GSCL: Generative Self-supervised Contrastive Learning for Vein-based Biometric Verification

Wei-Feng Ou, Lai-Man Po, IEEE Senior Member, Xiu-Feng Huang, Wing-Yin Yu, Yu-Zhi Zhao

***Abstract*— Vein-based biometric technology offers secure identity authentication due to the concealed nature of blood vessels. Despite the promising performance of deep learning-based biometric vein recognition, the scarcity of vein data hinders the discriminative power of deep features, thus affecting overall performance. To tackle this problem, this paper presents a generative self-supervised contrastive learning (GSCL) scheme, designed from a data-centric viewpoint to fully mine the potential prior knowledge from limited vein data for improving feature representations. GSCL first utilizes a style-based generator to model vein image distribution and then generate numerous vein image samples. These generated vein images are then leveraged to pretrain the feature extraction network via self-supervised contrastive learning. Subsequently, the network undergoes further fine-tuning using the original training data in a supervised manner. This systematic combination of generative and discriminative modeling allows the network to comprehensively excavate the semantic prior knowledge inherent in vein data, ultimately improving the quality of feature representations. In addition, we investigate a multi-template enrollment method for improving practical verification accuracy. Extensive experiments conducted on public finger vein and palm vein databases, as well as a newly collected finger vein video database, demonstrate the effectiveness of GSCL in improving representation quality.**

***Index Terms*—Biometric vein verification, representation learning, generative adversarial network (GAN), self-supervised contrastive learning (SCL), enrollment.**

I. Introduction

Vein recognition is an emerging biometric technology that exploits the pattern of blood vessels under the skin surface to identify a person. Compared with the traditional extrinsic biometrics (such as face, fingerprint, palmprint), vein biometric can essentially achieve more secure identity recognition due to its anti-counterfeiting and liveness detection ability [1], [2]. Generally, since the hemoglobin in the veins absorbs near-infrared light, the vein pattern inside the finger, palm, or hand can be captured by the near-infrared (NIR) imaging system, appearing as dark regions in a gray-scale image [3]. In practical applications, the quality of vein image could be affected by a variety of factors such as lighting, humidity, temperature, posture or sensor [1], which can cause large intra-class variations and high inter-class similarity of the vein samples and weaken the overall recognition performance. Therefore, it is critical to extract discriminative vein features to achieve robust performance.

Traditional feature extraction approaches [4]–[11] often need to manually design the feature algorithms, which are task-specific and usually have limited generalization ability. In contrast, deep learning (DL) provides a general data-driven paradigm for learning useful representations from massive training data, and has shown promising performance in various computer vision tasks [12], [13]. However, a major challenge in applying deep learning to vein recognition is the lack of vein image training data due to the sensitivity and privacy of the vein information. In fact, the existing public vein databases usually only comprise several thousands of images. Insufficient training data may cause overfitting issues of the deep networks, thereby impairing the performance. Generally, the data shortage issue of deep learning can be relieved by using transfer learning [14]–[17] or data augmentation [18]–[24]. However, the effects of transfer learning are limited because the vein data has a very different distribution from the general-vision datasets (e.g., ImageNet) commonly used for knowledge transfer. Conventional data augmentation can augment the training samples via category-preserving operations. However, it can only augment the intra-class samples, which brings limited invariance prior knowledge. A simple inter-class data augmentation technique using flipping was proposed in [18], but it can only generate fixed vein classes based on the existed classes in the training set. A GAN-based inter-class sample generation method was proposed in [24] for synthesizing arbitrary-pattern vein images, but this approach requires a relatively complicated synthetic procedure in progressive stages and relies on the traditional algorithms to extract the binary vein patterns. Some GAN-based data augmentation approaches [21]–[23] require training different generators for different vein categories, resulting in a more complicated training procedure.

In this paper, we attempt to tackle this vein data scarcity issue from a data-centric point of view by presenting a generative self-supervised contrastive feature learning framework, which aims to fully harness the underlying prior knowledge from limited training data through generative and discriminative modeling to improve the feature representation ability. The proposed framework is illustrated in Figure 1. GSCL first utilizes StyleGAN2 [25] to model the vein image

This work was supported in part by the National Key Research and Development Program of China under Grant 2022YFF0606303, the National Natural Science Foundation of China under Grant 62206054, Research Capacity Enhancement Project of Key Construction Discipline in Guangdong Province under Grant 2022ZDJS028, Dongguan Science and Technology of Social Development Program under Grant 20231800940522. Wei-Feng Ou is with the School of Computer Science and Technology, Dongguan University of Technology, Guangdong, China. Lai-Man Po, Xiu-Feng Huang, Wing-Yin Yu and Yu-Zhi Zhao are with the Department of Electrical Engineering, City University of Hong Kong, Hong Kong, China. (e-mail: ouwf@dgut.edu.cn; eelmpo@cityu.edu.hk; kxhuang3-c@my.cityu.edu.hk; wingyinyu8-c@my.cityu.edu.hk; yzzhao2-c@my.cityu.edu.hk).

This article has been accepted for publication in IEEE Transactions on Biometrics, Behavior, and Identity Science. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2024.3364021

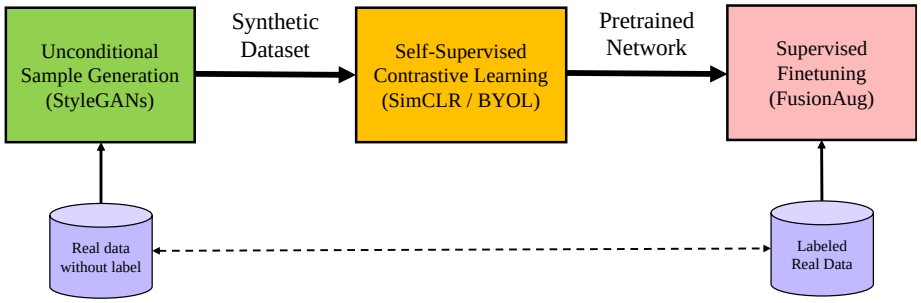


Figure 1. The proposed GSCL framework.

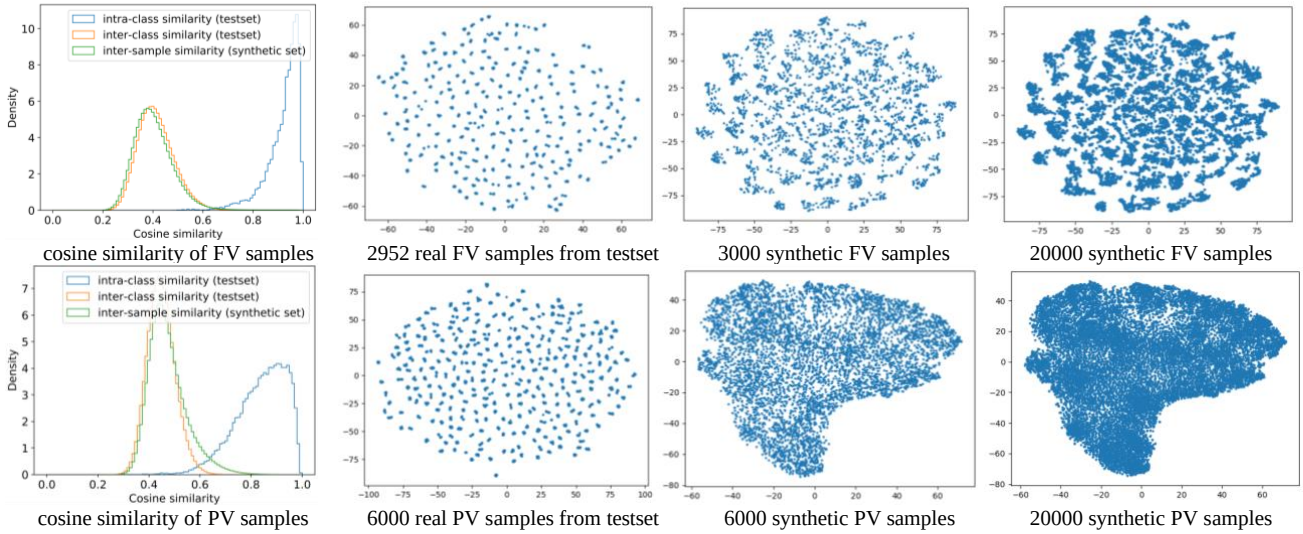


Figure 2. Visualization of (a) cosine similarity distribution of vein image pairs (first column) and (b) feature distribution by t-SNE (second to forth column). For cosine similarity, we train a feature embedding network on the trainset and use the embedding network to evaluate the cosine similarity of intra-class pairs and inter-class pairs on the testset, as well as the inter-sample similarity on 10000 synthetic samples (49995000 image pairs). For fair comparison, each density is normalized to an under-curve area of one. For feature distribution, we use the same embedding networks as those of (a). The details of the trainset and testset can be found in Table 3.

distribution on a relatively small dataset and then uses the learned distribution to produce a larger-sized synthetic vein image dataset. The powerful modeling capabilities of StyleGAN2 enable it to generate realistic and diverse vein image samples, which may provide new semantic knowledge for feature learning. As shown in Figure 2 (a), the inter-sample similarity distribution of the synthetic samples roughly aligns with the inter-class similarity distribution of the real testset. As further shown in Figure 2 (b), the features of the real vein samples in the testset exhibit an obvious pattern of clusters with each cluster corresponding to one vein category, whereas the features of the generated samples show a more random and scattered distribution. This empirical evidence suggests that the generated vein samples are generally discriminative, indicating that the synthetic samples exhibit significant semantic distinctiveness. Therefore, by approximately treating each synthetic sample as an individual class during pretraining and utilizing them for self-supervised contrastive learning, the network can acquire additional semantic knowledge to enhance the feature representation. We investigate two representative self-supervised contrastive learning (SCL) frameworks: SimCLR [26] and BYOL [27]. SimCLR employs both positive and negative samples for contrastive learning to foster data augmentation-invariant and instance-separating representation. In contrast, BYOL eliminates the need for negative samples by employing the stop-gradient trick and a non-symmetric predictor to prevent collapsed solutions. In the context of self-supervised learning, positive samples refer to different augmented views of a given sample, implying that features from various augmented views of the same synthetic image should be invariant. Conversely, negative samples refer to different instances within a min-batch of samples, implying that the features of each instance in a mini-batch should be separated from all the other instances within the batch. Our investigation demonstrates that the distinctive generated samples can provide SimCLR with informative negative instances that promote discriminative feature learning, yielding higher-quality feature representations compared to BYOL. Following the pretraining phase with synthetic samples, we further conduct supervised finetuning using FusionAug [18] on the original labeled training set. This progressive training process allows the network to thoroughly excavate the limited training data to acquire different forms of discriminative a priori knowledge, thereby improving the generalization of learned features. In addition, we present an enrollment technique that employs multiple templates to represent a finger vein identity through feature clustering. This improved enrollment strategy effectively improves the verification accuracy in practical applications. Extensive experiments conducted on the public finger vein (FV) and palm vein (PV) image databases as well as

an in-house finger vein video database demonstrate that the proposed GSCL effectively improves the learning of feature representations and obtains better performance than the recent DL-based methods for biometric vein verification. Our main contributions are summarized as follows:

1) From a data-centric viewpoint, we present a generative self-supervised contrastive feature learning framework, which systematically incorporates generative vein image modeling, self-supervised contrastive learning and supervised finetuning, to fully excavate useful prior knowledge from limited vein data for discriminative feature learning.

