law transformations to the original image. Experimental results demonstrate the quality of the acquired images of our fabricated prostheses, as illustrated in the Figure 9 below.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

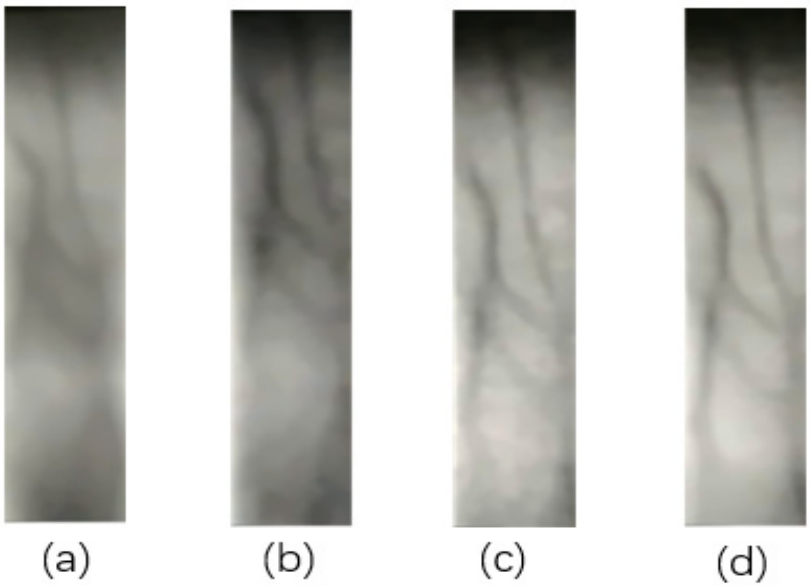

Figure 9. (a) Living sample; (b) A4 paper; (c)PVC plastic; (d) Laser printing film.

Before the model training, all video samples are preprocessed and converted into MSTmap representation as input data. The size of MSTmap image is (224,180,3). Finally, 400 MSTmap images are sent to the network or used in the experiment.

4.2. Model parameters

In the process of network training, the following parameters are set: the batch size is 32; the number of parallel tasks is 8; the number of training rounds is set to 300; the initial learning rate is 0.0001 and dynamically adjusted according to the cosine annealing strategy. The optimizer is the adaptive moment estimation method; the loss function is cross entropy loss.

The loss function is used to measure the network output and label error. The purpose of neural network back propagation is to reduce the error between output and label. The main task of the network in this paper is classification. According to the characteristics of the loss function, the loss function is cross entropy.

4.3. Evaluation indicators

In order to objectively evaluate the performance of the finger vein PAD algorithm and compare the performance between the implemented algorithms, researchers usually use the relevant evaluation indicators in international standards [32]. The indicators that evaluate the detection effect of the liveness detection system usually include Accuracy rate (ACR), Attack Presentation Classification Error Rate (APCER) and Bona fide Presentation Classification Error Rate (BPCER).

The calculation formulas of APCER and BPCER are as follows:

$$APCER_{PAIS} = 1 - \left(\frac{1}{N_{PAIS}}\right)\sum_{i=1}^{N_{PAS}} RES \tag{7}$$

$$BPCER = \frac{\sum_{i=1}^{N_{PAS}} RES}{N_{BF}} \tag{8}$$

Among them, the number of prosthesis samples is the number of real samples. If the i sample is classified as a prosthesis, it is 1, and vice versa is 0. The biometric anti-counterfeiting system should pursue higher security, so the higher the APCER, the better the security of the system, taking into account the two indicators of ACR and BPCER.

4.4. Experiment

Initially, we transform video samples into MSTmap images to extract dynamic liveness features. Subsequently, these MSTmap images are fed into an enhanced ViT network for classification to determine if the sample represents an authentic finger vein pattern. Compared to the DFT+SVM

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

method [33], our approach eliminates the need for intricate frequency domain transformations and exhibits superior generalization capabilities. In contrast to the EVM+MPR method [34], our technique is not dependent on the choice of filtering frequency and showcases heightened resistance to noise interference. Moreover, when compared to the LBP+WLD method [35], our methodology more effectively harnesses the dynamic information present in videos and avoids the information loss associated with texture feature extraction.

Table 1. Comparative Test of Common PAD Methods.

Experimental Method	ACR	APCER	BPCER
LBP+WLD	0.7875	0.325	0.2861
EVM+MPR	0.8292	0.2083	0.1583
DFT+SVM	0.9104	0.0833	0.0917
MSTmap+Light-ViT	**0.9963**	**0**	**0.0037**

Based on our experimental results, the liveness feature extraction method, MSTmap combined with Light-ViT classification, achieved the highest accuracy rate of 99.63%. In comparison, the DFT+SVM and EVM+MPR methods registered accuracy rates of 91.04% and 82.92%, respectively. The approach employing LBP+WLD yielded the lowest accuracy on the dataset, standing at a mere 78.75%.

