Study of Full-View Finger Vein Biometrics on Redundancy Analysis and Dynamic Feature Extraction

Junduan Huang, Sushil Bhattacharjee, Sébastien Marcel, Wenxiong Kang

***Abstract*—As a biometric trait drawing increasing attention, finger vein (FV) has been studied from many perspectives. One promising new direction in FV biometrics research is full-view FV biometrics, where multiple images, covering the entire surface of the presented finger, are captured. Full-view FV biometrics presents two main problems: increased computational load, and low performance-to-cost ratio for some views/regions. Both problems are related to the inherent redundancy in vascular information available in full-view FV images. In this work, we address this redundancy issue in full-view FV biometrics. Firstly, we propose a straightforward FV redundancy analysis (FVRA) method for quantifying the information redundancy in FV images. Our analysis shows that the redundancy ratio of full-view FV images is up to 83%-87%. Then, we propose a novel feature extraction model, named FV dynamic Transformer (FDT), whose architecture is configured based on the redundancy analysis results. The FDT focuses on both local (single-view) information as well as global (full view) information at different processing stages. Both stages provide the advantage of de-redundancy and noise avoidance. Additionally, the end-to-end architecture simplifies the full-view FV biometrics pipeline by enabling the direct, simultaneous processing of multiple input images, thus consolidating multiple steps into one. A series of rigorous experiments is conducted to evaluate the effectiveness of the proposed methods. Experimental results show that the proposed FDT achieves state of the art authentication performance on the MFFV-N dataset, yielding an EER of 0.97% on the development set and an HTER of 1.84% on the test set under the balanced protocol and EER criterion. The cross-domain generalization capability of FDT is also demonstrated on the LFMB-3DFB dataset, where it achieves an EER of 7.24% and an HTER of 7.34% under the same protocol and criterion. Code for the proposed methods can be access via: https://github.com/SCUT-BIP-Lab/FDT.**

***Keywords*—Biometrics, vein recognition, finger vein, full-view, dynamic Transformer**

I. Introduction

Security applications relying on authentication of individual identities are increasingly relying on biometrics. In recent years finger vein (FV) biometrics has attracted growing attention, due to the three main advantages it offers, namely, liveness detection, difficulty of counterfeiting, and user-friendliness [2, 3].

In FV biometrics, multi-view FV authentication (FVA) is a topic drawing increasing interest and has the potential to be widely used. Compared to single-view FV biometrics, multi-view FV biometrics 1) captures more information, which can lead to higher recognition accuracy; 2) significantly increases the difficulty of constructing presentation attacks (PA), and thus is more secure; and 3) inherently solves the pose variation problem during application [4], making the biometrics system more robust. Figure 1 shows an example of a multi-view FV biometrics system, the Mirror-based Full-view FV device

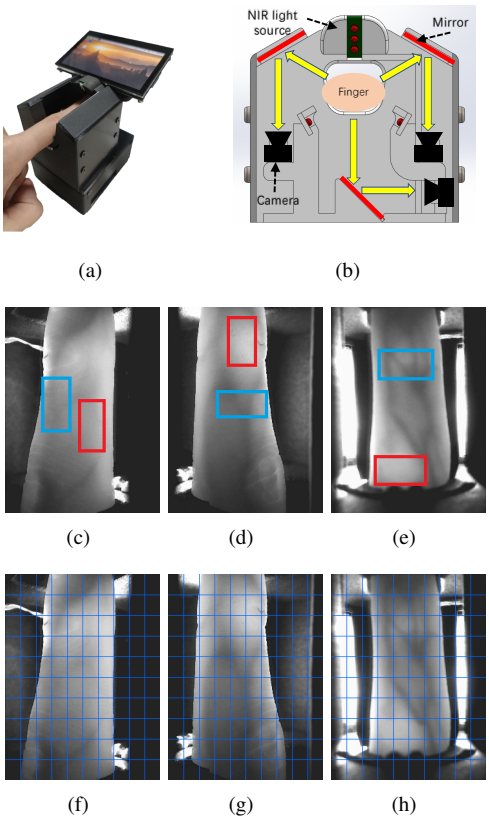


Fig. 1: Full-view FV biometrics, taking MFFV [1] as an example. (a) MFFV imaging device; (b) Schematic layout of MFFV imaging device; (c)-(e) Full-view FV images of View 1, View 2, and View 3, respectively (the blue boxes indicate regions with higher density of veins, and the red boxes indicate regions with fewer veins); (f) - (h) Patches division for View 1, View 2, and View 3, respectively.

This work was supported in part by the Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515110009, National Natural Science Foundation of China under Grant 61976095 and Grant 62376100. (Corresponding author: Wenxiong Kang.)

Junduan Huang is with School of Artificial Intelligence at South China Normal University, Foshan, 528225, China; and School of Automation Science and Engineering at South China University of Technology, Guangzhou, 510641, China; and Biometrics Security and Privacy Group at Idiap Research Institute, Martigny, 1920, Switzerland. (email: runrunjun@163.com).

Sushil Bhattacharjee and Sébastien Marcel are with Biometrics Security and Privacy Group at Idiap Research Institute, Martigny, 1920, Switzerland. (e-mail: sushil.bhattacharjee@idiap.ch; sebastien.marcel@idiap.ch).

Wenxiong Kang is with School of Automation Science and Engineering at South China University of Technology, Guangzhou, 510641, China. (e-mail: auwxkang@scut.edu.cn).

(MFFV) [1]. The external view of the MFFV device is shown in Figure 1(a) and the internal arrangement of cameras and mirrors in the device is shown in Figure 1(b). In Figure 1(c) - 1(e), we show examples of FV images captured using this device.

There are two kinds of taxonomy [1] for the current multi-view FV biometrics studies. The first taxonomy is modality-based, and groups the existing works into three categories: **partial multi-view FV**, **full-view FV**, and **three-dimensional (3D) FV**. The second taxonomy groups them by imaging technology, into **multi-camera-based systems** and **moving-camera-based systems**. Any multi-view FV biometrics system, regardless of taxonomy, needs to handle multiple FV images simultaneously. This requirement entails the following challenges: **1) Increased computational complexity:** the process of extracting features from multiple FV images will notably increase the computational complexity. **2) Low performance-to-cost ratio for certain views and regions:** due to the natural distribution property of FV trait, FV information is variable in different locations. As shown in Figure 1, the palmar view of finger (1(e)) has more FV information than the lateral views of the finger (Figure 1(c) and 1(d)). Also, certain regions, for example those marked with red rectangles, include relatively less FV information compared to other regions, such as those marked with blue rectangles in Figure 1(c) and 1(d). Therefore, if all image-regions are processed in the same way, this will result in low performance-to-cost ratio for regions, or even entire views containing low FV information.

These two challenges can be seen as different manifestations of redundancy in multi-view FV biometrics. The increased computational complexity reflects the macroscopic manifestation of the redundancy issue; whereas the low cost performance index for some views and regions reflects the microscopic embodiment. This redundancy issue is the main reason that most of the existing FV biometrics studies only focus on single-view FV paradigm. Simultaneously, it is the main obstacle to the development of multi-view FV biometrics.

To advance the state of the art (SOTA) in multi-view FV biometrics, the redundancy issue should be addressed. In this work we propose a series of methods for estimating and reducing redundancy in feature extraction of multi-view FV biometrics. Our proposed approach, based on dividing the FV images into patches, is inspired by the approach adopted in Vision Transformer (ViT) [5]. **Firstly, we propose a FV redundancy analysis (FVRA) method to estimate the redundancy ratio of both single-view and full-view FV data quantitatively.** As shown in Figure 1(f)-1(h), each view of FV image is divided into a set of patches for redundancy analysis, which is the manner for handling FV image with region variant and spatial texture characteristics. To the best of our knowledge, this is the first study on redundancy analysis for FV images. **Then, based on the redundancy analysis results, we propose a Full-view FV Dynamic Transformer (FDT) model. The proposed FDT model gradually reduces the token sequence during the feature extraction process, thereby, reducing the redundancy significantly, during feature extraction.** The proposed FDT incorporates discriminative information from individual single-view images along with relational information shared across multiple views. There are two kinds of feature extraction stages in FDT: the single-view stage (SV stage) and the full-view coupling stage (FCV stage), which are used as follows: **1) During the SV stages**, each view is processed independently of the other views and different views are treated independently. Hence enhance the feature extraction of specific characteristics inner each view. As the stage progresses, the number of tokens that contribute little to identity-discrimination are removed, and the tokens that are important to identity-discrimination are retained. **2) In the FVC stages**, the intermediate tokens from each SV stage are coupled together. This enables the FDT to extracted features that integrate information from all views. The FVC stages also continue to reduce the number of FV tokens, to further reduce redundancy. Finally, the class token output by the last FVC stage – a short (128-dimension) vector, is used as the biometrics template.

We have conducted a series of experiments to evaluate the performance of the proposed FDT on two publicly available full-view FV datasets. Experimental results show that the FDT achieves SOTA performance for full-view FV authentication, and effective cross-domain generalization performance.

The main contributions of this work are summarized hereunder:

1) This is the first work to address the redundancy issue, which is an significant obstacle in the development of FV biometrics, especially for multi-view FV biometrics. A straightforward and effective method, namely FVRA, for analyzing the redundancy for FV biometrics. The proposed FVRA provides a quantitative measure of the redundancy present in FV data.
2) Based on the redundancy analysis results, this work proposes a novel full-view FV feature extraction model, namely FDT. The proposed FDT considers both the information specific to each view, as well as the correlation among full views.
3) Experimental results on two publicly available full-view FV datasets demonstrate that the proposed FVRA and FDT effectively address the redundancy issue in FV biometrics. The proposed FDT achieves SOTA performance for full-view FV authentication, and effective cross-domain generalization performance.

The rest of this paper is organized as follows. We discuss the related works in Section II. Then, we describe the FVRA method in Section III, and introduce the design of the proposed FDT in Section IV. Experimental results for evaluating the proposed methods are presented in Section V. Finally, we conclude our work in Section VI.

II. Related Works

In this section, we mainly discuss the related works of FV biometrics, starting with the single-view FV biometrics, followed by a focus on multi-view FV approaches. We also provide a concise overview of vision Transformers as a relevant technical foundation.