2) Through comprehensive analysis on the generated vein samples and investigating two representative SCL approaches, we demonstrate that using unconditionally generated vein samples for self-supervised contrastive learning facilitates learning additional semantic knowledge during pretraining, improving the quality of feature representation.

3) We investigate an improved multi-template enrollment technique and build a finger vein video database for systematic performance evaluation. Through improved feature extraction and enrollment methods, our finger vein verification system shows promising performance in terms of accuracy and efficiency under practical working conditions.

The rest of the paper is organized as follows: In Section II, we provide an overview of literature on biometric vein recognition and self-supervised contrastive learning. Section III introduces the proposed GSCL framework. The proposed multi-template enrollment method is elaborated in Section IV. Section V presents comprehensive experimental results and discussions. Finally, Section VI concludes this paper.

II. Related Works

A. Biometric Vein Recognition

Vein-based biometric research has witnessed exciting progress in recent years. One popular line of research in vein recognition mainly focuses on extracting the binary vein pattern as feature representation through the design of novel algorithms [4]–[10] or novel learning tasks [28]–[32] and has achieved promising performance. Another line of research focuses on extracting discriminative feature vectors or descriptors as representations, such as the embedding learning-based approach [18], [19], [24], the optimization-based approach [33], [34], and the descriptor-based approach [35]–[37]. In particular, deep learning-based vein recognition has primarily faced the challenges of data shortage and the generalization ability of the neural networks, and prior endeavor has attempted to tackle these challenges by using transfer learning [14]–[17], data augmentation [18]–[24], novel network architecture design [3], [38]–[42], as well as better loss functions [18], [43], [44]. On the other hand, some literature focuses on improving the acquisition techniques, such as developing the pulsation-based finger vein acquisition [45], or the reflection-based finger vein imaging device [46] for better portability and user experience. Multimodal biometrics [47]–[50] fusing multiple kinds of hand biometrics also shows an upward trend with promising performance. Besides, recent studies have also investigated techniques of biometric template protection [51]–[56] and attack detection [57], to enhance the security and privacy of vein biometric systems. We briefly categorize the most related literature dedicated to vein biometrics in Table 1. Although previous work [24] has preliminarily explored using GAN to synthesize vein samples for contrastive learning, this approach requires training multiple conditional models and relies on the traditional algorithms to extract the binary vein patterns for constructing the training data of GANs. Differently, this paper explores a simpler unconditional generation scheme and presents new experimental evidence and insights to justify GSCL for tackling the data scarcity problem, as well as developing a new enrolment technique to advance the performance of biometric vein verification comprehensively.

B. Self-supervised Contrastive Learning

One of the earliest SCL approaches can be traced back to Exemplar CNN [58], which views each training example as a distinct class to perform exemplar classification using a parametric classifier. Later, Wu et al. [59] proposes to replace the parametric classifier with a memory bank of instance features, thereby formulating the contrastive loss directly in the feature space. By introducing a queue to store the instance features encoded by a smoothly-updated encoder, momentum contrast (MoCo) [60] can construct a large and consistent dictionary for contrastive learning without being limited by mini-batch size. Without using a momentum encoder, InvaSpread [61] formulates the contrastive loss within the current mini-batch of instances to learn data augmentation-invariant and instance-separating features. SimCLR [26] is similar to InvaSpread, but introduces a new projection head used only at training stage and uses strong data augmentation, which significantly improves the performance. By introducing the stop-gradient technique and a non-symmetric predictor to prevent collapsing solutions, BYOL [27] and SimSiam [62] further reveal that contrastive learning can be performed even without the participation of negative samples. Indeed, these evolving studies are gradually unveiling the essence of deep representation learning. It is worth noting that these pioneering SCL methods were primarily designed for general-vision tasks and assumed the availability of large-scale unlabeled datasets. Differently, vascular biometrics have a very different data domain and suffers from the data scarcity issue. Therefore, this paper attempts to investigate the application of SCL in the context of synthetic samples specific to the vein domain, with the purpose of acquiring additional vein semantic knowledge to enhance the quality of feature representations.

Table 1. Categorization of recent literature on biometric vein recognition.

Type		Reference
Feature extraction	Vein pattern-based	[4]–[10], [28]–[32]
	Embedding learning-based	[1], [3], [14], [16], [18]–[24], [38], [39], [41]–[44], [63]–[71]
	Optimization-based	[33], [34]
	Descriptor-based	[35]–[37]
Acquisition improvement		[45], [46]
Multimodal biometrics		[48]–[50]
Security & privacy enhancement		[51]–[57]

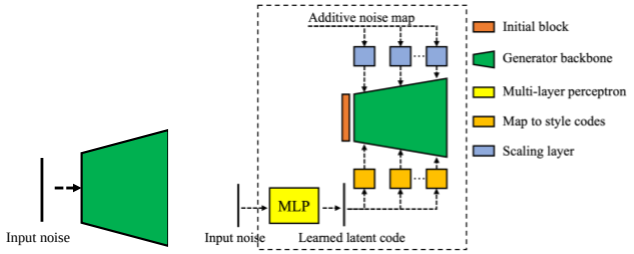


Figure 3. Comparison of (a) traditional generator and (b) style-based generator.

III. Proposed GSCL Framework

The idea of GSCL is to maximally excavate useful prior knowledge from limited training data to improve feature learning from a data-centric viewpoint. The overall diagram of GSCL is illustrated in Figure 1, which comprises three sequential learning steps - unconditional generative modeling of vein images, self-supervised contrastive learning and supervised finetuning. In the following subsections, we will introduce the details of each component.

A. Unconditional Vein Image Modelling

Style-based generative modelling has emerged as a state-of-the-art approach in the field of image generation. Figure 3 provides a visual comparison between the traditional generator architecture and the style-based generator architecture. In the conventional generator, random noise serves as the direct input. In contrast, the style-based generator employs either a constant or a learned input. To further enhance the flexibility of this model, an MLP network is introduced to map random noise to an intermediate latent space that controls the generator's outputs at various layers through learned style codes. Meanwhile, additive noises are further introduced to the feature maps to induce stochastic variations. This improved generator design empowers a more effective control over the generation process, yielding improved fidelity and diversity in the generated outputs. StyleGAN [72] directly controls the activations of the generator through adaptive instance normalization (AdaIN), while StyleGAN2 [25] was redesigned to control the weights of the convolutional kernels and achieve improved generation results. In this paper, we employ StyleGAN2 to achieve unconditional modelling of the vein image distribution. Formally, the training objective of unconditional GAN is given as:

$$\min_G \max_D L(G,D) = \mathbb{E}_x[\log D(x)] + \mathbb{E}_z[\log(1 - D(G(z)))] \quad (1)$$

where G and D denote the generator and discriminator. $x \sim P_d$ $z \sim P_z$, where P_d and P_z denote the real image distribution and the input noise distribution. After training, we generate a synthetic dataset $\tilde{X}=\{\tilde{x}_i\}|_{i=1}^{M}$, where $\tilde{x}_i = G(z)$, $z \sim P_z$. Given the randomness and diversity of the generation process as well as the powerful modelling ability of StyleGAN2, the trained generator can produce numerous vein samples distinct from each other, as analyzed in Figure 2. As a result, we can approximately treat different synthetic instances as different virtual classes to perform self-supervised contrastive learning and acquire additional semantic knowledge. In section V-C, we will perform a qualitative and quantitative assessment of the generated vein images.

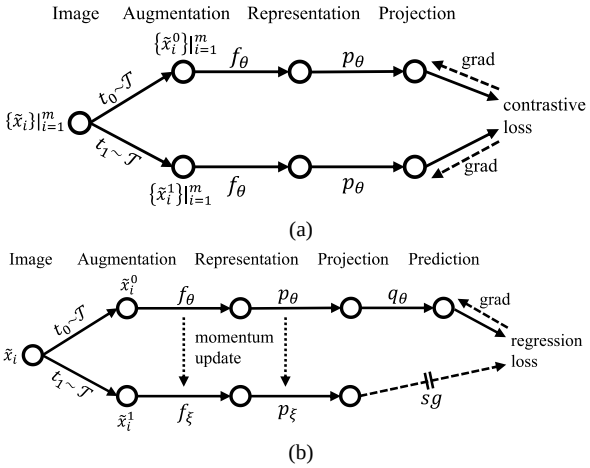


Figure 4. Training pipeline of (a) SimCLR and (b) BYOL.

B. Self-supervised Contrastive Learning on Generated Vein Samples

To evaluate the efficacy of self-supervised contrastive learning on the synthetic dataset, we study two representative approaches – SimCLR [26] and BYOL [27], with their training pipelines depicted in Figure 4. Formally, let $\{\tilde{x}_i\}|_{i=1}^{m}$ denote a mini-batch of m synthetic vein image examples, $\{\tilde{x}_i^0\}|_{i=1}^{m}$ and $\{\tilde{x}_i^1\}|_{i=1}^{m}$ denote two augmented views of $\{\tilde{x}_i\}|_{i=1}^{m}$ with $\tilde{x}_i^0 = t_0(\tilde{x}_i)$ and $\tilde{x}_i^1 = t_1(\tilde{x}_i)$, $t_0 \sim \mathcal{T}$ and $t_1 \sim \mathcal{T}$ denote two random data augmentation transformations drawn from $\mathcal{T}$. Some augmented sample pairs are shown in Figure 5. The data augmentation settings are given in Table 2 while a comparison of different data augmentation strategies will be conducted in Section V-D. The loss function of SimCLR is formulated as:

$$L_{SimCLR} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{\delta=0,1} l_{i,\delta} \quad (2)$$