The advent of lightweight networks provides crucial technical support for biometric systems. By integrating these lightweight networks into biometric systems, we can significantly reduce the consumption of computational resources while still capturing and identifying essential details within biometric features. To further assess the performance of the Light-ViT network in this task and make comparisons with other influential and representative networks in the field, this study selected the following networks for comparative analysis:(1)VGG-16 is a CNN-based image classification network comprising 16 layers of depth and the utilization of compact convolutional kernels [36].(2)ResNet50: A CNN-based architecture grounded in residual learning principles, capable of training networks with a depth of 50 layers while mitigating issues like gradient vanishing and degradation [37].(3)ViT: A vision transformer-based image classification network that leverages self-attention mechanisms to capture global features, demonstrating superior computational efficiency and accuracy compared to traditional CNNs [38].(4)MobileNetV2: An example of a lightweight convolutional neural network that substantially reduces network parameters through techniques like depth-wise separable convolutions [39]. The outcomes, including test set accuracy and loss curves, are detailed in Table 2 and illustrated in Figure 8, respectively.

Table 2. Comparative Experimental Results.

Network Name	ACR	APCER	BPCER
VGG16	0.9247	0.0803	0.0712
ResNet50	0.9687	0.0364	0.0367
ViT	0.9722	0.0294	0.0299
MobileNetV2	0.9725	0.0281	0.0307
Light-ViT	**0.9963**	**0**	**0.0037**

From the experimental results, it can be seen that Light-ViT is optimal on the MFVD dataset with 99.63% accuracy and 0 false acceptance rate. Among the tested networks, MobileNetV2, ResNet50, and ViT demonstrate the highest performance, achieving accuracies of 97.25%, 97.22%, and 96.87%, respectively. In contrast, VGG16 lags behind as the least effective option, achieving an accuracy of 92.47%. These results underscore the efficacy of the Light-ViT network in effectively amalgamating

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

the strengths of both CNN and ViT architectures, enabling the learning of both global and local features within an image, consequently enhancing detection accuracy. The variation of the network loss function in the above experiment is shown in Figure 10.

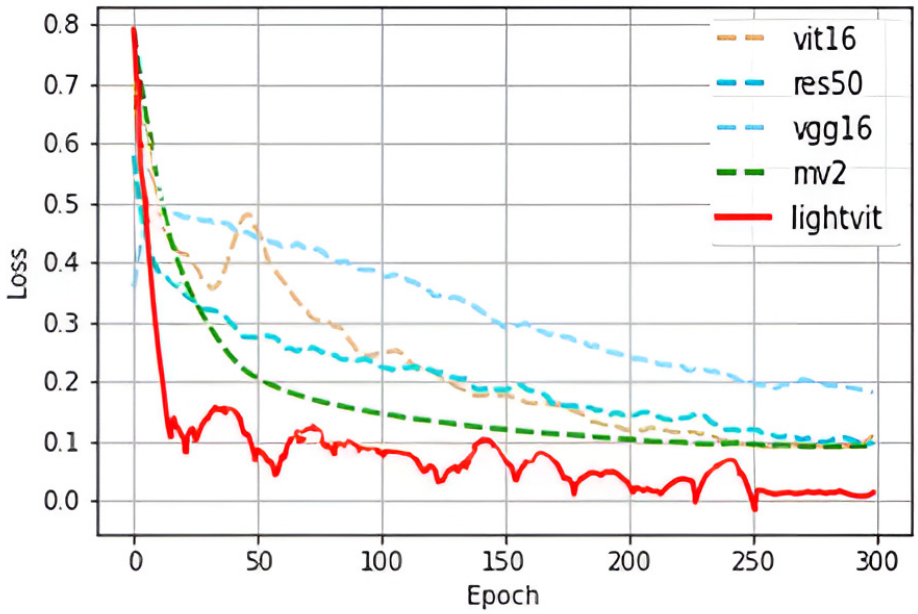

Figure 10. Loss Curve.

The parameter and computational load for each network are detailed in Table 3. In terms of network size and computational demand, Light-ViT exhibits a notable advantage. Its lightweight architecture can significantly reduce the system's response time.

Table 3. Network Parameters.