A. FV biometrics: from single-view to multi-view

Recent deep-learning-based methods have shown excellent performances for FVA. Das *et al*. [6] investigated the performance of convolutional neural networks (CNN) for FV identification tasks. Their experimental results show that it is possible to achieve rank-1 identification accuracy higher than 95% for some publicly available FV datasets. Yang *et al*. [7] proposed a feature-fusion method for small-area FV recognition. Recognition for partial FV image is common in practical application, especially for some low-constraint FV biometrics system. Kuzu *et al*. [8] proposed an on-the-fly FV recognition system, which uses a CNN for feature extraction and a recurrent neural network (RNN) to further extract the temporal dependence of the input features. Li *et al*. [9] presented a novel compact multi-representation feature descriptor (CMrFD) as FV feature, which has the characteristic of visual and semantic consistency.

Following the development of Vision-Transformer neural network architecture [5], Huang *et al*. [10] first adapted the Transformer encoder for FV authentication task. By adopting several novel local information enhanced techniques, they proposed a customized FV Transformer, namely FVT. Inspired by the difference of filtered results of FV images using different frequency filters, Huang *et al*. [3] proposed a frequency domain processing (FDP) module, which can adaptively learn the frequency filters through training. Experimental results show that their proposed FDP module can achieve satisfactory performance on FV authentication task. Further, they combine the FDP with a spatial CNN for coupling the frequency and spatial features of an FV image, and proposed a lightweight model, namely FVFSNet (FV frequency-spatial coupling network). Experimental results show that the FVFSNet achieved overall SOTA FVA performance on nine FV datasets.

With the extensive study of finger veins, some challenges have begun to emerge, such as the requirement for higher accuracy and the risk of being counterfeit attack. To tackle these two challenges, the multi-view biometric modality is a good response due to more information is captured, which will improve the accuracy of identity-discrimination and difficulty of being counterfeited. For FV biometrics, there are three kinds of multi-view modalities: partial multi-view FV, full-view FV, and 3D FV [1].

Partial multi-view FV systems capture a set of FV images from different angles, but the different views do not cover the entire surface of the finger. Bunda [11] reconstructed the 3D point cloud from multiple palmar FV images. These images were captured from different angles by the device of Rozendal [12]. Zhao *et al*. [13] also developed a partial multi-view FV imaging device. Their system is different from that of Rozendal [12] in two ways: (1) to capture different views of the presented finger, the user is required to rotate the finger in the device, and (2) the FV images are captured from the dorsal side.

Full-view FV systems capture a set of FV images that cover the entire finger surface. It cantains more information compared with partial multi-view FV, and therefore can make full use of the vein information on the fingers. Published works on full-view FV biometrics can be grouped into two categories: the moving-camera based methods [14–16] and the multi-camera based methods [1, 4, 17]. Prommegger *et al*. [14] have designed a moving-camera based full-view FV biometrics system. They have collected a new full-view FV dataset using their proposed device, for which they have presented a set of experimental results. Considering only single-view FVA, their experiments indicate that FV images from palmar and dorsal views lead to better results than those from other views. They further demonstrate that score-fusion based multi-view FVA can lead to better results than single-view FVA. Qin *et al*. [15] have also proposed a full-view FV imaging device based on the moving-camera idea. They have also proposed a Transformer-based model that models the features from the different views as a sequence to fit the Transformer architecture. Song *et al*. [16] also adopt the moving-camera design and proposed a full-view FV imaging system. They have used their prototype device to capture a dataset with eight views per finger. They also propose the MFV-FESNet architecture that extracts the context feature and the dominant feature from the full-view FV images and fuses them to generate the final feature vector.

Huang *et al*. [1] propose a mirror-based full-view FV (MFFV) imaging system. Their system follows the multi-camera design. There are two main contributions in this work: first, the mirror-based FV imaging design reduces the device size, which is important for practical applications; second, the multiple illumination FV capture strategy (MIFV) and the accompanying FV illumination adaption (FVIA) algorithm address the illumination issue caused by finger-thickness variations. Their proposed system is designed to use optimally illuminated FV images to improve both single-view and full-view FVA performance. Kauba *et al*. [18] also propose a mirror-based full-view FV imaging device. Their device only use a single camera and several mirrors to capture FV from different views. However, because of varying imaging distances for the different views, there are discrepancies among the different views captured by their system.

Another dimension of full-view FV biometrics is the 3D FV modality, where the full-view FV images are used to first reconstruct the 3D vein-pattern in the presented finger. After reconstructing the 3D representation of the finger, Kang *et al*. [4] unfold the 3D information into a 2D texture image and a 2D representation of the finger-shape. Their experimental results show that the combination of these two cues leads to superior FVA results, compared to using FV texture information alone. Yang *et al*. [17] propose a visual hull based 3D finger reconstruction method. Since the visual hull based 3D reconstruction requires as many views as possible, they have developed a six-view imaging device and have collected a finger multi-biometrics dataset namely LFMB-3DFB. Their experiments show that although the proposed visual-hull based method can reconstruct the 3D finger, there is room for improvement in some details such as edges and corners.

In contrast to the works discussed above, this work introduces a framework guided by an analysis of redundancy in full-view FV data. The redundancy analysis directly informs the design of an efficient, end-to-end feature extraction model for full-view FV recognition.

B. Vision Transformer

The Transformer implements the multi-head self-attention (MHSA) mechanism [19]. This architecture was first proposed for text-processing application. Dosovitskiy *et al*. [5] proposed a Vision Transformer (ViT) for image-processing applications by converting images into a sequence of 16 x 16 patches. This pioneering work, demonstrating the effectiveness of ViT in computer-vision (CV) tasks has inspired many CV applications of ViT [20, 21]. Liu *et al*. [22] propose a Swin Transformer, which uses a hierarchical architecture and applies the MHSA with local windows that are shifted across layers. Li *et al*. [23] propose a Local-ViT, which uses a locality-enhanced feed-forward network (LFFN) to gather the information from local tokens, thus enhance the local information extraction ability. Chu *et al*. [24] propose a position embedding generator (PEG) for generating the position embedding according to the given token and its adjacent tokens. These three works have proposed improvements to the original ViT by taking into account hierarchical structure, local information, and position embedding. Many contemporary Vision Transformer architectures draw inspiration from the key ideas behind these enhancements, even if they do not incorporate the techniques directly.

The recently proposed dynamic Tranformer (DT) architecture improves upon the ViT by processing token-sequences using a dynamic strategy that can reduce the sequence-length significantly. These dynamic strategies can generally be grouped into two categories: learning-based [25, 26] and class-token-based [27, 28]. The learning-based strategy uses a sub-network (also called decision-network) to assign a score to each token; tokens with scores lower than a preset threshold are *pruned*, that is, excluded from subsequent processing. Rao *et al*. [25] propose a DynamicViT, which prunes redundant tokens progressively by using a lightweight prediction module to estimate the importance score of each token. Meng *et al*. [26] propose an adaptive computation framework, namely AdaViT, which can dynamically adjust the number of patches, self-attention heads, and Transformer blocks, based on their corresponding decision networks. The Class-token based strategy generates the token scores directly from the attention value between the content tokens and the class token. Fayyaz *et al*. [27] propose a Adaptive Token Sampler (ATS) module, which can selects the most informative and distinctive tokens based on their attention values. Liang *et al*. [28] propose the EViT, which uses the attention values to identify the important tokens, thus reducing the sequence between MHSA and FFN modules.

Inspired by the dynamic mechanism of the DTs, this work introduces a novel full-view FV authentication approach that leverages this principle to reduce redundancy in intermediate features during feature extraction. By integrating this with a dedicated analysis of FV images redundancy, our method achieves efficient and focused feature learning.

III. Redundancy analysis for FV images

In this section, we present a method for analyzing redundancy in FV images that is equally applicable to both single-view and multi-view FV imagery. We refer to the proposed approach as FV redundancy analysis (FVRA). For FV biometrics, the location of veins can be considered as the distribution of identity-discriminant information, and redundancy is an inherent issue in FV images, mainly due to the uneven distribution of veins. As shown in Figure 1(c)-1(e), not all the regions in the FV images have veins. For example, the regions enclosed in red boxes do not show obvious vein, whereas the blue boxes indicate the vein-rich regions. Veins constitute a relatively small portion of the captured image. For example, Figures 1(f)-1(h) show FV images of three views of a single finger. In these figures, the FV image is divided into 100 non-overlapping rectangular sub-regions. We note from the figure that only about 10-15% of these sub-regions actually contains vein-structures. This is a typical distribution of vein-information in fingers. The uneven distribution of such identity-discriminant information implies that some FV regions are more salient than others [7], or even that some views are significantly better than others [1, 14], in terms of biometrics performance.

To the best of our knowledge, no previous study has addressed the issue of redundancy in FV images. In this work, we propose a straightforward method FVRA, for quantitatively estimating the redundancy in FV images. The redundancy ratio can be used as the basis for designing an efficient FV feature extraction model, which will be introduced in Section IV. The proposed FVRA method is described below.

1) Each FV image, of size 320 $\times$ 240, is divided into 100 non-overlapping patches of size 32 $\times$ 24, as illustrated in Figure 1(f)-1(h).
2) For single-view FV, each FV patch is unraveled into an 1-D vector, called a 'component'. Thus there are 100 components for a single view FV image.
3) For full-view FV, three complementary FV images that cover the whole finger surface are used. Hence, there are 3 $\times$ 100 components for a full-view FV image set.
4) Principal component analysis (PCA) is used to figure out the ratio of principal components. The FVRA framework assesses the information redundancy of the FV image by calculating the proportion of principal components necessary to account for 99% of its total information content. This 99% criterion serves to balance the dual objectives of maintaining recognition accuracy and minimizing redundancy.

Equation 1 shows the mathematical expression of the proposed FVRA metrics, where R is the redundancy-metric, P_{99} and P_t are the number of PCs retained and the number of total PCs.

$$R = 100 \times (1 - \frac{P_{99}}{P_t}) \tag{1}$$

The ratio of the number of selected tokens to the total number of tokens gives a quantitative measure of the redundancy that exists in a given set of FV images – the smaller the ratio, the larger the redundancy. The proposed FVRA method possesses two distinct advantages.

- Firstly, it takes into account the characteristics of image structure through patch-level analysis, thereby preserving the spatial properties of FV images.

- Secondly, by using the subspace learning method – PCA, the proposed FVRA can simulate the interaction between patches, as well as the dimensionality-reduction process implemented in the proposed FDT feature-extractor.

Using the proposed FVRA method conducive to select those tokens of an FV image that together include almost all (99%) of the FV information in the image (or set of full-view FV images).