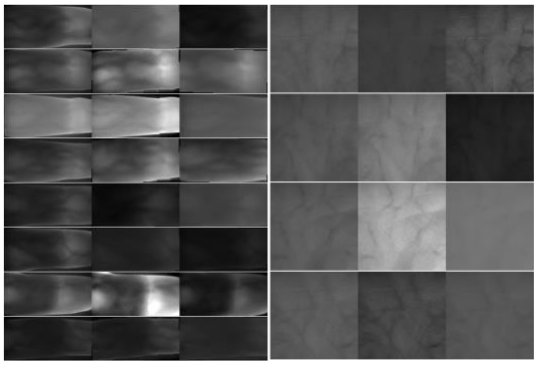
$$l_{i,\delta} = -\log \frac{\exp(\langle h_i^{\delta}, h_i^{1-\delta}\rangle/\tau)}{\sum_{j=1}^{m}\exp(\langle h_i^{\delta}, h_j^{1-\delta}\rangle/\tau) + \sum_{j=1, j\neq i}^{m}\exp(\langle h_i^{\delta}, h_j^{\delta}\rangle/\tau)} \quad (3)$$

where $h_i^{\delta} = p_\theta(f_\theta(\tilde{x}_i^{\delta}))$, f_θ denotes the feature embedding network and p_θ denotes the projection network, $\delta \in \{0,1\}$ is an indicator variable. τ is a temperature parameter, which is empirically set to 0.05 in our experiments, $\langle u, v\rangle \triangleq u^T v/||u||||v||$ denotes the cosine similarity between vectors u and v. On the other hand, BYOL alternately predicts the encoded outputs from one augmented view to the other view without the need for separating negative instances, expressed as:

$$L_{BYOL} = \frac{1}{2m}\sum_{i=1}^{m}\sum_{\delta=0,1} l\left(\tilde{x}_i^{\delta}, \tilde{x}_i^{1-\delta}\right) \quad (4)$$

$$l(a,b) = 2 - 2\langle q_\theta(p_\theta(f_\theta(a))), sg(p_\xi(f_\xi(b)))\rangle \quad (5)$$

where q_θ represents the prediction network and "sg" represents the stop-gradient operation, which are crucial for avoiding collapsed solutions when training without negative samples. f_ξ and p_ξ represent the feature embedding network and the



origin augmented pairs origin augmented pairs

Figure 5. Synthetic sample pairs for contrastive learning via strong data augmentation. The settings of data augmentation can be found in Table 2.

Table 2. Data augmentation settings for self-supervised contrastive learning.

Type	Data Augmentation Pipeline
Finger vein	RandomResizedCrop(size=(64,128, scale=(0.5,1.0), ratio=(1.5,2.5)), RandomRotation(degrees=3), ColorJitter(brightness=0.7, contrast=0.7);
Palm vein	ColorJitter(brightness=0.9, contrast=0.9)

projection network on the target encoder branch, which are updated by the exponential moving average of f_θ and p_θ with a momentum coefficient η, expressed as:

$$\xi \leftarrow \eta\xi + (1-\eta)\theta \tag{6}$$

where θ represents the set of network parameters for f_θ and p_θ, ξ represents the set of network parameters for f_ξ and p_ξ. We empirically set η=0.996 in our experiments. After training, only f_θ will be retained for feature extraction, all the other network components will be discarded.

C. Supervised Finetuning on Real Vein Samples

After self-supervised pretraining on the synthetic dataset, it is necessary to finetune the feature extraction network with the original labeled training set for further improvement. We employ FusionAug for supervised finetuning using fusion loss with intra-class and inter-class data augmentation [18]. Given a mini-batch of n training examples $\{x_i, y_i\}|_{i=1}^n$ comprising P vein classes and K image per class, where x_i and y_i denote the vein image and its class label. The fusion loss is given as:

$$\mathcal{L}_{fusion} = \alpha\mathcal{L}_{lmcl} + \beta\mathcal{L}_{tri} \tag{7}$$

where $\mathcal{L}_{lmcl}$ and $\mathcal{L}_{tri}$ denote the large margin cosine loss [73] and triplet loss [74]. α and β denote the weighting coefficients. $\mathcal{L}_{lmcl}$ is given as:

$$\mathcal{L}_{lmcl} = \frac{1}{n}\sum_{i=1}^{n} -\log\frac{\exp\left(s(\langle w_{y_i}, f_i\rangle - \rho)\right)}{\exp\left(s(\langle w_{y_i}, f_i\rangle - \rho)\right) + \sum_{j\neq y_i}^{c}\exp\left(s\langle w_j, f_i\rangle\right)} \tag{8}$$

where $f_i = f_\theta(x_i) \in R^d$ denotes the feature vector of x_i, $w \in R^{d\times c}$ denotes the classifier weights with c training classes. s and ρ denote the scale and margin parameters. $\mathcal{L}_{tri}$ is given as:

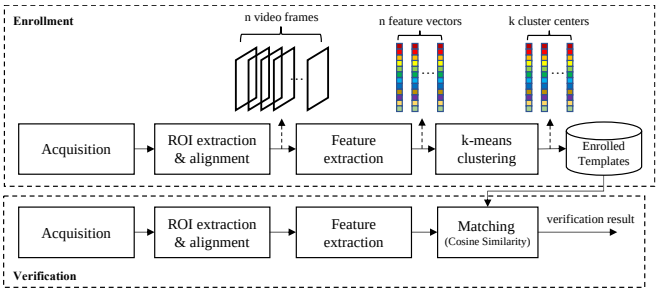


Figure 6. Our finger vein-based biometric verification system using multi-template enrollment.

$$\mathcal{L}_{tri} = \frac{1}{|\mathbb{T}|}\sum_{(a,p,n)\in\mathbb{T}}\left[\|f_a - f_p\|_2^2 - \|f_a - f_n\|_2^2 + \gamma\right]_+ \tag{9}$$

where f_a, f_p, f_n denote the feature vectors of the anchor, positive and negative, respectively. $\mathbb{T}$ denotes a set of semi-hard triplets and $|\mathbb{T}|$ is the cardinality of set $\mathbb{T}$, γ is a margin hyper-parameter and $[\cdot]_+ = \max(0,\cdot)$ denotes the rectifier function. This paper follows the parameter settings of fusion loss provided in [24].

IV. Multi-template Enrollment using Feature Clustering

Enrollment is an indispensable step for biometric applications and has a significant impact on performance. An effective enrollment strategy is essential for accurately and efficiently representing the identity being enrolled, consequently enhancing the overall verification accuracy. However, the investigation of enrollment methods is relatively underexplored in the existing literature. In a prior study [18], a real-time finger vein enrollment technique was presented, which relied on computing average features from consecutive frames. However, this single-template enrollment method exhibited sensitivity to vein pattern deformations induced by finger rolling, leading to the occurrence of false rejection errors. To address this issue, in this paper, we devise a multi-template enrollment method based on feature clustering, which is operated on the fly during enrollment. The overall procedure is shown in Figure 6. At the enrollment stage, we use our in-house finger vein acquisition device to capture n continuous frames from the enrolling finger at around 8 FPS, and then perform ROI extraction and alignment [7][75] to obtain the preprocessed frames $\{v_i\}|_{i=1}^n$. We then extract the corresponding feature vectors $\{f_i\}|_{i=1}^n$ from each frame with a ResNet-18 embedding network f_θ, where $f_i = f_\theta(v_i) \in R^d$. After that, we perform K-means clustering to the feature vectors and use the cluster centers as the enrollment templates for the given finger, expressed as:

$$\{\hat{f}_i\}|_{i=1}^k = kmeans(\{f_i\}|_{i=1}^n, k) \tag{10}$$

where k denotes the number of clusters, $\{\hat{f}_i\}|_{i=1}^k$ denote the k cluster centers, i.e., the enrolled templates. At the verification stage, the max cosine similarity score between the feature vector $f \in R^d$ of the probing finger and the enrolled templates $\{\hat{f}_i\}|_{i=1}^k$ will be computed to perform verification, expressed as:

$$score = max(\{\langle f, \hat{f}_i\rangle\}|_{i=1}^k) \tag{11}$$

Table 3. Overview of four finger vein databases and two palm vein databases as well as summary of our training and testing configurations. For the HKPU database, we only used the images from the first 105 subjects available for both sessions.

Attribute	**FV-USM**	**UTFVP**	**SDMULA**	**HKPU**	**Tongji**	**VERA**
Meta information of database						
# subjects	123	60	106	156	300	50
# fingers (palms) per subject	4	6	6	2	2	2
# images per finger (palm) per session	6	2	6	6	10	5
# sessions	2	2	1	1~2	2	2
# total images	5904	1440	3816	3132	12000	1000
Training & testing configurations						
# images in trainset	2952	0	0	0	6000	0
# images in testset	2952	1440	3816	2412	6000	1000
# genuine pairs from testset	8856	1440	9540	7560	30000	5500
# imposter pairs from testset	2169720	516960	201930	1580040	8970000	1204500