Network Name	Total params（M）	Params size（MB）	GFLOPS（M）
VGG-16	134.269	512.19	30932
ResNet	23.512	89.69	8263
ViT	85.648	326.72	33726
MobileNetV2	2.226	8.49	652.419
Light-ViT	**1.107**	**4.22**	**690.217**

To evaluate the influence of our network structure improvements on the final results, we replaced the L-ViT block in the Light-ViT network architecture with a standard convolution layer of size 3x3, modified the number of convolution channels, and removed the linear bottleneck structure, resulting in the baseline network (Basenet). The experimental results are presented in Table 4.

Table 4. Results of ablation experiment.

Network Structure	ACR	APCER	BPCER
Basenet	0.9090	0.0928	0.1005
Basenet + L-ViT block	0.9809	0.0147	0.0239

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 October 2023 doi:10.20944/preprints202310.1764.v1

Basenet+ bottleneck	0.9287	0.0603	0.0716
Basenet + all	**0.9963**	**0**	**0.0037**

The experimental results indicate that the proposed L-ViT block significantly enhances the network's performance. By introducing the ViT structure, the network gains the capability to learn global features. To validate the role of MSTmap in aiding subsequent network feature recognition, we designed the following comparative experiments. Firstly, we directly used frames extracted from the vein videos, randomly selecting 4 frames from each video sample, amounting to 7,056 training images. These were then trained using the Light-ViT network. The graph below illustrates the loss curve from the ablation study. The Loss plot is shown in Figure 11.

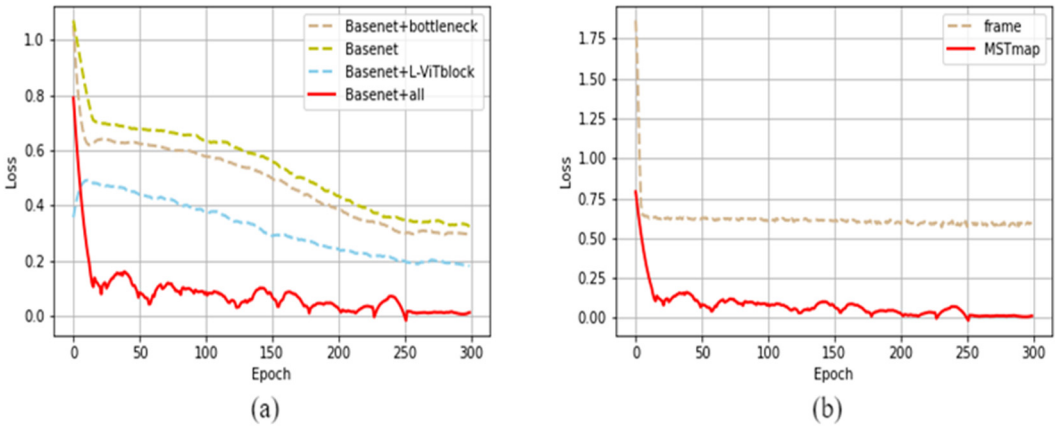

Figure 11. (a) : Loss curve of ablation experiment; (b) : MSTmap experimental loss curve.

To evaluate the performance of the vein anti-counterfeiting identification system proposed in this study, which incorporates a lightweight network, we chose finger vein recognition as a specific application scenario and compared it with other representative network structures. We employed the MFVD dataset, FV-USM dataset, and VERA dataset, conducting data augmentation on them and splitting them proportionally into training and validation sets, as detailed in Table 5.

Table 5. Dataset Composition.

Dataset	**Amount (picture)**	**Class**	**Proportions**
MFVD	17280	288	8：2
FV-USM	23616	492	8：2
VERA	14080	220	8：2

We also conducted comparative experiments on other lightweight networks, and the experiments demonstrated that our proposed Light-ViT achieved better performance on three different datasets. Table 6 is used to illustrate their comparative results with the use of ACR

Table 6. Identity Identification Experiment Results and Network Parameters.

Network Structure	**ACR**		
	MFVD	VERA	FV-USM
VGG16	0.9588	0.9172	0.9285
ResNet50	0.9687	0.9426	0.9428

 doi:10.20944/preprints202310.1764.v1

ViT-16	0.9699	0.9359	0.9471
MobileNetV2	0.9684	0.9428	0.9452
Light-ViT	**0.9881**	**0.9612**	**0.9702**

The Light-ViT network, as introduced in this paper, demonstrates commendable performance in the identification task, attaining an accuracy of 98.81% on the MFVD dataset. This achievement surpasses that of other conventional CNN and ViT networks. Notably, even though all the alternative network architectures achieved accuracy rates exceeding 95% on the MFVD dataset, the Light-ViT network distinguishes itself by offering a more compact parameter size and requiring less computational effort. These attributes render it well-suited for deployment and operation on mobile devices. Consequently, Light-ViT emerges as a neural network capable of effectively harnessing global information processing capabilities, with substantial advantages in identity recognition tasks.