The experimental details and results are presented in Section V-D, which verify that 1) the redundancy issue exists in FV image heavily; and 2) this issue is amplified in full-view FV. More importantly, in this work, we use the results of FV redundancy analysis to guild the configuration of the feature extraction model, which will be demonstrated in Section IV.

IV. Full-view FV Dynamic Transformer

In this section we describe the proposed Full-view FV Dynamic Transformer (FDT) in detail. Compared to previous works on FV authentication (FVA), the proposed FDT offers three advantages:

1) FDT processes the input data dynamically. During the process of feature extraction, the tokens that contribute less to the identity-discrimination are progressively eliminated, while the tokens that provide significant identity-discrimination are retained for the next layer, making the feature extraction procedure both focused and efficient.
2) The FDT is inherently consider both the local information from individual single-view FV images and the global information from the full-view set. It first processes each view independently to capture view-specific characteristics, and subsequently merges the resulting intermediate features to form a consolidated full-view feature sequence.
3) The FDT offers an end-to-end processing solution for full-view FV image sets. By integrating the entire procedure into a single model, it eliminates the need for dedicated stages — such as preprocessing and a separate fusion step — that are typically required to adapt single-view systems for multi-view data.

There are five key modules in the proposed FDT: the multi-layer perceptron based patch embedding (MLP-P) module, the batch attention (BatchAtten) module, the position embedding generator (PEG), the de-redundancy multi-head self-attention (DeRedun-MHSA), and the local-enhanced feed-forward network (LFFN). In the following, we describe these modules in detail.

A. Overall Architecture and Working Mechanism

As shown in Figure 2, the pipeline of the proposed FDT includes two kinds of stages: single-view (SV) stages and full-view coupling (FVC) stages. During the SV stages, the FV features are extracted from different views independently. As indicated in Figure 1, FV images captured from different views carry complementary information. By processing each view independently in the SV stages, the model captures both the view-specific information and the inter-view complementarity. The input tokens to the SV stages are image patches consisting of pixels that represent low-level information. The SV stages process these image-patches and produce tokens that encode higher-level information. The FVC stages process tokens from all views together, thus aggregating the full-view information into a unified representation. It is these two kinds of information processing stages that realize the feature extraction strategically and efficiently. The full-view coupling dynamic Transformer (FVCDT) and single-view dynamic Transformer (SVDT) layers are built upon the same structural design. They differ exclusively in their input tokens: SVDT handles view-specific tokens, whereas FVCDT processes a combined token set from all views to achieve full-view representation.

The FDT is trained in an end-to-end manner, including both the auxiliary and classifier stages shown in Figure 2. The auxiliary part is described in section IV-F. while the classifier and the losses consist with the works [3, 10]. The cost-function of training FDT is a combination of the cross-entropy loss and the center loss [29], as shown in in Equation 2.

$$L = \alpha L_{CELoss} + \beta L_{CenterLoss} \tag{2}$$

In the inference phase, both the auxiliary and classifier parts are discarded. For a given full-view FV image set, the embedding produced by FVC Stage 6 of the FDT is taken as the corresponding feature vector. Two feature vectors, the enrolled feature and the probe feature, are compared using the Cosine-Similarity measure, to generate an authentication score and output the authentication result by comparing the authentication score with a pre-defined threshold.

B. Multi-layer Perceptron based Patch Embedding

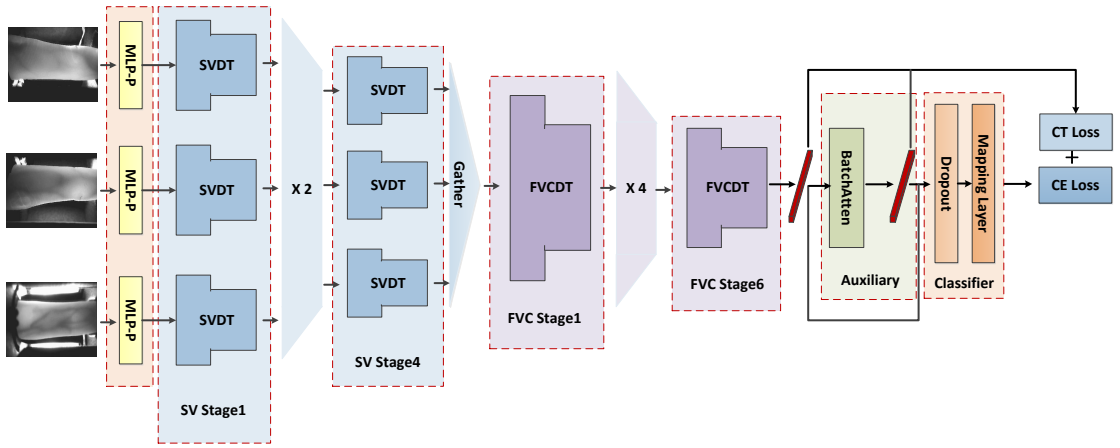
The proposed FDT framework begins by converting the input full-view finger vein (FV) images into token sequences. This transformation is achieved using a multi-layer perceptron-based patch embedding (MLP-P) module, which differs from the tokenization method employed in the original Vision Transformer (ViT).

In the original ViT, tokenization involves a sequence of operations: dividing the image into patches, flattening them, and mapping via a linear layer. This procedure can be viewed as a transformation based on a single linear layer. We argue that this approach has a limitation: it fails to fully capture the information within each patch. To address this, the MLP-P module replaces the single linear layer with a multi-layer perceptron, offering a simple yet effective solution to enhance intra-patch feature extraction.

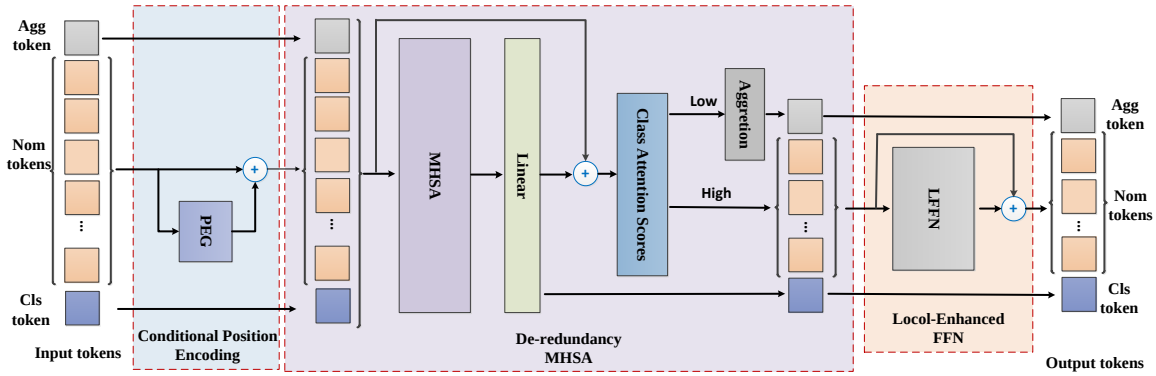
The MLP-P module, as illustrated in Figure 3, is implemented through a specific sequence of operations: a convolutional layer with kernel sizes and strides matching the patch dimensions, a Leaky ReLU activation layer, and a subsequent 1×1 convolutional layer. The mathematical expression for the MLP-P is shown in Equation 3

$$\boldsymbol{X} = \boldsymbol{W_{1\times1}} * \sigma(\boldsymbol{W_{PS-KS}} * \boldsymbol{I} + \boldsymbol{b_1}) + \boldsymbol{b_2} \tag{3}$$

where $\boldsymbol{W_{PS-KS}}$ and $\boldsymbol{b_1}$ denote the kernel and bias of the first convolutional layer; the $\boldsymbol{W_{1\times1}}$ and the $\boldsymbol{b_2}$ represent the



(a)



(b)

Fig. 2: Architecture of the proposed FDT. (a) Overall architecture. MLP-P: the Multi-Layer Perceptron based Patch embedding module, SVDT denotes the Single View Dynamic Transformer module, FVCDT: the Full-View Coupling Dynamic Transformer module, BatchAtten: the Batch Attention module, CT loss: Center loss, CE loss: the Cross-entropy loss; (b) Dynamic Transformer module in FDT. PEG: Position Embedding Generator, MHSA: the Multi-Head Self-Attention module, LFFN: the Local-enhanced Feed Forward Network, Agg token: the aggregation token, Nom tokens: the normal tokens, Cls token: the class token.

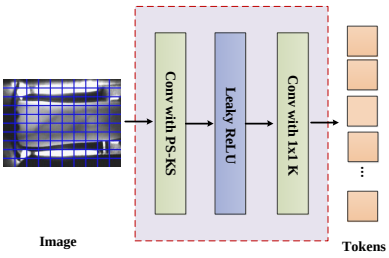


Fig. 3: Architecture of Multi-layer perceptron based patch embedding (MLP-P). PS-KS denotes patch-sizes kernel and stride, 1x1 K denotes one-by-one size kernel.

kernel and bias of the second convolutional layer, $\boldsymbol{I}$ denotes the input image and $\sigma(\cdot)$ denotes the activation function.

Ablation experiments to study the effectiveness of the MLP-P stage are discussed in Section V-G.

C. Position Embedding Generator

In this work, we have used conditional position embedding to represent the position information of FV tokens. The working mechanism of conditional position embedding is based on a *position embedding generator* (PEG) [24], which generates the position embedding based not only on the token itself, but also on its nearest neighbors. For the proposed FDT, we have adapted the PEG in the following two specific details:

- The PEG is used for all dynamic Transformer modules in the proposed FDT. There are two reasons. On one hand, the PEG can enhance the local position information extraction ability with the given token, which is an advantage when processing FV images that are abundant by texture information. On the other hand, as the tokens are dynamically reduced by FDT, the position of each token changes from one stage to the next. Hence it is necessary to update the position information for each stage.
- To use PEG, the token sequence should be reshaped to the 2D feature map format. In the proposed FDT,

before using PEG, each token is reshaped as a pixel in the intermediate feature map. Note that PEG is applied only to normal tokens. For special tokens, namely, class tokens and aggregation tokens, position embedding is not necessary.