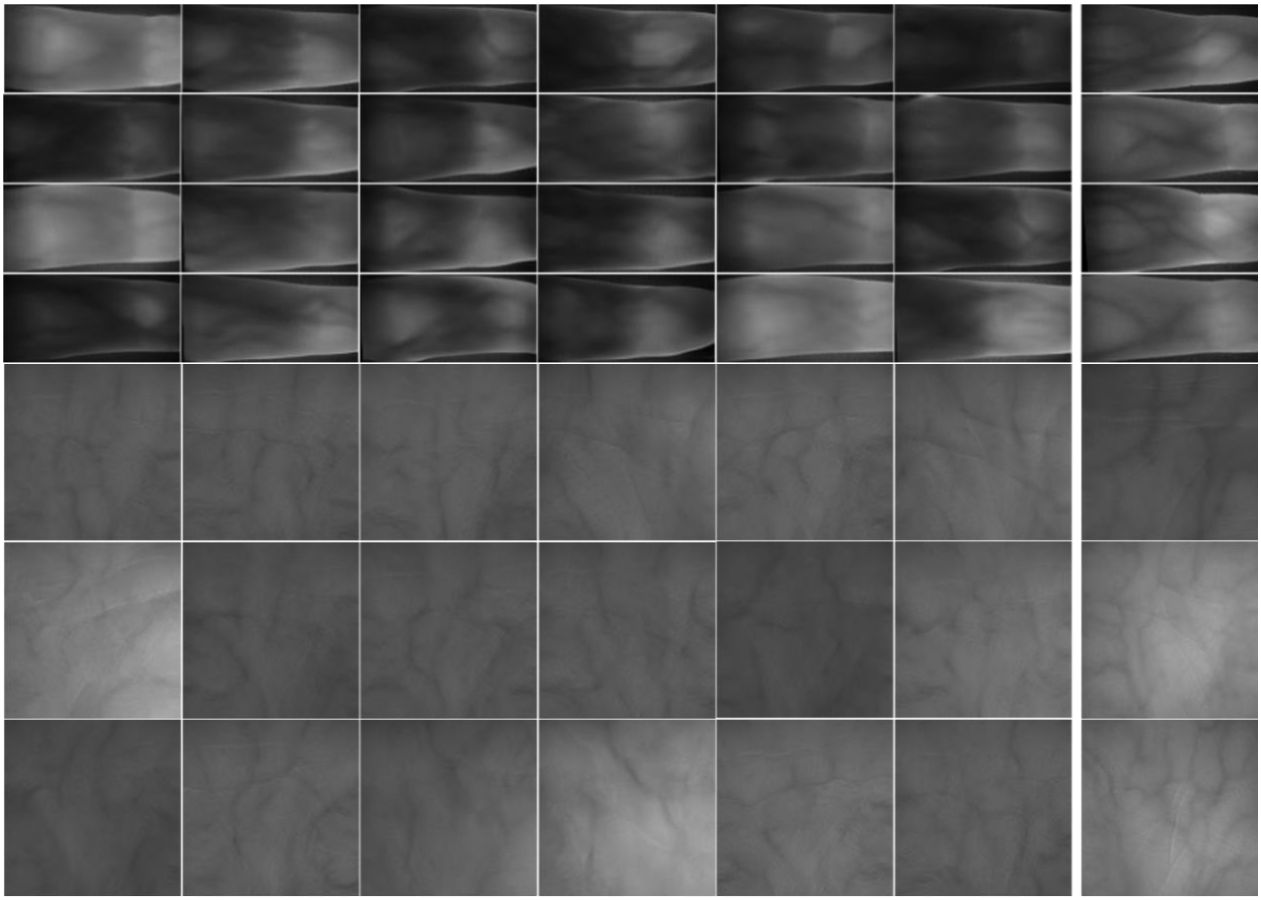


Figure 7. Visual quality assessment. Column 1~6: synthetic vein examples. Column 7: real vein examples; Row 1~4: Finger vein; Row 5~7: Palm vein.

By using this enrollment technique, we can more effectively represent the enrolled finger with multiple feature templates, resulting in improved verification accuracy. We will analyze the settings of n and k in the experimental section.

V. Experiments

A. Datasets and Experimental Setup

We conduct experiments on four finger vein databases (FV-USM [76], UTFVP [77][78], SDMULA [79], HKPU [4]), two palm vein databases (Tongji [63], VERA [80]), and an in-house finger vein video database. Table 3 provides an overview of the public vein image databases, while the in-house database will be detailed in Section V-G. For finger vein, we preprocess all images to obtain the horizontally aligned ROI images with size 64×128 using the classic methods [7], [75]. For palm vein, we used the ROI images released by the authors and resized them to 128×128. Different fingers or palms are treated as different vein classes. Our main experiments are conducted on FV-USM and Tongji. Each database is divided into a trainset (50%) and a testset (50%) according to vein classes. The trainset is used for training the generative and feature extraction models, and the testset is used for performance evaluation. For the UTFVP, SDMULA, HKPU and VERA databases, the whole database is treated as testset for evaluating the generalization performance, which will be detailed in section V-H. To perform biometric verification, we pair the images from different sessions in the testset to construct the genuine and imposter image pairs, which aligns closer with the actual working situation [24]. Cosine

This article has been accepted for publication in IEEE Transactions on Biometrics, Behavior, and Identity Science. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2024.3364021

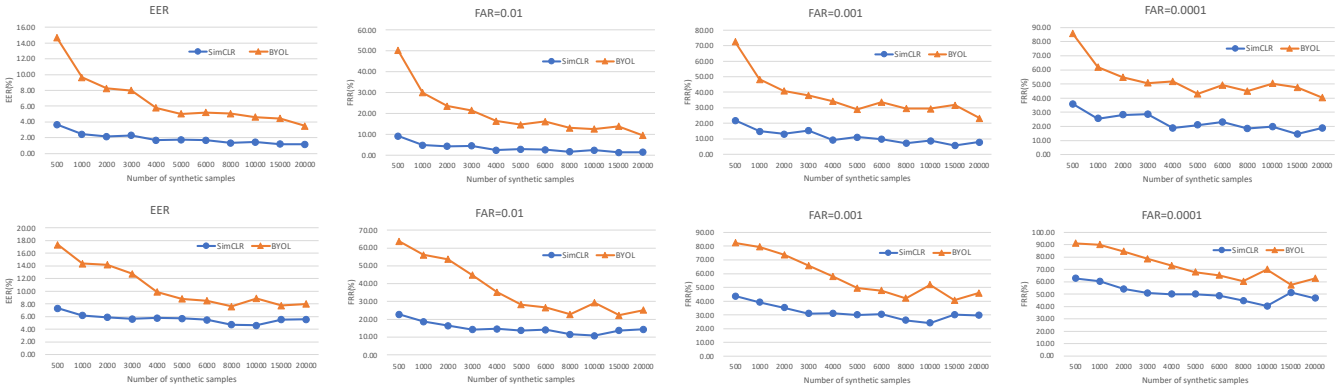


Figure 8. Pretraining performance of GSCL based on different numbers of synthetic training samples. Upper row: FV-USM; Lower row: Tongji.

similarity is used as the distance metric between the feature vectors. We employ receiver operating characteristics (ROC), equal error rate (EER), and false rejection rates (FRR) under different false acceptance rates (FAR) as performance metrics. The training and testing configurations are summarized in Table 3.

B. Network Architectures and Training Details

For vein image generation, we adapt the StyleGAN2 model from [25], which employed a skip-connected generator and a residual discriminator. The generator consists of an initial convolutional layer and six modulated convolutional blocks and finally outputs the vein image of size 128×128 for palm vein or 64×128 for finger vein. The discriminator consists of seven convolutional residual blocks, a final convolutional layer as well as a fully-connected layer which outputs the predicted logits of real or fake. We use Adam optimizer with momentum parameters β_1=0.5, β_2=0.9 and a learning rate of 0.0002 for training the generator, and a learning rate of 0.0003 for training the discriminator, with a mini-batch size of 30. The models were trained for 150000 iterations.

The embedding network used for biometric verification follows the standard ResNet-18 [81], which is initialized by ImageNet-pretrained weights. The projection network of SimCLR follows the "FC(512)+PReLU" architecture, while the projection network and prediction network of BYOL employ the same architecture of "FC(512)+BN(512)+ReLU+FC(128)", where the number represents the output feature dimension. The models were trained with SGD (Stochastic Gradient Descent) using a learning rate of 0.01, a momentum of 0.9 and a weight decay factor of 0.0005 for SimCLR, and using a learning rate of 0.03, a momentum of 0.9 and a weight decay factor of 0.0003 for BYOL. For supervised finetuning, we use SGD with a learning rate of 0.01 without using momentum and weight decay. The learning rates were decayed by multiplying 0.1 at 60^{th} epoch, and the training completes in 80 epochs.

C. Assessment of Generated Vein Images

In Figure 7, we perform a qualitative assessment on the synthetic vein images by demonstrating some typical vein image samples generated by StyleGAN2. The results reveal that the synthetic vein images have a good fidelity and diversity compared to the real vein images. Some detailed textures, such as the palmprint and wrinkles, can be faithfully synthesized. A good variety in the vein patterns, contrast, and brightness as well as the various shapes of the fingers can also be observed. These results show that the generator achieves relatively good modeling of the real image distribution. Therefore, the sample generation process from the learned distribution may be "imagined" as collecting more vein images from some random "virtual" human subjects, thereby may introduce new semantic knowledge that can be learned by using self-supervised contrastive learning due to the distinction of the synthetic vein samples.

In Table 4, we perform a quantitative assessment of the generated vein images using the Fréchet Inception Distance (FID), which is a widely used metric for assessing the realism and diversity of generated images by measuring the Fréchet distance between the features of two sets of images extracted by an Inception-v3 model. We interpret the FID between the training and testing splits of the same database as ground truth. Meanwhile, the FID between two different vein databases serves as the comparative reference. We can observe that the FID of generated vein samples is generally reasonable, lying between the ground truth and the cross-database reference. The HKPU has a larger FID due to its observable visual difference compared to FV-USM. The FID of synthetic palm vein images is larger than that of finger vein, which may be because the palm vein images have more complex details and present a more challenging task for generative modeling.

D. Pretraining Performance of GSCL

Figure 8 shows the pretraining performance of GSCL when training with different number of synthetic vein samples using SimCLR and BYOL respectively. The results showed that SimCLR consistently outperformed BYOL with a lower error rate under different false acceptance rates. This result is different from the prior study [27] where BYOL usually achieves similar or better performance than SimCLR in the general-object datasets. We speculate that this is because the sample difference of the synthetic vein samples is relatively small compared to that of the general-object images. As a result, the explicit participation of negative samples in contrastive learning can help SimCLR more effectively learn to distinguish between different samples, which can enhance the inter-class separation ability of the learned features, thereby achieving more discriminative power than BYOL. In contrast, since the images in the general-object datasets usually have larger sample

Table 4. Evaluation of FID for the generated vein images. (lower is better)

Type	Setting	FID↓
Finger vein	FV-USM_trainset vs. FV-USM_testset (GT)	13.1
	FV-USM_real vs. FV-USM_generated	48.1
	FV-USM vs. UTFVP	68.2
	FV-USM vs. HKPU	221.7
Palm vein	Tongji_trainset vs. Tongji_testset (GT)	6.6
	Tongji_real vs. Tongji_ generated	72.7
	Tongji vs. VERA-Palmvein	93.0

Table 5. Comparison of EER (%) by performing self-supervised contrastive learning on real training set and synthetic samples, respectively.

Type	# samples	SimCLR	BYOL
Real FV training set	2952	3.77	6.03
Synthetic FV samples	3000	2.29	7.99
Synthetic FV samples	20000	1.16	3.49
Real PV training set	6000	8.86	11.00
Synthetic PV samples	6000	5.48	8.51
Synthetic PV samples	10000/15000*	4.62	7.77

(*: 10000 for SimCLR, 15000 for BYOL)

Table 6. Comparison of different data augmentation strategies for self-supervised contrastive learning on synthetic samples based on SimCLR.

Type	ColorJitter	RandRsCrop	RandRotation	EER (%)
FV	✓	✓	✓	1.16
	✓	✓		1.23
	✓		✓	10.66
		✓	✓	20.93
PV	✓	✓	✓	7.87
	✓	✓		8.01
	✓			4.62

Table 7. Performance of SimCLR under different temperature (τ) settings.

Temperature		0.03	0.04	0.05	0.06	0.07	0.08
EER (%)	FV-USM	1.21	1.03	1.16	1.08	1.19	1.38
	Tongji	4.21	4.34	4.62	5.33	5.02	5.02

Table 8. Performance of BYOL under different momentum (η) settings.

Momentum		0	0.98	0.99	0.993	0.996	0.999
EER (%)	FV-USM	4.74	5.00	4.97	3.90	3.49	4.27
	Tongji	12.55	11.49	9.80	10.46	7.77	7.39

Table 9. Performance of SimCLR under different batch sizes. The learning rates are adjusted for different batch sizes for a fair comparison.