5. Conclusions

This paper proposes a video detection system for vein pattern anti-counterfeiting recognition. In this system, we propose a lightweight network and enhance its capacity to learn global features by introducing L-ViT blocks. Additionally, we replace standard convolutions with depth-wise separable convolutions to construct MN blocks, reducing the computational cost of convolution operations and consequently reducing the network's size and computational load. We embed this lightweight network into the system and transform video data into multi-scale spatio-temporal graphs to allow the network to extract more feature information. Experimental results demonstrate that Light-ViT offers fast computation, a compact parameter footprint, and achieves higher recognition accuracy compared to other lightweight networks. The integrated detection system with this network exhibits superior real-time performance and usability.

Author Contributions: Conceptualization, L.C. and T.G.; methodology, L.C. and H.J.; software, H.J and T.G.; validation, W.L. and L.L.; formal analysis, L.C. and Z.L.; data curation, W.L. and L.L.; writing—original draft preparation, T.G.; writing—review and editing, L.C. and T.G.; supervision, L.C. and Z.L.; All authors have read and agreed to the published version of the manuscript.

Funding: This research of this paper is supported by these funds: (1) The Natural Science Foundation of Chongqing, No. cstc2020jcyj-msxmX0818, CSTB2023NSCQ-MSX0760 and cstc2021ycjh-bgzxm0071. (2) The Science Technology Research Program of Chongqing Municipal Education Commission, No. KJZD-M202301502, KJQN201901530, KJQN202300225, and Chongqing postgraduate education 'curriculum ideological and political' demonstration project, No. YKCSZ23193. (3) Graduate Innovation Program Project of Chongqing University of Science & Technology, No. YKJCX2220803, ZNYKJCX2022008, and ZNYKJCX2022017. (4) The Foundation of Wuhan Maritime Communication Research Institute, China State Shipbuilding Corporation, No. KCJJ2020-6.

Data Availability Statement: Not applicable

Acknowledgments: Not applicable

Conflicts of Interest: All authors declared that there are no conflicts of interest.

References

1. Ali S F, Khan M A, Aslam A S. Fingerprint matching, spoof and liveness detection: classification and literature review[J]. Frontiers of Computer Science, 2021.
2. Mitsutoshi Himaga and Katsuhiro Kou. Finger vein authentication technology and financial applications. In Advances in Biometrics, pages 89–105. Springer, 2008.
3. Liu Y, Jia Y, Zhang Z, et al. Deep learning for face anti-spoofing: a survey[J]. IEEE Transactions on Biometrics Behavior and Identity Science, 2021.
4. Tome P, Vanoni M, Marcel S. On the vulnerability of finger vein recognition to spoofing[C], Biometrics Special Interest Group. IEEE, 2014:1-10.
5. Remya Krishnan, Image Quality Assessment For Fake Biometric Detection: Application To Finger-Vein Images. International Journal of advanced research, 4(8), 2016 :2015-2021.
6. Keke，Hitachi debuts a smartphone with vein recognition technology [in Chinese]，Sohu, 2016.10.30，https://www.sohu.com/a/117647702_122331.