We have implemented the PEG as a convolution layer, as expressed in Equation 4, where $\boldsymbol{X}$ denotes the normal tokens input; (X_{PE}) denotes the conditional positional embedding output; $\boldsymbol{W_{PE}}$ and $\boldsymbol{b}$ denote the conditional weights and bias for position embedding (here a 3 $\times$3 kernel is used); M2S-reshape$(\cdot)$ and S2M-reshape$(\cdot)$ denote the map-to-sequence and sequence-to-map reshape operation, respectively.

$$\boldsymbol{X_{PE}} = \text{M2S-reshape}(\boldsymbol{W_{PE}} * \text{S2M-reshape}(\boldsymbol{X}) + \boldsymbol{b}) \quad (4)$$

The effectiveness of using PEG in FDT is studied with the ablation experiment presented in Section V-G.

D. De-Redundancy Multi-Head Self-Attention

As discussed in Section III, redundancy in FV image is an inherent issue in multi-view FV biometrics. The Dynamic Transformer is well suited to handle redundant tokens, because it can dynamically trim the token-sequence. As the sequence of tokens progresses through the DT layers, its length is reduced by removing token of less importance. In this work, the class-token-based token reduction strategy [28] is used to dynamically reduce the number of tokens. There are two advantages of using class-token based strategy:

- In the proposed FDT working mechanism, we use the class token as the feature vector that represents the full-view FV images. Hence the class token is trained to carry identity-discrimination information, and therefore, its attention value can serve to determine whether to retain a token at a given layer.
- The class-token based strategy requires no additional parameters and introduces extra computational complexity at a negligible cost, which remains crucial for the lightweight design.

The structure of the DeRedun-MHSA is illustrated in the center part of Figure 2(b) and its working mechanism is described as follows:

1) The input tokens are first processed by MHSA and a linear layer. The mathematical expression is shown in Equation 5, where the $\hat{X}$ denotes the resulting intermediate tokens; $MHSA(\cdot)$ and $F_{linear}(\cdot)$ denote the MHSA and linear mapping operation, respectively.

$$\hat{\boldsymbol{X}} = F_{\text{linear}}(MHSA(\boldsymbol{X})) + \boldsymbol{X} \quad (5)$$

2) The attention values between the class token and normal tokens are considered as the importance scores of the tokens. These scores are sorted in descending order.
3) The top K tokens with highest scores are selected as the normal tokens, denoted by $\boldsymbol{X_{Nom}}$; and the tokens with lower scores are considered as redundant tokens, denoted by $\boldsymbol{X_R}$. Thus, the input token sequence results in the normal tokens, redundant tokens, and the class tokens. The mathematical expression is shown in Equation 6, where the Sort_Split$(\cdot)$ denotes the sort and split operation.

$$\boldsymbol{X_{Nom}}, \boldsymbol{X_R}, \boldsymbol{X_{Cls}} = Sort\_Split(\hat{\boldsymbol{X}}) \quad (6)$$

4) The redundant tokens are aggregated into a single token, named aggregation token, $\boldsymbol{X_{Agg}}$. In the aggregation process, the attention values between the class token and each of the redundant tokens are used as the aggregation weights, denoted by $\boldsymbol{W_{att}}$. The mathematical expression is shown as Equation 7.

$$\boldsymbol{X_{Agg}} = \boldsymbol{W_{att}} \times \boldsymbol{X_R} \quad (7)$$

5) Finally, the normal tokens $\boldsymbol{X_{Nom}}$, aggregation token $\boldsymbol{X_{Agg}}$, as well as the class token $\boldsymbol{X_{Cls}}$ are combined to construct the token sequence for next layer.

The effectiveness of the DeRedun-MHSA is studied by the ablation experiments in Section V-G.

E. Locally-Enhanced Feed-Forward Network

In Transformer architectures, the output of the self-attention stages is processed by a feed-forward network (FFN). The purpose of the FFN is further refine the features extracted by the MHSA stages. The MLP based FFN in the original Transformer architecture processes each input token in isolation. Li *et al*. [23] propose a simple and effective Locally-enhanced feed-forward network (LFFN), which mainly consists of token-to-map reshape operation and convolution layer. By contrast, the LFFN processes each token together with its neighboring tokens, thus enhancing the local interaction between neighboring tokens. This enables the LFFN to extract identity-discriminant contextual information from FV images. The effectiveness of this module in FV biometrics is demonstrated in the study of FVT (Finger Vein Transformer) [10]. Therefore, in the proposed FDT, we also adopt LFFN to post-process the intermediate tokens output by DeRedun-MHSA.

In our work, only the normal tokens are processed by the LFFN section of the proposed FDT. class token and aggregation tokens are not processed by the LFFN. The reason is as follows: The class token is associated with an entire sequence of normal tokens, and the aggregation token is used for mitigating redundancy in normal tokens. Therefore, the notion of local information does not apply to class tokens and Aggregation tokens. Hence, in the proposed FDT, tokens of the latter two types are directly transferred to the stage following the LFFN. This is illustrated in the right part of Figure 2(b). The mathematical expression of LFFN is shown as Equation 8. Where the Concatenate$(\cdot)$ and LFFN$(\cdot)$ denote the concatenation and LFFN operation, respectively, and the $\boldsymbol{X_{out}}$ denotes the output of the current layer.

$$\boldsymbol{X_{out}} = Concatenate(\boldsymbol{X_{Agg}}, LFFN(\boldsymbol{X_{Nom}}), \boldsymbol{X_{Cls}}) \quad (8)$$

In Section V-G we describe ablation experiments to study the effectiveness of the LFFN in the proposed FDT.

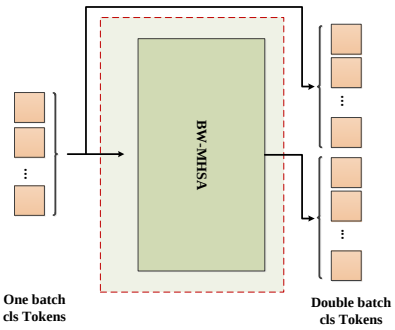


Fig. 4: Architecture of BatchAtten. BW-MHSA denotes the batch-wise MHSA. After the the BatchAtten step, the number of class tokens is doubled, one from the output of BatchAtten, and another from the original class tokens.

F. Batch attention

Inspired by BatchFormer [30], in this work we propose a novel batch attention (BatchAtten) module, aimed at further improving the discrimination ability of the FDT. BatchAtten is used as a training auxiliary, as shown in Figure 2. The BatchAtten module, illustrated in Figure 4, is based on the MHSA mechanism, whose function is to facilitate the interaction among class tokens by attention-fusion.

When training Transformer networks, it is common practice to provide the class tokens as input to the classifier (for the cost-function evaluation). In our case, to train the FDT, we provide the class tokens as input to the BatchAtten module first. The use of the BatchAtten module doubles the number of class tokens provided as input to the classifier, as the BatchAtten module produces the same number of class tokens as the number of original class tokens. The main idea is based on the conjuncture: For biometrics task, the feature extracted should always be distinguishable, even after they interact with other biometrics samples' feature. Here the feature is the class tokens output by the FDT model, and the other biometrics samples are the class tokens of other samples within the training batch.

The mathematical expression of BatchAtten is given by Equation 9. Here, $\boldsymbol{X_{2Cls}}$ denotes the double batch class tokens output of BatchAtten, and Norm$(\cdot)$ denotes the layer normalization operation.

$$\boldsymbol{X_{2Cls}} = Concat(\boldsymbol{X_{Cls}}, \boldsymbol{X_{Cls}} + Norm(MHSA(\boldsymbol{X_{Cls}}))) \quad (9)$$

Since the proposed BatchAtten is applied batch-wise, the additional computational load (in the training phase) of the BatchAtten module is negligibly small, as the module uses only a single MHSA layer. In the inference phase, both the auxiliary and classifier are removed, and the class token output by the last FVCDT layer is used as the feature vector to represent the input full-view FV images.

In Section V-G we present ablation experiments to analyze the performance of the proposed BatchAtten module.

V. Experiments

In this section we present the experiments designed to evaluate the performance of the proposed FDT for full-view FVA. First we describe two publicly available full-view FV datasets, as well as the corresponding experimental protocols used in this work. Then, the metrics and criteria for evaluating full-view FVA performance are outlined. We also describe the software setup for reproducing our experimental results.

We present five sets of experiments. First, we conduct redundancy-analysis experiments for both single-view and full-view FV data. Then we evaluate the performance of the proposed FDT. We also compare the performance of the FDT to that of other relevant, reproducible works published previously. This is followed by a set of ablation experiments to study the utility of key modules in the FDT. Finally, we present an analysis of the computational complexity of the proposed FDT, which provides some intuition for practical applications.

A. Full-View FV Datasets and Experimental Protocols

We have used two full-view FV datasets, namely, MFFV and LFMB-3DFB, in this work:

1) MFFV [1] includes FV data from 320 fingers, each captured in 3 orientation poses (10 presentations per orientation), 6 illumination levels, and 3 camera views, totaling 540 images per finger. The orientations include normal pose, 60° clockwise, and 60° anticlockwise rotations. In this work we have used the MFFV-N dataset, which is a subset of the full MFFV dataset, consisting of only the normal presentations of each finger. We follow the two experimental protocols provided with the dataset, balanced protocol and normal protocol, as shown in the upper part of Table I.

2) LFMB-3DFB [17] includes multi-biometrics data (FV and finger-skin), corresponding to 695 fingers. For each finger this dataset includes images from six views (named A – F). Since three cameras evenly spaced around the finger are sufficient to capture the full view FV [1, 4], we use FV images of three views A, C, and E in this study. In this way, we make the full-view FV data from this dataset compatible with the FDT proposed in this work. We follow the protocol provided with the dataset, as shown in the lower part of Table I.

The main results of this study are reported on the MFFV-N dataset. That is, we report the results of all experiments on the MFFV-N dataset. The LFMB-3DFB dataset is used only for evaluating the cross-domain generalization of the proposed FDT. In FVA experiments, we train the FDT model on the training partition. Then, we use the development partition with the balanced protocol to select the optimal model, and the score-threshold corresponding to the selected criterion. Finally, we apply the selected optimal model, with the selected score threshold, to the test partition, to compute the final FVA performance of the FDT.

B. Evaluation Metrics and Criteria

In our experiments we use several metrics and criteria to comprehensively evaluate the FVA performance of the proposed FDT. The metrics and criteria used in this work, as well as their functions are briefly introduced below.

1) Metrics: We have used the following five metrics: false match rate (FMR), false non match rate (FNMR), half total error rate (HTER), equal error rate (EER), and true match rate (TMR). FMR indicates the proportion of impostor matches that

TABLE I: Experimental protocols. There are two protocols in our experiments, the balanced protocol and the normal protocol.