Batch size		16	32	64	128	256	512
EER (%)	FV-USM	1.22	1.13	1.16	1.19	1.30	1.54
	Tongji	7.13	5.93	4.62	4.28	4.81	7.34

Table 10. Performance of BYOL under different batch sizes. The learning rates are adjusted for different batch sizes for a fair comparison.

Batch size		16	32	64	128	256	512
EER (%)	FV-USM	3.54	4.74	3.49	4.49	4.00	6.32
	Tongji	10.58	9.85	7.77	7.15	9.95	17.67

Table 11. Performance of SimCLR and BYOL when using different MLP architectures as the projector for SimCLR and as the projector and predictor for BYOL.

Method	MLP architecture	EER (%)	
		FV-USM	Tongji
SimCLR	FC+PReLU	1.16	4.62
	FC+BN+ReLU+FC	1.65	6.25
BYOL	FC+PReLU	6.11	14.82
	FC+BN+ReLU+FC	3.49	7.77

difference, therefore, BYOL can easily learn the inter-class separability even without the explicit participation of negative samples. With the increase of synthetic samples, the performance can be gradually improved, and gradually become saturated beyond a certain amount (around 2~3 times of the original training set size). This is generally reasonable because, after all, the generator was trained on a relatively small dataset, it is unrealistic to perfectly model the real image distribution to generate new semantic knowledge infinitely.

In Table 5, we also compare with applying contrastive learning directly on the real training set, which shows inferior performance compared to contrastive learning on the synthetic samples. This may be because the real vein samples from the same category usually have a high similarity that may be harmful for self-supervised contrastive learning, as it implicitly treats different samples as a unique category to learn discriminative features. In contrast, the distinctness of the synthetic samples naturally satisfies the instance separating assumption of self-supervised contrastive learning, thereby helpful for learning discriminative features. In Table 6, we examine the efficacy of different data augmentation strategies for self-supervised contrastive learning. The results show that a combination of geometric and photometric augmentations is important for synthetic finger vein images. In contrast, color jittering alone yields the best performance for synthetic palm vein images. This variation in data augmentation strategies is mainly attributed to the distinct characteristics of the vein databases. The palm vein testset primarily presents photometric variations, such as changes in lighting and contrast. Color jittering augmentation effectively captures these variations,

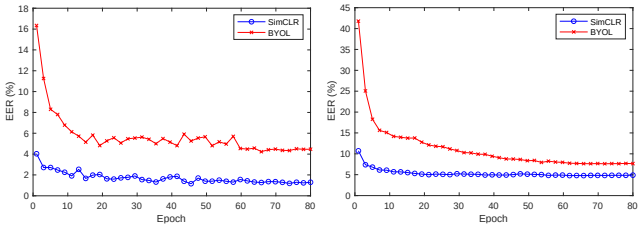


Figure 9. Convergence of SimCLR and BYOL when trained on the synthetic dataset. Left: FV-USM; Right: Tongji.

enhancing model performance on this specific data. Conversely, rigid geometric augmentations, like rotations, scaling and cropping, can distort the palm vein structure and disrupt the data distribution, leading to suboptimal results. In contrast, the finger vein testset exhibits both photometric and geometric variations. The latter likely arises from varying finger positions during image acquisition. To effectively model this sample diversity, an integrated augmentation approach incorporating both color jittering and geometric augmentations is important. This strategy enhances the model's ability to generalize to unseen data.

On the other hand, we also experimented with different hyper-parameter settings when pretraining on the synthetic samples to better understand the performance of GSCL. The performance of SimCLR under different temperature τ and the performance of BYOL under different momentum η are summarized in Table 7 and Table 8 respectively, showing a relative stable performance across different parameter settings. The performance of SimCLR and BYOL under various batch

This article has been accepted for publication in IEEE Transactions on Biometrics, Behavior, and Identity Science. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TBIOM.2024.3364021

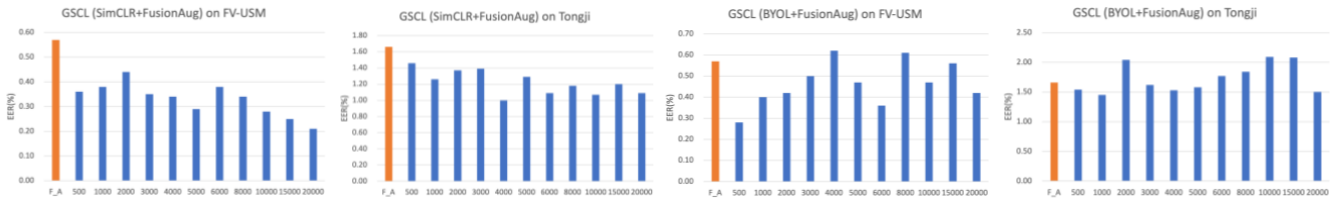


Figure 10. Finetuning performance of GSCL when pretrained with different number of synthetic samples. The results in orange represent the supervised baseline FusionAug (F_A) which is only trained on the real training set.

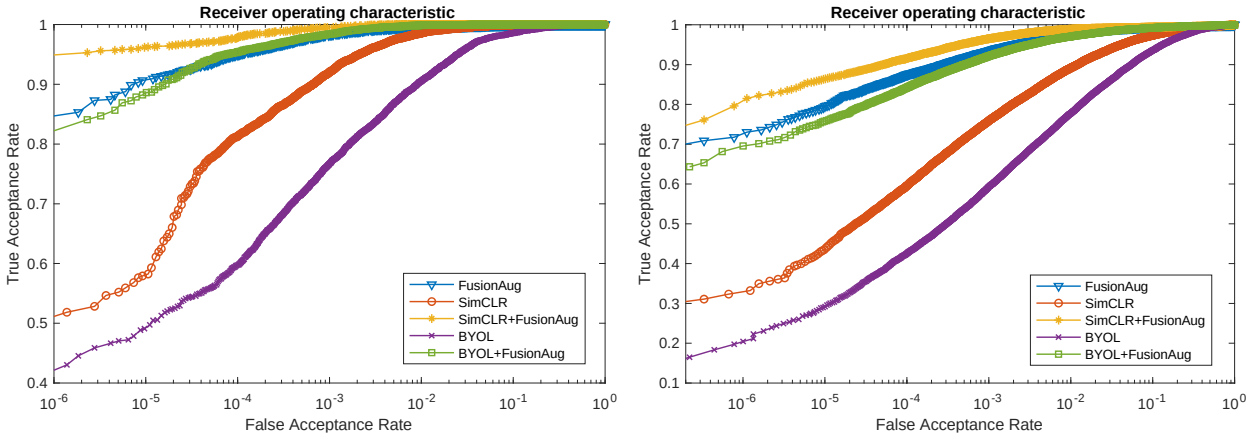


Figure 11. ROC curves of different training schemes. Left: FV-USM; Right: Tongji

Table 12. Verification performance (%) of different training schemes. (FusionAug: train on original training set by FusionAug; SimCLR (20000): train on 20000 synthetic samples with SimCLR; SimCLR+FusionAug: pretrain with SimCLR (20000) and finetune with FusionAug; The same notations apply to BYOL.)

Database	Method	EER	FRR@FAR=10^{-2}	FRR@FAR=10^{-3}	FRR@FAR=10^{-4}
FV-USM	FusionAug	0.45 ±0.11	0.28 ±0.18	1.50 ±0.22	4.96 ±0.72
	SimCLR (20000)	1.18 ±0.11	1.36 ±0.19	6.37 ±1.38	15.14 ±3.22
	SimCLR+FusionAug	**0.24 ±0.03**	**0.03 ±0.04**	**0.66 ±0.14**	**2.32 ±0.19**
	BYOL (20000)	3.95 ±0.42	9.00 ±0.40	22.54 ±0.87	41.84 ±2.89
	BYOL+FusionAug	0.50 ±0.09	0.22 ±0.09	2.02 ±0.61	7.13 ±2.28
Tongji	FusionAug	1.56 ±0.14	2.07 ±0.35	5.91 ±0.31	11.69 ±0.73
	SimCLR (10000)	4.74 ±0.37	11.56 ±1.33	25.78 ±2.80	43.55 ±4.78
	SimCLR+FusionAug	**1.11 ±0.12**	**1.20 ±0.22**	**3.66 ±0.26**	**8.40 ±0.53**
	BYOL (15000)	8.29 ±0.53	25.36 ±2.69	45.85 ±4.76	63.64 ±6.19
	BYOL+FusionAug	1.93 ±0.10	2.83 ±0.12	7.71 ±0.16	15.86 ±0.20

sizes are summarized in Table 9 and Table 10 respectively. A batch size of 64 usually works well for both finger vein and palm vein samples and thus is adopted in our final pretraining scheme. Table 11 shows the performance of using different projector (predictor) architectures for SimCLR and BYOL. The results showed that a simple projector architecture of FC+PReLU works well for SimCLR, while BatchNorm is necessary for BYOL's projector and predictor to achieve better performance, which is also consistent with the empirical findings of [27]. Figure 9 illustrates the training process, showing that SimCLR and BYOL can be effectively converged within 80 training epochs.

E. Finetuning Performance of GSCL

Figure 10 summarizes the performance of GSCL when further finetuned on the original training set with category labels. We use the prior work FusionAug [18] as the supervised baseline for finetuning. We can find that SimCLR+FusionAug obtained a consistent performance improvement compared to the supervised baseline under different numbers of synthetic samples, indicating the benefit of SSL pretraining on synthetic samples. However, BYOL+FusionAug cannot achieve consistent performance improvement compared to the supervised baseline due to its weak pretraining performance. In Table 12, we report the average performance of different training schemes by repeating the training three times with different random seeds. The corresponding ROC curves are shown in Figure 11. We can again find that SimCLR-based GSCL (SimCLR+FusionAug) achieved consistent performance improvement compared to the supervised baseline FusionAug under different false acceptance rates, whereas the BYOL-based GSCL showed less effective performance improvement compared to the supervised baseline. Moreover, finetuning on the original training set is also critical to achieve better performance, since the category labels can provide extra supervision and semantic knowledge. Based on these results, we employ SimCLR+FusionAug as the final scheme for GSCL.

Additionally, as presented in Table 13, we evaluate the performance of GSCL under two distinct pretraining paradigms:

Table 13. Performance of GSCL under supervised pretraining and self-supervised pretraining with the synthetic samples, respectively.

Setting	EER (%)	
	FV-USM	Tongji
GSCL (supervised pretraining)	0.33	1.30
GSCL (self-supervised pretraining)	**0.21**	**1.07**

Table 14. Performance comparison of different approaches for biometric vein verification.