7. Tome P, Raghavendra R, Busch C. On the vulnerability of palm vein recognition to spoofing attacks[C]//2015 International Conference on Biometrics (ICB). IEEE, 2015: 1-8.
8. Ramachandra R , Busch C . Presentation Attack Detection Methods for Face Recognition Systems - A Comprehensive Survey[J]. ACM Computing Surveys (CSUR), 2017, 50(1):8.1-8.37.
9. Rehman, Yasar, Abbas, et al. LiveNet: Improving features generalization for face liveness detection using convolution neural networks[J]. Expert Systems with Application, 2018.
10. Hu M , Qian F , Guo D , et al. ETA-rPPGNet: Effective Time-Domain Attention Network for Remote Heart Rate Measurement[J]. IEEE Transactions on Instrumentation and Measurement, 2021, PP(99):1-1.
11. Yu L , Du B , Hu X , et al. Deep spatio-temporal graph convolutional network for traffic accident prediction[J]. Neurocomputing, 2021, 423:135-147.
12. Dosovitskiy A , Beyer L , Kolesnikov A , et al. An Image is Worth 16 × 16 Words: Transformers for Image Recognition at Scale[J]. 2020.
13. Chen X , Huang M , Fu Y . Simultaneous acquisition of near infrared image of hand vein and pulse for liveness dorsal hand vein identification[J]. Infrared Physics & Technology, 2021, 115(1):103688..
14. J Määttä, Hadid A , M Pietikäinen. Face spoofing detection from single images using micro-texture analysis[C]//International Joint Conference on Biometrics. IEEE, 2011
15. Chingovska I , Anjos A , Marcel S . On the Effectiveness of Local Binary Patterns in Face Anti-spoofing[J]. IEEE, 2012.
16. Tome P , Raghavendra R , Busch C , et al. The 1st Competition on Counter Measures to Finger Vein Spoofing Attacks[C]//2015 International Conference on Biometrics (ICB). IEEE, 2015.
17. Fang Y, Wu Q, Kang W. A novel finger vein verification system based on two-stream convolutional network learning[J]. Neurocomputing, 2018, 290: 100-107.
18. Bok J Y, Suh K H, Lee E C. Detecting fake finger-vein data using remote photoplethysmography[J]. Electronics, 2019, 8(9): 1016.
19. Pouyanfar S, Sadiq S, Yan Y, et al. A survey on deep learning: Algorithms, techniques, and applications[J]. ACM Computing Surveys (CSUR), 2018, 51(5): 1-36.
20. Assim O M, Alkababji A M. CNN and genetic algorithm for ginger vein recognition[C].2021 14th International Conference on Developments in eSystems Engineering (DeSE). IEEE, 2021: 503-508.
21. Baweja Y, Oza P, Perera P, et al. Anomaly detection-based unknown face presentation attack detection[C]. 2020 IEEE International Joint Conference on Biometrics (IJCB). IEEE, 2020: 1-9.
22. Zeng J, Wang F, Deng J, et al. Finger vein verification algorithm based on fully convolutional neural network and conditional random field[J]. IEEE access, 2020, 8: 65402-65419.
23. T Tao Z, Wang H, Hu Y, et al. DGLFV: deep generalized label algorithm for finger-vein recognition[J]. IEEE Access, 2021, 9: 78594-78606.
24. Gionfrida, L., Rusli, W.M., Kedgley, A.E., & Bharath, A.A. (2022). A 3DCNN-LSTM Multi-Class Temporal Segmentation for Hand Gesture Recognition. Electronics.
25. Zhou, L., Yang, L., Fu, D., & Yang, G. (2022). SIFT-Flow-Based Virtual Sample Generation for Single-Sample Finger Vein Recognition. Electronics.
26. Kono M, Ueki H, Umemura S. Near 鄄 infrared finger vein patterns for personal identification [J]. Appl Opt, 2002, 41 (35): 7429-36.
27. Yang J, Zhang X. Feature -level fusion of fingerprint and finger 鄄 vein for personal identification [J]. Pattern Recogn Lett, 2012, 33(5): 623-628
28. Sarwar, Syed Shakib et al. “Gabor filter assisted energy efficient fast learning Convolutional Neural Networks.” 2017 IEEE/ACM International Symposium on Low Power Electronics and Design (ISLPED) (2017): 1-6.
29. Geirhos, Robert et al. “ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.” ArXiv abs/1811.12231 (2018): n. pag.
30. Boughida A, Kouahla M N, Lafifi Y. A novel approach for facial expression recognition based on Gabor filters and genetic algorithm[J]. Evolving Systems, 2022, 13(2): 331-345.
31. Howard A G, Zhu M, Chen B, et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications[J]. arXiv preprint arXiv:1704.04861, 2017.
32. ISO/IEC 30107-3-2017 Information technology — Biometric presentation attack detection — Part 3: Testing and reporting.
33. Chen X, Huang M, Fu Y. Simultaneous acquisition of near infrared image of hand vein and pulse for liveness dorsal hand vein identification[J]. Infrared Physics & Technology, 2021, 115: 103688.
34. Lin B, Li X, Yu Z, et al. Face liveness detection by rppg features and contextual patch-based cnn[C].Proceedings of the 2019 3rd international conference on biometric engineering and applications. 2019: 61-68.
35. Park K R. Finger vein recognition by combining global and local features based on SVM[J]. Computing and Informatics, 2011, 30(2): 295-309.

36. Song J M, Kim W, Park K R. Finger-vein recognition based on deep DenseNet using composite image[J]. Ieee Access, 2019, 7: 66845-66863.
37. Chen L, Li S, Bai Q, et al. Review of image classification algorithms based on convolutional neural networks[J]. Remote Sensing, 2021, 13(22): 4712.
38. Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16 × 16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.
39. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]. Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.