Dataset	Num	Balanced protocol			Normal protocol		
		Training set	Development set	Test set	Training set	Development set	Test set
MFFV-N	Num of fingers	160	64	96	160	64	96
	Num of genuines	/	2880	4320	/	576	864
	Num of imposters	/	2880	4320	/	36288	82080
	Num of total comparison	/	5760	8640	/	36864	87944
LFMB-3DFB	Num of fingers	347	139	209	/	/	/
	Num of genuines	/	6255	9405	/	/	/
	Num of imposters	/	6255	9405	/	/	/
	Num of total comparison	/	12510	18810	/	/	/

Note: these protocols are provided with the datasets. For MFFV-N, there are balanced and normal protocols. For LFMB-3DFB, there is one protocol, balanced protocol.

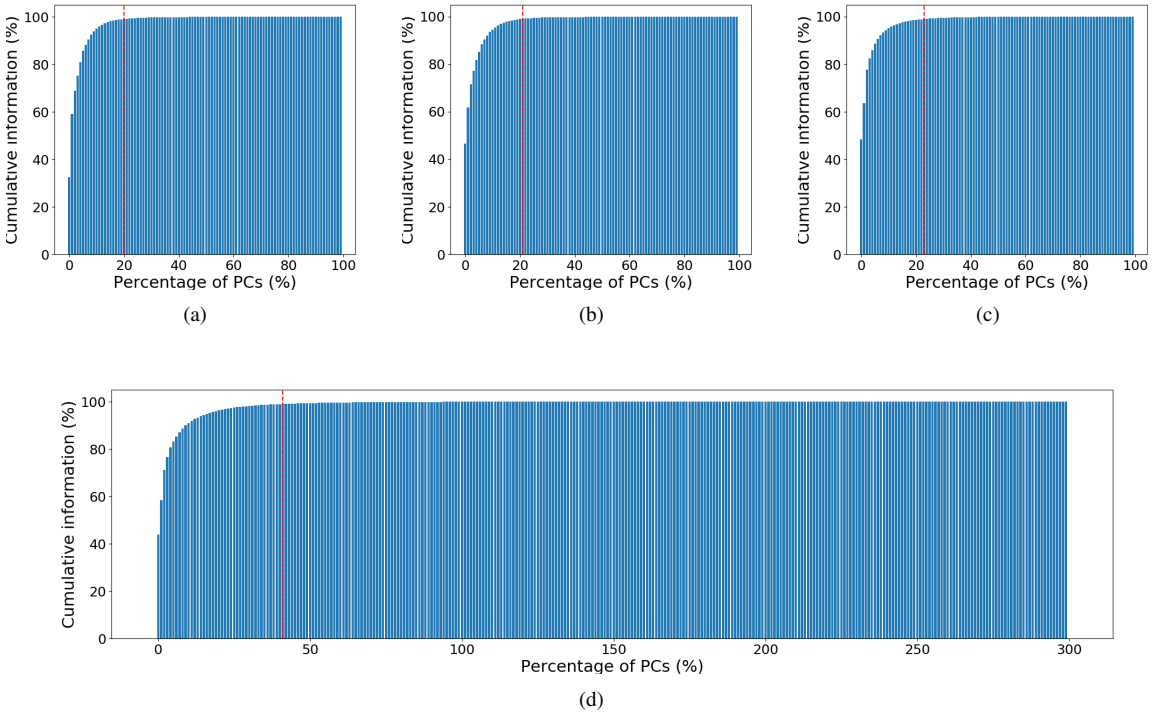


Fig. 5: Cumulative information distribution of FV image components. The red dotted line indicates the location where the cumulative information reaches 99%. (a) View 1; (b) View 2; (3) View 3; (4) Full-view FV images.

are incorrectly accepted by the system, whereas FNMR denotes the proportion of genuine matches incorrectly rejected by the system. HTER is the average of FMR and FNMR. EER is the error rate corresponding to the operating point (score-threshold over the development set) where FMR and FNMR are equal. TMR is indicates the true matching rate at a fixed FMR.

2) Criteria: Three criteria are used in this work: minimum HTER (min-HTER) criterion, EER criterion, and fixed-FMR. The min-HTER criterion is used to compare experiments with different parameters, to find the parameters that lead to the lowest misclassification error. The EER criterion is used in scenarios when both FMR an FNMR are considered to be equally important. In most real-world applications, false-matches are considered to be far more important than false non-matches. In such scenarios, the biometrics system is tuned to limit the FMR to a certain acceptable value, typically 1% or even 0.1%. The fixed-FMR criterion is useful for evaluating the biometrics system in this scenario.

C. Experimental Environment and Setup

We have conducted the experiments using the PyTorch framework. The models are trained and tested on the NVIDIA RTX 3090 GPU; the maximum number of training epoch is set to 5000. During the last 3000 epochs of training, the model is evaluated on the development set with balanced protocol in the later 3000 epochs, where the model with the lowest EER is selected as the best performing model. The learning rate is set to 0.001. The cosine annealing schedule with the maximum of interactions set to 25 is used as the learning rate change scheduler. The batch size is set to 32. The weights of cross-entropy loss and center loss are set to 1 and 0.01, respectively. Note that there is a separate learning rate for updating the centers in the center-loss function, which is set to 0.001. The stochastic gradient descent is used as the optimizer, where the momentum and weight-decay parameters are set to 0.9 and 0.01, respectively. We have used the data augmentation, including random color jitter with the coefficient of brightness, contrast, saturation, and hue all set to 0.2; random translation

transform with both the coefficient of the horizontal and vertical are set to 0.2; and random perspective with the coefficient of scale is set to 0.8-1.2.

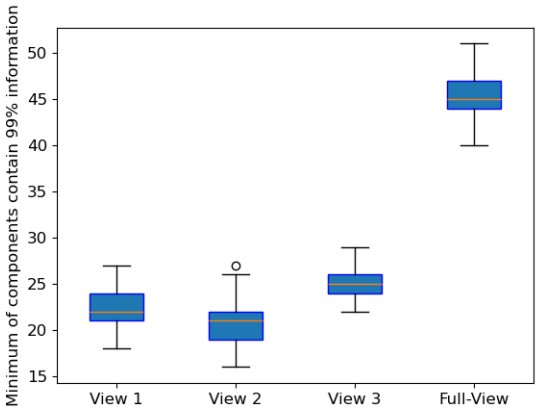


Fig. 6: Redundancy analysis results.

D. Redundancy Analysis Experiments

In this section, we present the redundancy analysis experiment for both single-view and full-view FV images. To the best of our knowledge, this is the first experiment for FV image redundancy analysis.

The experiment is conducted using the MFFV-N dataset. First, we perform the FV image redundancy analysis for each view independently. As described in Section III, the given single-view FV image, of size $320 \times 240 \times 1$, is divided into 100 non-overlapping patches of size $32 \times 24 \times 1$ each. Each patch is flattened, that is, represented by a 1D vector of 768 elements. Thus, each single-view FV image is represented by matrix of shape 100×768. We perform PCA on this matrix, and retain the top N principal components such that 99% of the original information is conserved. Figure 5(a) - 5(c) show the cumulative information distributions of one full-view sample (three views). From these plots we can see that the cumulative information of a given FV image is close to 100% even when N (the number of retained principal components) is relatively small (20%-30%) of the total number of PCs (which is 768)). **This result indicates that there is significant redundancy in FV images, up to 70%-80%.**

To perform redundancy analysis of FV images in the full-view scenario, we combine the components of all the three views, totally 300 components. Then all the components are fed into the PCA process as for single-view. Figure 5(d) shows the cumulative information distribution for PCs of a given set of full-view FV images. From this plot we can see that the ratio of components when cumulative information is obviously less than the single-view's – the ratio is less than 20%. This result is consistent with our conjecture, in Section I that **the redundancy issue is more serious in full-view FV images.**

We have performed a statistical analysis of the redundancy in FV images using the training partition of the MFFV-N dataset. Specifically, the redundancy-analysis procedure is applied to the first FV sample recorded for each finger in the training partition, that is, to a total of 480 FV images (160 FV images per view). For each finger we record the number of PCs required to retain 99% of the original image-information in both, single-view and full-view scenarios. In Figure 6 we show box-plots summarizing the distribution of the number of PCs retained, for each single-view scenario as well as for the full-view scenario. From this figure we see that for single-view scenarios (View 1, View 2 and View 3) a minimum of 20-30% of the total number of PCs is retained. This corresponds to a redundancy in the range of 70-80%. Among these three views, View 3 requires the overall most minimum of components that contain 99% information compared to View 1 and View 2, indicating that the redundancy of View 3 is lower than other two views. This phenomenon might be due to the fact that the finger palmar side contains richer veins, which is consist with the experimental results of work [1]. **For the full-view scenario we see that the minimum number of PCs retained falls in the range of 13-17%, corresponding to a redundancy of 83-87%.**

E. FDT Configuration Based on The FVRA Result

The architecture configuration of the proposed FDT model is based on the redundancy analysis results. During the feature extraction process, the FDT model gradually trims the unimportant tokens, so as to achieve de-redundant FV feature extraction. The configuration details are as explained in follows:

Firstly, we set the number of normal tokens to be retained in the FDT output layer (the last FVCDT module) to 49. This setting is guided by the redundancy-analysis results (Section V-D), and fulfills two important conditions: (a) it falls within the range (40 – 50) of number of PCs typically retained in full-view FV redundancy analysis, and (b) it is an integer result of a square operation, which is convenient for the sequence-to-map operation (in the LFFN and PEG modules) in FDT.

Secondly, we also set the number of retained normal tokens in the last SVDT module (Stage 4 in Figure 2(a) to 49. This choice is guided by two requirements: (a) the number should have an integer square-root, and (b) the number should be approximately twice the minimum number of tokens necessary, according to the preceding redundancy analysis. We elaborate further on the second requirement. In the single-view scenario, we have seen that 20-30 tokens may be sufficient to represent the identity-discriminant information available in an FV image. The SVDT modules form the intermediate stages of the FDT, where we cannot guarantee that the identity-discriminant information is most efficiently represented. In other words, to encapusulate all the identity information available in the input FV image, at this stage we may need more tokens than the absolute minumum necessary according to the redundancy-analysis. By setting the number of retained normal tokens to approximately twice the number indicated by the preceding redundancy-analysis, we aim to ensure that important identity information is not inadvertently excluded at the intermediate stages.