	Approach	EER (%)	Time for Feature extraction (ms)
FV-USM	MC	3.74	69.8
	PC	3.44	**4.3**
	RLT	5.24	954.1
	WLD	5.55	10.9
	Gabor	4.41	11.6
	cGAN	3.15	23.4
	CycleGAN	4.61	27.4
	CNN	1.69*	-
	FusionAug	0.45	7.8
	InterSample	0.32	7.8
	GSCL	**0.21**	7.8
Tongji	MC	1.25	132.3
	PC	1.28	**5.4**
	RLT	2.16	1028.8
	WLD	2.36	23.7
	Gabor	1.37	16.1
	cGAN	1.03	44.3
	CycleGAN	2.27	49.4
	$PalmR_{CNN}$	2.30*	-
	CR_CompCode	4.41	7.2
	FusionAug	1.56	9.2
	InterSample	1.19	9.2
	GSCL	**1.02**	9.2

Table 15. Verification performance (FV-USM) and model complexity of GSCL using different embedding network architectures.

Network architecture		ResNet-18	ResNet-34	ResNet-50
Verification performance (%)				
Pretrain	EER	1.16	1.17	1.20
	FRR@FAR=10^{-2}	1.43	1.46	1.42
	FRR@FAR=10^{-3}	7.96	8.05	5.48
	FRR@FAR=10^{-4}	18.71	17.85	11.90
Finetune	EER	0.21	0.23	0.21
	FRR@FAR=10^{-2}	0.00	0.03	0.00
	FRR@FAR=10^{-3}	0.50	0.54	0.49
	FRR@FAR=10^{-4}	2.18	1.66	1.83
Model complexity				
Feature extraction time (ms)		7.8	13.0	19.7
G-FLOPs		0.6	1.2	1.34
Model parameters (million)		11.4	21.6	24.6

supervised pretraining and self-supervised pretraining on synthetic samples. The results showed that supervised pretraining yields a reduced performance, indicating the efficacy of the proposed scheme. This is because using supervised learning for the synthetic samples requires a large instance-based parametric classifier, potentially leading to inefficient optimization and reduced performance [61]. In contrast, SCL performs representation learning explicitly in the feature space by learning augmentation-invariant and instance-separating features, achieving improved learning efficiency and enhanced representation quality.

F. Comparison of Different Approaches

We compare the proposed method with different approaches for biometric vein verification, including the classic approaches

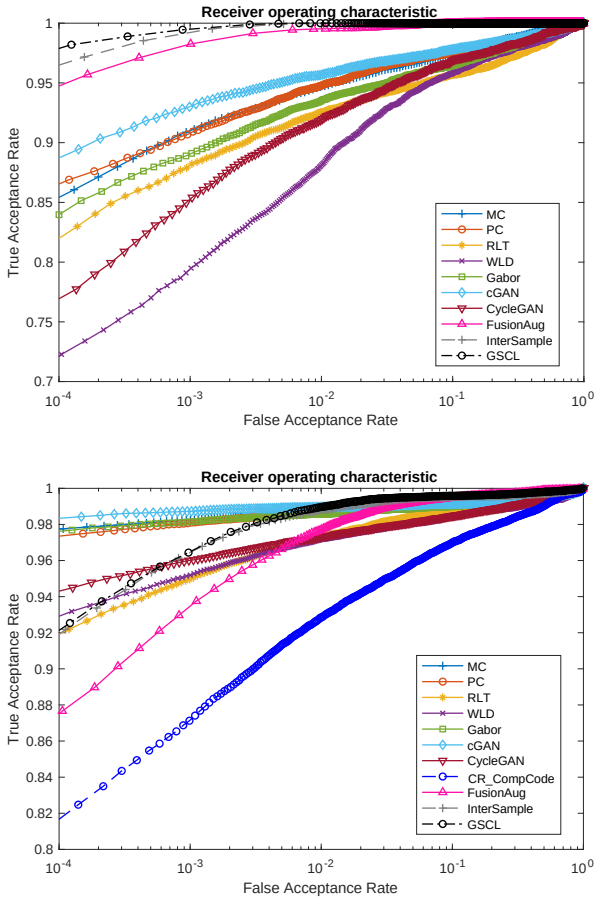


Figure 12. ROC curves of different approaches for biometric vein verification. Up: FV-USM; Down: Tongji.

[10] based on binary vein pattern extraction (MC [8], PC [82], RLT [6], Gabor [4], WLD [7], cGAN [30], [83], CycleGAN [29], [84], CNN [28]), and the deep learning-based approaches based on embedding vector extraction (FusionAug [18], InterSample [24], $PalmR_{CNN}$ [63], CR_CompCode [85]). For finger vein verification, the embedding network of GSCL is pretrained on 20000 synthetic finger vein samples by SimCLR and then finetuned on the original training set by FusionAug. For palm vein verification, GSCL's embedding network is pretrained on 10000 synthetic palm vein samples by SimCLR and then finetuned on the original palm vein training set by FusionAug. The EER performance and feature extraction time of different approaches are summarized in Table 14, while the ROC curves are shown in Figure 12. The feature extraction time is measured by the average time (run by 100 times) for extracting features from a single image under a 3.6GHz Inter CPU. We can find that GSCL outperformed the other approaches for finger vein verification, while the vein pattern-based approaches tend to achieve better performance for palm vein verification at lower false acceptance rates, as can be observed from the ROC curves in Figure 12. This may be because the palm vein images contain additional identity information (e.g., palmprint and the palm wrinkles) that can be indistinguishably extracted by the vein pattern-based approaches to enrich the features, whereas the embedding learning approaches fail to capture such subtle and useful features [24]. Moreover, the approaches of cGAN and

Table 16. Overview of the in-house finger vein video database.

# subjects	# fingers	# sessions	# videos	# frames
21	126	2	252	30240

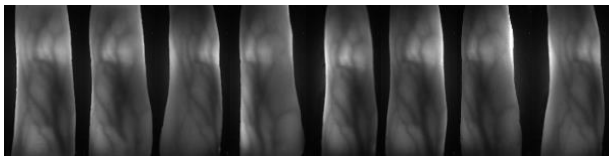


Figure 13. Example frames from a finger vein sample video.

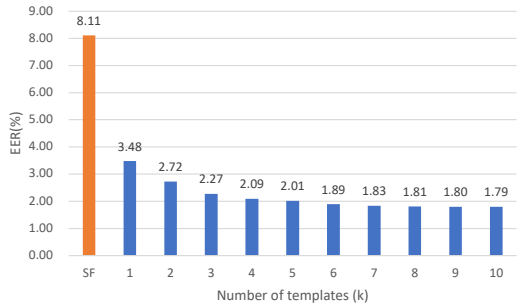


Figure 14. Performance of using different number of templates based on n=120 enrolled frames.

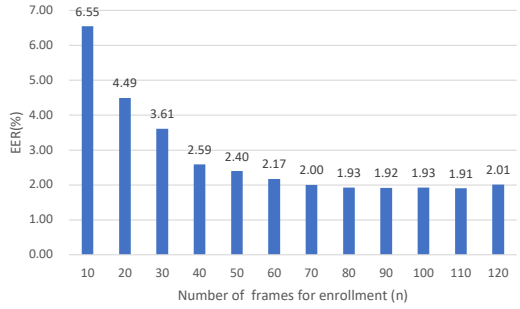


Figure 15. Performance of using different number of frames for enrollment based on k=5 templates.

CycleGAN requires using traditional algorithms to extract the binary vein patterns to construct the training data, whereas the GSCL has no such reliance. It is worth noting that GSCL achieved superior performance to InterSample [24], which is most related to our approach but using a more complicated training process based on conditional vein image generation in multiple synthetic stages. In addition, the feature extraction time of GSCL is lower than most of the approaches, requiring less than 10 ms for single image feature extraction under a regular desktop CPU. By fully mining the semantic prior knowledge of the vein data through generative and contrastive modeling, the proposed GSCL approach effectively alleviates the data shortage issue and obtains superior performance compared to the latest deep learning approaches for vein-based biometric verification.

In addition to the Resnet-18, we also tried using deeper model as the feature embedding network, such as Resnet-34 and Resnet-50. The verification performance and model complexity of different embedding networks trained by GSCL are summarized in Table 15. We can observe that the ResNet-18 generally achieves the best trade-off between verification accuracy and efficiency. Indeed, due to the relatively small size of the vein databases, the model capacity of Resnet-18 may be enough for vein-based biometric verification applications.

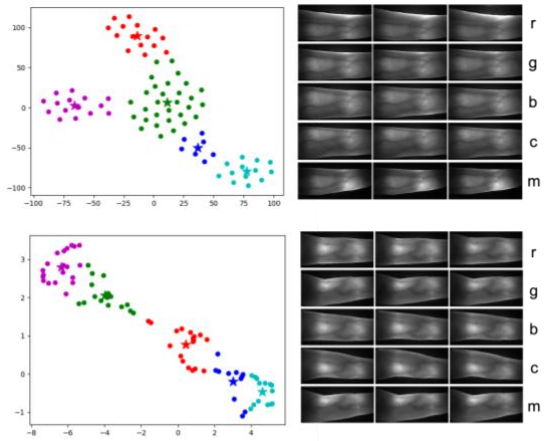


Figure 16. Qualitative analysis of feature clustering based on two different enrolled fingers from different subjects. On the left, we visualize the clustered features by tSNE based on n=80 frames and k=5 templates. Different colors represent different clusters, the cluster centers (enrolled templates) are marked by star. On the right, we show the top-3 vein images closest to each cluster center. "r, g, b, c, m" corresponds to red, green, blue, cyan, magenta, respectively.

G. Analysis of Multi-template Enrollment

In order to provide a quantitative performance evaluation of the proposed multi-template enrollment method, we collected a finger vein video database to simulate on-the-fly enrollment. The details of the database can be found in Table 16. The database was collected from 21 subjects in two separate sessions. The time interval of the two sessions is between 1~3 days. In each session, we acquired one video from each of the six fingers (index, middle, ring of both hands) of each subject, resulting in 252 videos in total. Each video contains 120 continuous vein image frames (around 15 seconds), which cover a variety of geometric variations of the finger (such as translation, rotation, pitching, yawing and rolling within a sensible range). Some example frames from a sample video are shown in Figure 13. The whole database contains 30240 image frames at a resolution of 200×400. The first session is used for enrollment, while the second session is used for testing. As a result, the verification set contains 15120 genuine matchings and 1890000 imposter matchings for performance evaluation. To obtain better performance at practical application, we trained the feature embedding network (ResNet-18) on the entire FV-USM database based on the GSCL scheme. Specifically, we first trained a StyleGAN2 generator using the entire FV-USM database. We then used this generator to generate 20000 finger vein images to pretrain the embedding network with SimCLR. After that, we further finetuned the embedding network on the entire FV-USM database using FusionAug. It is worth noting that the training database is completely irrelevant to our in-house testing database, which is a challenging open-set scenario. For enrollment, we run K-means clustering on each vein class (120 frames) separately and consider the same number of clusters for each class. The dimension of the embedding feature vector is 512.

Table 17. Verification performance of the proposed method under different generalization settings.

Generalization setting	Test database	EER (%)
Inner-DB	FV-USM	0.21
Cross-DB	UTFVP	2.64
	SDMULA	3.09
	HKPU	5.26
Cross-modality	Tongji	10.11
	VERA	10.22

Figure 14 showed that our multi-template approach significantly outperformed the baseline approach SF, which uses a single frame (SF) of the finger at standard frontal position for enrollment. As the number of templates was increased, the performance was improved accordingly, and gradually became plateaued beyond 5 templates. Since using more templates would cost extra storage and increase the matching computation, we empirically suggest a template number between 3~6 for practical use as a trade-off between verification performance and computational complexity. In Figure 15, we evaluated the performance of capturing different numbers of frames for enrollment based on 5 templates. The results showed that increasing the number of frames effectively improved the performance, since it can record more variations of the enrolled finger. The performance improvement gradually saturates at round 80 frames. In Figure 16, we visualize the distribution of the clustered features and the corresponding vein image samples. We can observe a large intra-class variation within each enrolled finger, which can be sufficiently represented through different feature templates. In essence, our enrollment method is built upon the principles of feature-level fusion (using the cluster centers as templates for enrollment) and score-level fusion (using the max-fusion rule during verification). Adapting these fusion principles into the context of enrollment facilitates fusing diverse and valuable information from a finger vein video sequence to represent a person's identity more effectively, thereby leading to improved verification performance.

H. Generalization Performance

In this section, we examine the performance of GSCL under different generalization scenarios. The feature extraction model was trained on the trainset of FV-USM, and evaluated on different vein databases, covering different generalization scenarios: inner-database, cross-database and cross-modality generalization. The results are summarized in Table 17. We can observe that GSCL has impressive inner-database generalizability. The performance remains promising for cross-database generalization. However, a significant performance drop is observed for cross-modality generalization from finger vein to palm vein. This is largely because different vascular modalities have different discriminating characteristics and image distributions. Exploration of universal feature representation that can generalize to different vascular modalities is an interesting future work.

VI. Conclusions

This paper presents a generative self-supervised contrastive learning framework, designed from a data-centric viewpoint for fully mining the potential prior knowledge in the context of limited vein data to enhance feature representations for vascular biometric verification. The proposed GSCL framework incorporates three key components: unconditional generative image modeling, self-supervised contrastive learning, and supervised finetuning. The generative modeling can provide contrastive learning with numerous discriminative synthetic vein image samples carrying rich semantic knowledge, while the discriminative modeling with self-supervised pretraining and supervised finetuning enables the network to effectively learn rich prior knowledge from synthetic and real image samples for enhancing feature representations. Additionally, we introduce a clustering-based multi-template enrollment method, which contributes to more reliable enrollment and effectively improves the practical biometric verification performance. Comprehensive experiments performed on seven vein databases demonstrate the efficacy of the proposed method in improving the feature representation ability and its promising performance.

Future work shall explore more effective self-supervised learning techniques (e.g., mask image modeling) to tackle the data scarcity challenge more effectively. Furthermore, the generative self-supervised contrastive learning approach offers an alternative method for utilizing synthetic data in lieu of raw data when training deep neural networks. This eliminates the need for direct access to real biometric data, thereby promoting privacy protection for AI research.

Reference

[1] W. Kang, H. Liu, W. Luo, and F. Deng, "Study of a full-view 3D finger vein verification technique," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 1175–1189, 2019.

[2] W. Jia *et al.*, "A survey on dorsal hand vein biometrics," *Pattern Recognit.*, vol. 120, p. 108122, 2021.

[3] R. Das, E. Piciucco, E. Maiorana, and P. Campisi, "Convolutional neural network for finger-vein-based biometric identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 2, pp. 360–373, 2018.

[4] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. image Process.*, vol. 21, no. 4, pp. 2228–2244, 2011.

[5] Y. Zhou and A. Kumar, "Human identification using palm-vein images," *IEEE Trans. Inf. forensics Secur.*, vol. 6, no. 4, pp. 1259–1274, 2011.

[6] N. Miura, A. Nagasaka, and T. Miyatake, "Feature extraction of finger-vein patterns based on repeated line tracking and its application to personal identification," *Mach. Vis. Appl.*, vol. 15, no. 4, pp. 194–203, 2004.

[7] B. Huang, Y. Dai, R. Li, D. Tang, and W. Li, "Finger-vein authentication based on wide line detector and pattern normalization," in *2010 20th international conference on pattern recognition*, 2010, pp. 1269–1272.

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

[9] L. Yang, G. Yang, Y. Yin, and X. Xi, "Finger vein recognition with anatomy structure analysis," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 8, pp. 1892–1905, 2017.

[10] C. Kauba and A. Uhl, "An available open-source vein recognition framework," in *Handbook of Vascular Biometrics*, Springer, Cham, 2020, pp. 113–142.

[11] A. Uhl, C. Busch, S. Marcel, and R. Veldhuis, Eds., *Handbook of Vascular Biometrics*. Cham: Springer International Publishing, 2020.

[12] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[14] R. S. Kuzu, E. Maioranay, and P. Campisi, "Vein-based biometric verification using transfer learning," in *2020 43rd International Conference on Telecommunications and Signal Processing (TSP)*, 2020, pp. 403–409.

[15] N. A. Al-johania and L. A. Elrefaei, "Dorsal hand vein recognition by convolutional neural networks: feature learning and transfer learning approaches," *Int. J. Intell. Eng. Syst,* vol. 12, no. 3, pp. 178–191, 2019.

[16] Z. Pan, J. Wang, G. Wang, and J. Zhu, "Multi-scale deep representation aggregation for vein recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 1–15, 2020.

[17] K. Shaheed *et al.*, "DS-CNN: A pre-trained Xception model based on depth-wise separable convolutional neural network for finger vein recognition," *Expert Syst. Appl.*, vol. 191, p. 116288, Apr. 2022, doi: 10.1016/j.eswa.2021.116288.

[18] W. F. Ou, L. M. Po, C. Zhou, Y. A. U. Rehman, P. F. Xian, and Y. J. Zhang, "Fusion loss and inter-class data augmentation for deep finger vein feature learning," *Expert Syst. Appl.*, vol. 171, Jun. 2021, doi: 10.1016/j.eswa.2021.114584.

[19] H. Hu *et al.*, "FV-Net: learning a finger-vein feature representation based on a CNN," in *2018 24th International Conference on Pattern Recognition (ICPR)*, 2018, pp. 3489–3494.

[20] J. M. Song, W. Kim, and K. R. Park, "Finger-vein recognition based on deep DenseNet using composite image," *IEEE Access*, vol. 7, pp. 66845–66863, 2019.

[21] J. Zhang, Z. Lu, M. Li, and H. Wu, "GAN-Based Image Augmentation for Finger-Vein Biometric Recognition," *IEEE Access*, vol. 7, pp. 183118–183132, 2019.

[22] H. Qin, M. A. El-Yacoubi, Y. Li, and C. Liu, "Multi-Scale and Multi-Direction GAN for CNN-Based Single Palm-Vein Identification," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 2652–2666, 2021.

[23] G. Wang, C. Sun, and A. Sowmya, "Learning a Compact Vein Discrimination Model With GANerated Samples," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 635–650, 2019.

[24] W. F. Ou, L.-M. Po, C. Zhou, P.-F. Xian, and J.-J. Xiong, "GAN-based Inter-Class Sample Generation for Contrastive Learning of Vein Image Representations," *IEEE Trans. Biometrics, Behav. Identity Sci.*, p. 1, 2022, doi: 10.1109/TBIOM.2022.3152345.

[25] T. Karras, S. Laine, M. Aittala, J. Hellsten, J. Lehtinen, and T. Aila, "Analyzing and improving the image quality of stylegan," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 8110–8119.

[26] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*, 2020, pp. 1597–1607.

[27] J.-B. Grill *et al.*, "Bootstrap your own latent: A new approach to self-supervised learning," *arXiv Prepr. arXiv2006.07733*, 2020.

[28] H. Qin and M. A. El-Yacoubi, "Deep representation-based feature extraction and recovering for finger-vein verification," *IEEE Trans. Inf. Forensics Secur.*, vol. 12, no. 8, pp. 1816–1829, 2017.

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

[30] H. Qin and M. A. El Yacoubi, "End-to-End Generative Adversarial Network for Palm-Vein Recognition," in *International Conference on Pattern Recognition and Artificial Intelligence*, 2020, pp. 714–724.

[31] L. Yang, G. Yang, K. Wang, F. Hao, and Y. Yin, "Finger Vein Recognition via Sparse Reconstruction Error Constrained Low-Rank Representation," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 4869–4881, 2021, doi: 10.1109/TIFS.2021.3118894.

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

[33] L. Yang, X. Liu, G. Yang, J. Wang, and Y. Yin, "Small-Area Finger Vein Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1914–1925, 2023, doi: 10.1109/TIFS.2023.3258252.

[34] P. Zhao *et al.*, "Single-Sample Finger Vein Recognition via Competitive and Progressive Sparse Representation," *IEEE Trans. Biometrics, Behav. Identity Sci.*, pp. 1–1, 2022, doi: 10.1109/TBIOM.2022.3226270.

[35] S. Li, R. Ma, L. Fei, and B. Zhang, "Learning Compact Multirepresentation Feature Descriptor for Finger-Vein Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 1946–1958, 2022, doi: 10.1109/TIFS.2022.3172218.

[36] C.-G. Ri, K.-I. Ri, and J.-H. Ri, "Robust and Efficient Finger Vein Recognition Using Binarized Difference of Gabor Jet on Grid," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 177–189, 2023, doi: 10.1109/TIFS.2022.3217396.

[37] G. Wang, C. Sun, and A. Sowmya, "Multi-weighted co-occurrence descriptor encoding for vein recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 375–390, 2019.

[38] R. S. Kuzu, E. Maiorana, and P. Campisi, "Vein-Based Biometric Verification Using Densely-Connected Convolutional Autoencoder," *IEEE Signal Process. Lett.*, vol. 27, pp. 1869–1873, 2020.

[39] B. Hou and R. Yan, "Convolutional autoencoder model for finger-vein verification," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 5, pp. 2067–2074, 2019.

[40] R. S. Kuzu, E. Maiorana, and P. Campisi, "On the intra-subject similarity of hand vein patterns in biometric recognition," *Expert Syst. Appl.*, vol. 192, p. 116305, 2022.