Thirdly, for the same two reasons mentioned previously, we set the number of retained normal tokens in the SVDT modules in the first stage at 100. This implies that the SVDT modules

TABLE II: Architecture configuration of FDT.

Stage	Retained normal tokens			
	SVDT		FVCDT	
	Num	Reduction ratio	Num	Reduction ratio
Stage 1	100×3	0%	144	52%
Stage 2	81×3	19%	121	59.67%
Stage 3	64×3	36%	100	66.67%
Stage 4	49×3	51%	81	73%
Stage 5	/		64	78.67%
Stage 6 (Final)	/		**49**	**83.67%**

in the first stage do not reduce the number of normal tokens at all. Similarly, the number of tokens retained after the first FVCDT module is set to 144, which represents a very small reduction compared to the number of input normal tokens.

Fourthly, the number of normal tokens retained in the rest intermediate SVDT and FVCDT modules are gradually reduced from the first stage to the last stage, in accordance with the square-root principle.

The configuration of the proposed FDT, based on the factors described above, is shown in Table II. **As indicated in the table, the model achieves a high overall reduction of 83.67% in the full-view scenario.** At the individual view level, the SVDT stages contribute to this by achieving a reduction of up to 51% for normal tokens per view.

There are two advantages of token reduction in the proposed FDT:

1) Token reduction is beneficial to reduce the model complexity, *i.e.* computational load;
2) Trimming the low information tokens also reduces noise interference.

We evaluate these two aspects through ablation experiments and the application complexity experiment, presented later.

F. Authentication Experiments of the Proposed FDT

In this section we present the FVA experiments using the proposed FDT as well as the baseline methods. For comparisons, the MVCNN [31] and the MVT [32] are adopted as the baseline methods in our experiments, we also adopt the baseline method (MC+MM+SVM) results in work [1]. As neither the MVCNN nor the MVT are designed specifically for FV biometrics, they are hard to be trained in this task. Therefore, we have modified both models for FVA applications, based on some straightforward but effective principles. The MVCNN is modified by replacing the original VGG backbone by ResNet [33]. In the MVT, the original parameter-shared local Transformer module is replaced by a parameter-independent local Transformer.

Figure 7 shows the receiver operator characteristic (ROC) curves of the FVA experiments. To comprehensively evaluate and compare the performances of the proposed FDT and the comparative methods in all application scenario, both balanced and normal protocols and all the three kinds of criteria are used: EER criterion, min-HTER criterion, and fixed FMR criterion at FMR=1%, and FMR=0.1%. Table III shows the FVA results, and the best values of each case is denoted in bold font. **From this table we can see that the proposed FDT achieves the overall best performances on full-view FV authentication task** – most of the best results (among the different experiments) are produced by the FDT. Even in the experiments where the FDT does not produce the best results, its performance is very close to the best. We highlight here some key results.

1) Using the EER criterion, FDT achieves 0.97% EER on development set and 1.84% HTER on test set, under the balanced protocol; whereas under the normal protocol it achieves 2.22% EER on development set and 3.04% HTER on test set. Considering the min-HTER criterion, FDT achieves 0.94% HTER on development set and 1.85% on test set, under the balanced protocol; and HTERs of 2.05% and 2.92% on the development set and the test set, respectively, under the normal protocol. **Both the EER and min-HTER criteria evaluate the biometrics system's performance in the standard evaluation scenario – that the determinant threshold of a biometrics system is selected when the two error rates, FMR and FNMR, are equal, or their mean is smallest. These results demonstrate that the proposed FDT is a very effective tool for full-view FVA.**

2) Using the FMR=1% criterion, the FDT achieves 98.96% TMR on development set and 1.37% FMR on test set, under the balanced protocol; and 95.49% TMR on development set and 0.98% FMR on test set, under the normal protocol. Considering the FMR=0.1% criterion, FDT achieves 92.62% TMR on development set and 0.16% FMR on test set, under the normal protocol; whereas it achieves 75.52% TMR on the development set and 0.13% FMR on the test set, under the normal protocol. The fixed FMR criteria evaluate the performance of the biometrics system in application scenarios where FMR is constrained to be below a specified threshold. The TMR metrics indicate that the proposed FDT achieves high pass rate for genuine FV samples under the predefined-FMR constraint. **In both protocols, the FMR values for the test-sets are quite close to the FMR limits set on the corresponding development sets. These results demonstrate that the FDT generalizes well from the development set to the test set in both protocols.**

G. Ablation Experiments for the Proposed FDT

In this section, we present the ablation experiments for the proposed FDT. The performances of the various ablated models of the FDT are shown in Table IV. As introduced in Section IV, there are five key modules in the proposed FDT, namely PEG, MLP-P, LFFN, DeRedun-MHSA, and BatchAtten. We ablate these modules one by one to evaluate their effectiveness. For the ablation of PEG, it is simply removed from the FDT. For the ablation of MLP-P, since the patch-to-token embedding is necessary in Transformer architecture, it is replaced by the original linear method. For the ablation of LFFN, since the FFN model is one of the two key modules of Transformer, it is replaced by the original FFN method. For the ablation of DeRedun-MHSA, since the MHSA is also one of the two key modules of Transformer, it is also replaced by the original MHSA method. For the ablation of BatchAtten, it is removed from the training framework of the FDT.

Table IV shows the evaluation results of the ablation experiments. The performances without ablation are also

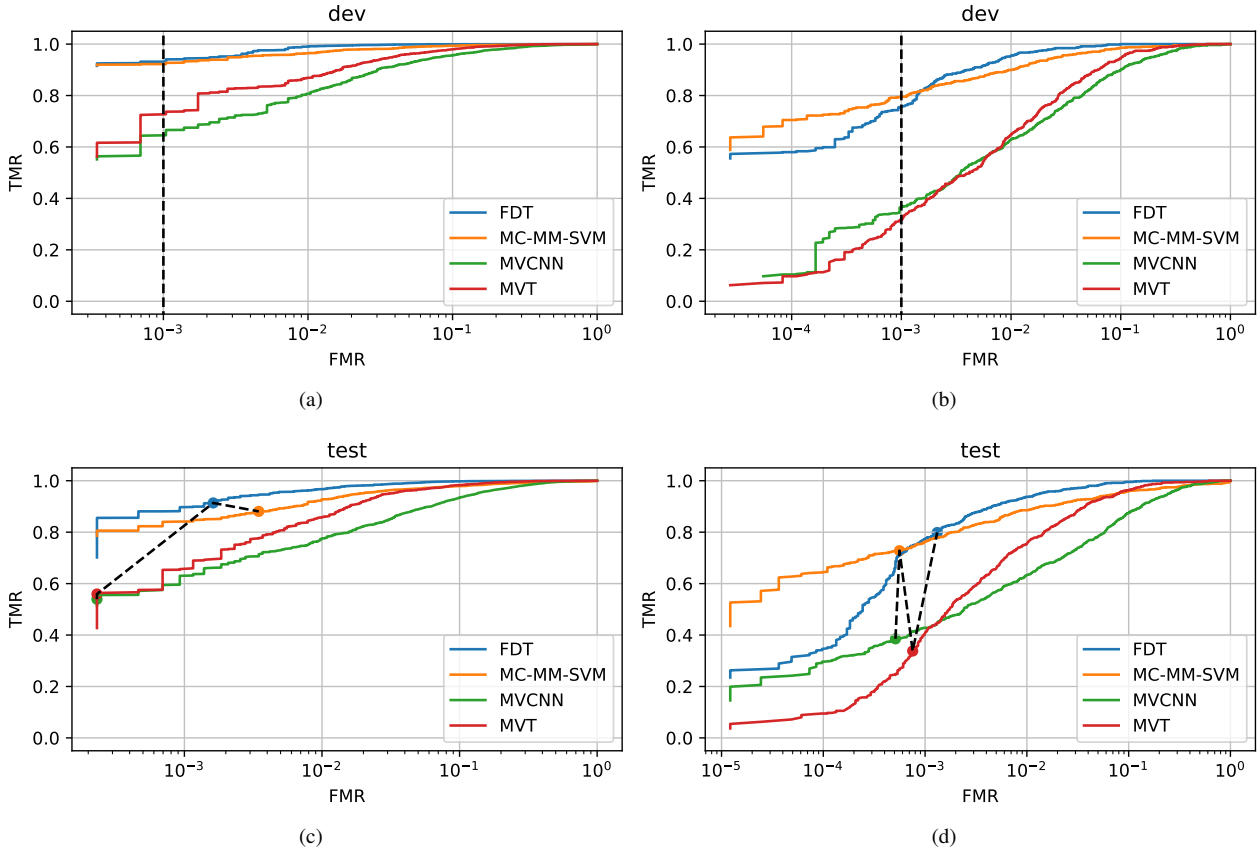


Fig. 7: ROC plots for FVA experiments using the proposed FDT and other baseline methods. (a) ROC plots for balanced protocol on development set; (b) ROC plots for normal protocol on development set; (c) ROC plots for balanced protocol on test set; (d) ROC plots for normal protocol on test set.

TABLE III: FVA experiments for the proposed FDT and the baseline methods. The best results are indicated in bold font.

Criterion	Method	Balanced protocol							Normal protocol						
		Development set				Test set			Development set				Test set		
		EER	FMR	FNMR	HTER	FMR	FNMR	HTER	EER	FMR	FNMR	HTER	FMR	FNMR	HTER
EER	MC+MM+SVM	2.08	2.08	2.08	2.08	2.29	4.51	3.40	3.97	3.94	3.99	3.97	2.32	8.56	5.44
	MVCNN	6.15	6.15	6.15	6.15	6.64	9.42	8.03	10.07	10.07	10.07	10.07	9.67	12.62	11.14
	MVT	4.27	4.27	4.27	4.27	3.08	4.98	4.03	7.63	7.62	7.64	7.63	5.48	6.94	6.21
	FDT(proposed)	**0.97**	**0.97**	**0.97**	**0.97**	**1.39**	**2.29**	**1.84**	**2.22**	**2.18**	**2.26**	**2.22**	**2.02**	**4.05**	**3.03**
		EER	FMR	FNMR	HTER	FMR	FNMR	HTER	EER	FMR	FNMR	HTER	FMR	FNMR	HTER
min-HTER	MC+MM+SVM	/	1.77	2.19	1.98	2.08	4.65	3.37	/	2.97	4.34	3.66	**1.67**	9.72	5.69
	MVCNN	/	4.27	7.88	6.08	4.70	11.55	8.13	/	8.56	11.11	9.83	8.15	14.93	11.54
	MVT	/	4.27	4.24	4.25	3.15	4.88	4.02	/	7.80	6.94	7.37	5.61	6.60	6.10
	FDT(proposed)	/	**1.01**	**0.87**	**0.94**	**1.46**	**2.25**	**1.85**	/	**2.53**	**1.56**	**2.05**	2.25	**3.59**	**2.92**
		TMR	FMR	FNMR	HTER	FMR	FNMR	HTER	TMR	FMR	FNMR	HTER	FMR	FNMR	HTER
FMR=1%	MC+MM+SVM	96.39	0.97	3.61	2.29	1.34	6.25	3.80	89.93	1.00	10.07	5.53	0.49	15.74	8.12
	MVCNN	80.35	0.97	19.65	10.31	1.00	22.55	11.77	62.85	1.00	37.15	19.08	0.83	38.43	19.63
	MVT	86.70	0.97	13.30	7.14	**0.79**	15.53	8.16	64.76	1.00	35.24	18.12	0.63	29.40	15.02
	FDT(proposed)	**98.96**	0.97	**1.04**	**1.01**	1.37	**2.38**	**1.88**	**95.49**	1.00	**4.51**	**2.76**	0.98	**6.25**	**3.61**
		TMR	FMR	FNMR	HTER	FMR	FNMR	HTER	TMR	FMR	FNMR	HTER	FMR	FNMR	HTER
FMR=0.1%	MC+MM+SVM	92.05	0.07	7.95	4.01	0.35	11.92	6.13	**79.51**	0.10	**20.49**	**10.29**	0.06	27.20	13.63
	MVCNN	56.60	0.07	43.40	21.74	0.02	46.06	23.04	36.46	0.10	63.54	31.82	0.05	61.46	30.75
	MVT	61.77	0.07	38.23	19.15	0.02	44.05	22.04	31.77	0.10	68.23	34.16	0.08	66.20	33.14
	FDT(proposed)	**92.60**	0.07	**7.40**	**3.73**	**0.16**	**8.63**	**4.40**	75.52	0.10	24.48	12.29	0.13	**20.14**	**10.14**

*The results of MC+MM+SVM method are reproduced from work [1].

TABLE IV: Ablation experiments for the proposed FDT and the comparative methods. The best results are indicated in bold font.

Criterion	Ablation	Balanced protocol							Normal protocol						
		Development set				Test set			Development set				Test set		
		EER	FMR	FNMR	HTER	FMR	FNMR	HTER	EER	FMR	FNMR	HTER	FMR	FNMR	HTER
EER	PEG	1.35	1.35	1.35	1.35	**0.97**	3.61	2.29	2.81	2.84	2.78	2.81	1.73	7.18	4.45
	MLP-P	1.74	1.74	1.74	1.74	1.50	2.13	**1.82**	4.34	4.34	4.34	4.34	2.79	4.40	3.59
	LFFN	3.65	3.65	3.65	3.65	1.99	8.56	5.28	5.76	5.78	5.73	5.76	2.46	14.58	8.52
	DeRedun-MHSA	1.70	1.70	1.70	1.70	2.57	2.27	2.42	3.48	3.49	3.47	3.48	3.25	5.79	4.52
	BatchAtten	1.08	1.08	1.08	1.08	1.39	2.48	1.93	2.26	2.26	**2.26**	2.26	**1.63**	5.44	3.54
	No Ablation(FDT)	**0.97**	**0.97**	**0.97**	**0.97**	1.39	**2.29**	1.84	**2.22**	**2.18**	**2.26**	**2.22**	2.02	**4.05**	**3.03**
		EER	FMR	FNMR	HTER	FMR	FNMR	HTER	EER	FMR	FNMR	HTER	FMR	FNMR	HTER
min-HTER	PEG	/	1.22	1.35	1.28	**0.95**	3.68	2.31	/	2.92	2.43	2.67	1.77	6.94	4.36
	MLP-P	/	1.28	2.15	1.72	1.18	2.62	1.90	/	2.70	5.38	4.04	1.81	5.56	3.68
	LFFN	/	3.40	3.82	3.61	1.78	8.84	5.31	/	4.52	6.94	5.73	1.90	17.25	9.57
	DeRedun-MHSA	/	1.49	1.91	1.70	2.27	2.38	2.33	/	3.81	3.13	3.47	3.56	4.98	4.27
	BatchAtten	/	1.28	**0.80**	1.04	1.60	2.29	1.94	/	**2.01**	2.43	2.22	**1.50**	5.67	3.59
	No Ablation(FDT)	/	**1.01**	0.87	**0.94**	1.46	**2.25**	**1.85**	/	2.53	**1.56**	**2.05**	2.25	**3.5**9	**2.92**
		TMR	FMR	FNMR	HTER	FMR	FNMR	HTER	TMR	FMR	FNMR	HTER	FMR	FNMR	HTER
FMR=1%	PEG	98.02	0.97	1.98	1.48	**0.83**	4.77	2.80	94.10	1.00	5.90	3.45	**0.67**	13.43	7.05
	MLP-P	97.29	0.97	2.71	1.84	0.97	3.10	2.04	90.45	1.00	9.55	5.27	0.76	9.61	5.18
	LFFN	91.49	0.97	8.51	4.74	0.65	16.30	8.47	76.74	1.00	23.26	12.13	0.41	38.43	19.42
	DeRedun-MHSA	96.56	0.97	3.44	2.20	1.48	3.40	2.44	86.81	1.00	13.19	7.10	1.10	11.00	6.05
	BatchAtten	98.40	0.97	1.60	1.28	1.09	2.94	2.01	92.53	1.00	7.47	4.23	0.85	8.68	4.77
	No Ablation(FDT)	**98.96**	0.97	**1.04**	**1.01**	1.37	**2.38**	**1.88**	**95.49**	1.00	**4.51**	**2.76**	0.98	**6.25**	**3.61**
		TMR	FMR	FNMR	HTER	FMR	FNMR	HTER	TMR	FMR	FNMR	HTER	FMR	FNMR	HTER
FMR=0.1%	PEG	91.84	0.07	8.16	4.11	0.12	13.08	6.60	**79.17**	0.10	**20.83**	**10.47**	**0.13**	27.31	13.72
	MLP-P	89.76	0.07	10.24	5.16	0.14	12.55	6.34	71.18	0.10	28.82	14.46	0.12	29.28	14.70
	LFFN	55.80	0.07	44.20	22.14	**0.00**	56.34	28.17	40.45	0.10	59.55	29.82	0.04	70.72	35.38
	DeRedun-MHSA	84.31	0.07	15.69	7.88	0.07	17.20	8.63	72.57	0.10	27.43	13.76	0.21	25.00	12.61
	BatchAtten	85.24	0.07	14.76	7.41	0.07	18.17	9.12	75.69	0.10	24.31	12.20	0.14	26.39	13.26
	No Ablation(FDT)	**92.60**	0.07	**7.40**	**3.73**	0.16	**8.63**	**4.40**	75.52	0.10	24.48	12.29	**0.13**	**20.14**	**10.14**

included in this table for intuitive comparison, and the best value in each case is denoted in bold font. From this table we can make the following observations:

1) In most cases the (full, non-ablated) FDT performs better than the ablated models. From these results we can conclude that all the key modules contribute to the efficacy of the proposed FDT model;
2) In a few cases the ablated version of the FDT achieves slightly lower FMR or FNMR than the full FDT. We can leverage this phenomenon by adjusting the FDT's structure to optimize its performance for specific application scenarios. For example, in applications requiring very low FMR, it may be useful to remove the PEG module.

H. Cross Dataset FVA Experiments Using FDT

In this section we discuss the results of using the proposed FDT model (trained only using MFFV-N data) for FVA on the LFMB-3DFB dataset. Our purpose here is to study the generalization properties of the FDT in cross-dataset scenarios.

FV images of the three views A, C, and E of the LFMB-3DFB dataset are used to construct the 3D input sample required for the proposed FDT. The MC+MM+SVM method [1]) is taken as the baseline. Experimental results are shown in Table V. From these results we can draw the following three conclusions:

1) The performance of the proposed FDT is much better than the baseline method, thus demonstrating its cross-dataset generalization, *e.g*. the proposed FDT achieves an EER of 7.24% and an HTER of 7.34% under the balanced protocol and EER criterion.
2) The MC+MM+SVM method performs poorly on the LFMB-3DFB dataset. This can be explained by the following analysis: in the LFMB-3DFB dataset the finger-orientation was not controlled during data collection (as described in [1, 17]). Therefore, for a given view (say, A), we are not guaranteed that every presentation of the same finger will result in the same vein pattern. In other words, there is high intra-class variation in this dataset for individual views.
3) This result indicates that there is still room for developing orientation-invariant methods for FV biometrics, and also indicates that there still room for cross-dataset generalization of our proposed method.

TABLE V: Cross-domain Experiments on LFMB-3DFB (%).

Criterion	Method	Balanced protocol						
		Development set				Test set		
EER		EER	FMR	FNMR	HTER	FMR	FNMR	HTER
	MC+MM+SVM	35.38	35.38	35.38	35.38	32.93	35.23	34.08
	FDT (proposed)	**7.24**	**7.24**	**7.24**	**7.24**	**7.09**	**7.23**	**7.37**
min-HTER		EER	FMR	FNMR	HTER	FMR	FNMR	HTER
	MC+MM+SVM	/	13.61	52.69	33.15	13.04	53.33	33.18
	FDT (proposed)	/	**6.19**	**8.09**	**7.14**	**6.20**	**8.46**	**7.33**
FMR=1%		1-FNMR	FMR	FNMR	HTER	FMR	FNMR	HTER
	MC+MM+SVM	25.15	0.99	74.85	37.92	**0.72**	74.75	37.74
	FDT (proposed)	**71.37**	0.99	**28.63**	**14.81**	1.02	**28.59**	**14.81**
FMR=0.1%		1-FNMR	FMR	FNMR	HTER	FMR	FNMR	HTER
	MC+MM+SVM	18.10	0.10	81.90	41.00	0.07	81.81	40.94
	FDT (proposed)	**36.28**	0.10	**63.73**	**31.91**	**0.06**	**64.23**	**32.15**

*The results of MC+MM+SVM method are reproduced from work [1].

I. Complexity Experiments

We have also conducted the complexity experiments, the metrics including parameters (Params), floating-point operations (FLOPs), and the saved template dimension. Results of the complexity-analysis are shown in Table VI, from which we can see that the proposed FDT model and its sub-models (ablation models) have the relatively small Params (11.11M), FLOPs (0.87G), and template dimension (128). These complexity

TABLE VI: Complexity Experiments

Model	Params	FLOPs	Template dimension
MVCNN	21.28M	16.72G	512
MVT	50.92M	6.53G	384
FDT without LFFN	9.69M	0.85G	128
FDT without MLP-P	10.37M	0.80M	128
FDT without PEG	8.45M	0.62G	128
FDT without DeRedun-MHSA	11.11M	1.59G	128
FDT without BA	11.11M	0.87G	128
FDT	11.11M	0.87G	128

metrics are considerably lower than those of the comparator models, highlighting the lightweight advantage of the proposed FDT model. Compared to the ablated models, the Params and FLOPs of the full (non-ablated) FDT are only slightly higher. These results show that the proposed FDT not only achieves SOTA performance, but also application potential.

VI. Conclusions and future works

Full-view FV authentication (FVA) systems offer higher accuracy than single-view FVA systems and it is much more difficult to be constructed presentation attacks. However, Full-view FVA systems present two main challenges: increased computational loads, and low performance-to-cost ratio. Both these challenges stem from the high degree of redundancy in FV images. In this work we have proposed a dynamic transformer model, named Finger-Vein Dynamic Transformer (FDT), that addresses the core issue of redundancy in FV samples. Successive stages of the FDT gradually trim redundant information in the input FV samples (full-view or single-view). To guide this process, we initially developed the FV Redundancy Analysis (FVRA) method to conduct a redundancy analysis, which estimates redundancy in both single-view and full-view FV samples. To the best of our knowledge, this is the first work to conduct redundancy analysis in FV biometrics, and to use it to design a FV feature extraction method.

We have presented a set of rigorous experiments to comprehensively evaluate the performance of the proposed FDT and FVRA methods. Specifically, the proposed FDT achieves SOTA full-view FV authentication performance: on the MFFV-N dataset [1]: 0.97% EER on development set and 1.84% HTER on test set, under the balanced protocol. The FDT also exhibits better cross-dataset generalization on the LFMB-3DFB dataset [17] than the baseline method.

In future works, we plan to further develop FV biometrics mainly on the following two aspects: 1) cross-view FV biometrics and orientation invariant full-view FV biometrics; and 2) more lightweight and higher performance feature extraction models.

References

[1] Junduan Huang, Zifeng Li, Sushil Bhattacharjee, Sébastien Marcel, and Wenxiong Kang. Mirror-based full-view finger vein authentication with illumination adaptation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(3):2057–2073, 2025.

[2] Andreas Uhl, Christoph Busch, Sébastien Marcel, and Raymond Veldhuis. *Handbook of vascular biometrics*. Springer Nature, 2020.

[3] Junduan Huang, An Zheng, M. Saad Shakeel, Weili Yang, and Wenxiong Kang. Fvfsnet: Frequency-spatial coupling network for finger vein authentication. *IEEE Transactions on Information Forensics and Security*, 18:1322–1334, 2023.

[4] Wenxiong Kang, Hongda Liu, Wei Luo, and Feiqi Deng. Study of a full-view 3d finger vein verification technique. *IEEE Transactions on Information Forensics and Security*, 15:1175–1189, 2020.

[5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

[6] Rig Das, Emanuela Piciucco, Emanuele Maiorana, and Patrizio Campisi. Convolutional neural network for finger-vein-based biometric identification. *IEEE Transactions on Information Forensics and Security*, 14:360–373, 2019.

[7] Lu Yang, Xingbo Liu, Gongping Yang, Jun Wang, and Yilong Yin. Small-area finger vein recognition. *IEEE Transactions on Information Forensics and Security*, 18:1914–1925, 2023.

[8] Ridvan Salih Kuzu, Emanuela Piciucco, Emanuele Maiorana, and Patrizio Campisi. On-the-fly finger-vein-based biometric recognition using deep neural networks. *IEEE Transactions on information Forensics and Security*, 15:2641–2654, 2020.

[9] Shuyi Li, Ruijun Ma, Lunke Fei, and Bob Zhang. Learning compact multirepresentation feature descriptor for finger-vein recognition. *IEEE Transactions on Information Forensics and Security*, 17:1946–1958, 2022.

[10] Junduan Huang, Weijian Luo, Weili Yang, An Zheng, Feng-Zhao Lian, and Wenxiong Kang. Fvt: Finger vein transformer for authentication. *IEEE Transactions on Instrumentation and Measurement*, 71:1–13, 2022.

[11] Szilvia Bunda. 3d point cloud reconstruction based on the finger vascular pattern. 2018.

[12] Sjoerd Rozendal. Redesign of a finger vein scanner. In *University of Twente*, 2018.

[13] Pengyang Zhao, Shuping Zhao, Luyang Chen, Wenming Yang, and Qingmin Liao. Exploiting multiperspective driven hierarchical content-aware network for finger vein verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 32: 7938–7950, 2022.

[14] Bernhard Prommegger, Christof Kauba, and Andreas Uhl. Multi-perspective finger-vein biometrics. *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, pages 1–9, 2018.

[15] Huafeng Qin, Rongshan Hu, Mounîm A. El-Yacoubi, Yantao Li, and Xinbo Gao. Local attention transformer-based full-view finger-vein identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[16] Yizhuo Song, Pengyang Zhao, Siqi Wang, Qingmin Liao, and Wenming Yang. Study of 3d finger vein biometrics on imaging device design and multi-view verification. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[17] Weili Yang, Zhuoming Chen, Junduan Huang, and Wenxiong Kang. A novel system and experimental study for 3d finger multibiometrics. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 4:471–485, 2022.

[18] Christof Kauba, Martin Drahanský, Marie Nováková, Andreas Uhl, and Stepán Rydlo. Three-dimensional finger vein recognition: A novel mirror-based imaging device. *Journal of Imaging*, 8, 2022.

[19] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[20] Kai Han, Yunhe Wang, Hanting Chen, Xinghao Chen, Jianyuan Guo, Zhenhua Liu, Yehui Tang, An Xiao, Chunjing Xu, Yixing Xu, et al. A survey on vision transformer. *IEEE transactions on pattern analysis and machine intelligence*, 45(1):87–110, 2022.

[21] Salman Khan, Muzammal Naseer, Munawar Hayat, Syed Waqas

Zamir, Fahad Shahbaz Khan, and Mubarak Shah. Transformers in vision: A survey. *ACM computing surveys (CSUR)*, 54(10s): 1–41, 2022.

[22] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 9992–10002, 2021.

[23] Yawei Li, K. Zhang, Jie Cao, Radu Timofte, and Luc Van Gool. Localvit: Bringing locality to vision transformers. *ArXiv*, abs/2104.05707, 2021.

[24] Xiangxiang Chu, Zhi Tian, Bo Zhang, Xinlong Wang, Xiaolin Wei, Huaxia Xia, and Chunhua Shen. Conditional positional encodings for vision transformers. *arXiv preprint arXiv:2102.10882*, 2021.

[25] Yongming Rao, Wenliang Zhao, Benlin Liu, Jiwen Lu, Jie Zhou, and Cho-Jui Hsieh. Dynamicvit: Efficient vision transformers with dynamic token sparsification. In *Neural Information Processing Systems*, 2021.

[26] Lingchen Meng, Hengduo Li, Bor-Chun Chen, Shiyi Lan, Zuxuan Wu, Yu-Gang Jiang, and Ser Nam Lim. Adavit: Adaptive vision transformers for efficient image recognition. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 12299–12308, 2021.

[27] Mohsen Fayyaz, Soroush Abbasi Koohpayegani, Farnoush Rezaei Jafari, Sunando Sengupta, Hamid Reza Vaezi Joze, Eric Sommerlade, Hamed Pirsiavash, and Juergen Gall. Adaptive token sampling for efficient vision transformers. In *European Conference on Computer Vision*, 2021.

[28] Youwei Liang, Chongjian Ge, Zhan Tong, Yibing Song, Jue Wang, and Pengtao Xie. Evit: Expediting vision transformers via token reorganizations. In *International Conference on Learning Representations*, 2022.

[29] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, 2016.

[30] Zhi Hou, Baosheng Yu, and Dacheng Tao. Batchformer: Learning to explore sample relationships for robust representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7256–7266, 2022.

[31] Hang Su, Subhransu Maji, Evangelos Kalogerakis, and Erik G. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 945–953, 2015.

[32] Shuo Chen, Tan Yu, and Ping Li. Mvt: Multi-view vision transformer for 3d object recognition. In *British Machine Vision Conference*, 2021.

[33] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

Junduan Huang is currently an associate researcher in School of Artificial Intelligence, South China Normal University, Foshan, China. He received the B.S. degree and M.S. degree from South China Agriculture University, Guangzhou, China, in 2017 and 2020, respectively, and the Ph.D degree in South China University of Technology, Guangzhou, China, in 2024, respectively. From 2022 to 2023, he was a research intern in Idiap Research Insititute, Martigny, Switzerland. His research interests include biometrics, computer vision, audio signal processing, deep learning and agricultural engineering.

Sushil Bhattacharjee is a member of the Biometrics Security and Privacy Group at the Idiap Research Institute, where he works on biometrics recognition and presentation attack detection problems for a variety of biometrics modalities. His academic interests include signal processing and computer vision. He received a Ph.D. degree from the School of Communication Systems at the École Polytechnique Fédérale de Lausanne (EPFL) in 1999.

Sébastien Marcel heads the Biometrics Security and Privacy group at Idiap Research Institute (Switzerland) and conducts research on face recognition, speaker recognition, vein recognition, attack detection (presentation attacks, morphing attacks, deepfakes) and template protection. He received his Ph.D. degree in signal processing from Université de Rennes I in France (2000) at CNET, the research center of France Telecom (now Orange Labs). He is Professor at the University of Lausanne (School of Criminal Justice) and a lecturer at the École Polytechnique Fédérale de Lausanne. He is also the Director of the Swiss Center for Biometrics Research and Testing, which conducts certifications of biometric products.

Wenxiong Kang received the M.S. degree from Northwestern Polytechnical University, Xi'an, China, in 2003, and the Ph.D. degree from the South China University of Technology, Guangzhou, China, in 2009. He is currently a Professor with the School of Automation Science and Engineering, South China University of Technology. His research interests include biometrics, image processing, pattern recognition, and computer vision.