[41] J. Huang, A. Zheng, M. S. Shakeel, W. Yang, and W. Kang, "FVFSNet: Frequency-Spatial Coupling Network for Finger Vein Authentication," *IEEE Trans. Inf. Forensics Secur.*, vol. 18, pp. 1322–1334, 2023, doi: 10.1109/TIFS.2023.3238546.

[42] T. Chai, J. Li, S. Prasad, Q. Lu, and Z. Zhang, "Shape-driven lightweight CNN for finger-vein biometrics," *J. Inf. Secur. Appl.*, vol. 67, p. 103211, Jun. 2022, doi: 10.1016/j.jisa.2022.103211.

[43] R. S. Kuzu, E. Maiorana, and P. Campisi, "Loss functions for CNN-based biometric vein recognition," in *2020 28th European Signal Processing Conference (EUSIPCO)*, 2021, pp. 750–754.

[44] B. Hou and R. Yan, "ArcVein-Arccosine Center Loss for Finger Vein Verification," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021.

[45] A. Krishnan, T. Thomas, and D. Mishra, "Finger Vein Pulsation-Based Biometric Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 5034–5044, 2021, doi: 10.1109/TIFS.2021.3122073.

[46] Z. Zhang, F. Zhong, and W. Kang, "Study on Reflection-Based Imaging Finger Vein Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2298–2310, 2022, doi: 10.1109/TIFS.2021.3093791.

[47] D. Zhong, H. Shao, and X. Du, "A hand-based multi-biometrics via deep hashing network and biometric graph matching," *IEEE Trans. Inf. Forensics Secur.*, vol. 14, no. 12, pp. 3140–3150, 2019.

[48] H. Ren, L. Sun, J. Guo, and C. Han, "A Dataset and Benchmark for Multimodal Biometric Recognition Based on Fingerprint and Finger Vein," *IEEE Trans. Inf. Forensics Secur.*, vol. 17, pp. 2030–2043, 2022, doi: 10.1109/TIFS.2022.3175599.

[49] S. Li and B. Zhang, "Joint Discriminative Sparse Coding for Robust Hand-Based Multimodal Recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 16, pp. 3186–3198, 2021, doi: 10.1109/TIFS.2021.3074315.

[50] W. Yang, Z. Chen, J. Huang, and W. Kang, "A Novel System and Experimental Study for 3D Finger Multibiometrics," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 4, no. 4, pp. 471–485, Oct. 2022, doi: 10.1109/TBIOM.2022.3181121.

[51] H. O. Shahreza and S. Marcel, "Towards Protecting and Enhancing Vascular Biometric Recognition methods via Biohashing and Deep Neural Networks," *IEEE Trans. Biometrics, Behav. Identity Sci.*, 2021.

[52] W. Yang, S. Wang, J. Hu, G. Zheng, J. Yang, and C. Valli, "Securing deep learning based edge finger vein biometrics with binary decision diagram," *IEEE Trans. Ind. Informatics*, vol. 15, no. 7, pp. 4244–4253, 2019.

[53] C. Kauba, S. Kirchgasser, V. Mirjalili, A. Uhl, and A. Ross, "Inverse Biometrics: Generating Vascular Images from Binary Templates," *IEEE Trans. Biometrics, Behav. Identity Sci.*, 2021.

[54] F. Ahmad, L.-M. Cheng, and A. Khan, "Lightweight and privacy-preserving template generation for palm-vein-based human recognition," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 184–194, 2019.

[55] S. Kirchgasser, C. Kauba, Y.-L. Lai, J. Zhe, and A. Uhl, "Finger Vein Template Protection Based on Alignment-Robust Feature Description and Index-of-Maximum Hashing," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 2, no. 4, pp. 337–349, Oct. 2020, doi: 10.1109/TBIOM.2020.2981673.

[56] C. Kauba *et al.*, "Towards practical cancelable biometrics for finger vein recognition," *Inf. Sci. (Ny).*, vol. 585, pp. 395–417, Mar. 2022, doi: 10.1016/j.ins.2021.11.018.

[57] J. Schuiki, M. Linortner, G. Wimmer, and A. Uhl, "Attack Detection for Finger and Palm Vein Biometrics by Fusion of Multiple Recognition Algorithms," *IEEE Trans. Biometrics, Behav. Identity Sci.*, vol. 4, no. 4, pp. 544–555, Oct. 2022, doi: 10.1109/TBIOM.2022.3212836.

[58] A. Dosovitskiy, P. Fischer, J. T. Springenberg, M. Riedmiller, and T. Brox, "Discriminative unsupervised feature learning with exemplar convolutional neural networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 9, pp. 1734–1747, 2015.

[59] Z. Wu, Y. Xiong, S. X. Yu, and D. Lin, "Unsupervised feature learning via non-parametric instance discrimination," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3733–3742.

[60] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 9729–9738.

[61] M. Ye, X. Zhang, P. C. Yuen, and S.-F. Chang, "Unsupervised embedding learning via invariant and spreading instance feature," in *Proceedings of the IEEE Conference on computer vision and pattern recognition*, 2019, pp. 6210–6219.

[62] X. Chen and K. He, "Exploring Simple Siamese Representation Learning," *arXiv Prepr. arXiv2011.10566*, 2020.

[63] L. Zhang, Z. Cheng, Y. Shen, and D. Wang, "Palmprint and palmvein recognition based on DCNN and a new large-scale contactless palmvein dataset," *Symmetry (Basel).*, vol. 10, no. 4, p. 78, 2018.

[64] C. Xie and A. Kumar, "Finger vein identification using convolutional neural network and supervised discrete hashing," in *Deep Learning for Biometrics*, Springer, 2017, pp. 109–132.

[65] Y. Fang, Q. Wu, and W. Kang, "A novel finger vein verification system based on two-stream convolutional network learning," *Neurocomputing*, vol. 290, pp. 100–107, 2018.

[66] H. G. Hong, M. B. Lee, and K. R. Park, "Convolutional neural network-based finger-vein recognition using NIR image sensors," *Sensors*, vol. 17, no. 6, p. 1297, 2017.

[67] W. Kim, J. M. Song, and K. R. Park, "Multimodal biometric recognition based on convolutional neural network by the fusion of finger-vein and finger shape using near-infrared (NIR) camera sensor," *Sensors*, vol. 18, no. 7, p. 2296, 2018.

[68] H. Huang, S. Liu, H. Zheng, L. Ni, Y. Zhang, and W. Li, "DeepVein: Novel finger vein verification methods based on deep convolutional neural networks," in *2017 IEEE International Conference on Identity, Security and Behavior Analysis (ISBA)*, 2017, pp. 1–8.

[69] W. Yang, W. Luo, W. Kang, Z. Huang, and Q. Wu, "FVRAS-net: An embedded finger-vein recognition and AntiSpoofing system using a unified CNN," *IEEE Trans. Instrum. Meas.*, vol. 69, no. 11, pp. 8690–8701, 2020.

[70] R. S. Kuzu, E. Piciucco, E. Maiorana, and P. Campisi, "On-the-Fly Finger-Vein-Based Biometric Recognition Using Deep Neural Networks," *IEEE Trans. Inf. Forensics Secur.*, vol. 15, pp. 2641–2654, 2020.

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

[72] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4401–4410.

[73] H. Wang *et al.*, "Cosface: Large margin cosine loss for deep face recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5265–5274.

[74] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[75] E. C. Lee, H. C. Lee, and K. R. Park, "Finger vein recognition using minutia-based alignment and local binary pattern-based feature extraction," *Int. J. Imaging Syst. Technol.*, vol. 19, no. 3, pp. 179–186, 2009.

[76] M. S. M. Asaari, S. A. Suandi, and B. A. Rosdi, "Fusion of band limited phase only correlation and width centroid contour distance for finger based biometrics," *Expert Syst. Appl.*, vol. 41, no. 7, pp. 3367–3382, 2014.

[77] P. Tome, M. Vanoni, and S. Marcel, "On the vulnerability of finger vein recognition to spoofing," in *2014 international conference of the biometrics special interest group (BIOSIG)*, 2014, pp. 1–10.

[78] B. T. Ton and R. N. J. Veldhuis, "A high quality finger vascular pattern dataset collected using a custom designed capturing device," in *2013 International Conference on Biometrics (ICB)*, Jun. 2013, pp. 1–5, doi: 10.1109/ICB.2013.6612966.

[79] Y. Yin, L. Liu, and X. Sun, "SDUMLA-HMT: a multimodal biometric database," in *Chinese Conference on Biometric Recognition*, 2011, pp. 260–268.

[80] P. Tome and S. Marcel, "On the vulnerability of palm vein recognition to spoofing attacks," in *2015 International Conference on Biometrics (ICB)*, May 2015, pp. 319–325, doi: 10.1109/ICB.2015.7139056.

[81] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[82] J. H. Choi, W. Song, T. Kim, S.-R. Lee, and H. C. Kim, "Finger vein extraction using gradient normalization and principal curvature," in *Image Processing: Machine Vision Applications II*, 2009, vol. 7251, p. 725111.

[83] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[84] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2223–2232.

[85] L. Zhang, L. Li, A. Yang, Y. Shen, and M. Yang, "Towards contactless palmprint recognition: A novel device, a new benchmark, and a collaborative representation based identification approach," *Pattern Recognit.*, vol. 69, pp. 199–212, 2017.

Wei-Feng Ou received his B.E. degree in Information Engineering from Guangdong University of Technology in 2013, his M.E. degree in Signal & Information Processing from South China University of Technology in 2016 and his Ph.D. degree in Electrical Engineering from City University of Hong Kong in 2021. He was with Huawei as a Research & Development Engineer from 2016 to 2018 and was with SenseTime as an Algorithm Researcher from 2021~2023. He is currently a Lecturer with the School of Computer Science and Technology, Dongguan University of Technology. His research interests include biometrics, computer vision and deep learning.

Lai-Man Po (M'92–SM'09) received the B.S. and Ph.D. degrees in electronic engineering from the City University of Hong Kong, Hong Kong, in 1988 and 1991, respectively. He has been with the Department of Electronic Engineering, City University of Hong Kong, since 1991, where he is currently an Associate Professor of Department of Electrical Engineering. He has published over 170 technical journal and conference papers with more than 8,000 citations. His research interests include image and video coding with an emphasis deep learning based computer vision algorithms. Dr. Po is a member of the Technical Committee on Multimedia Systems and Applications and the IEEE Circuits and Systems Society. He was the Chairman of the IEEE Signal Processing Hong Kong Chapter in 2012 and 2013. He was an Associate Editor of HKIE Transactions in 2011 to 2013. He also served on the Organizing Committee, of the IEEE International Conference on Acoustics, Speech and Signal Processing in 2003, and the IEEE International Conference on Image Processing in 2010.

Xiu-Feng Huang received the B.E. degree in Information Engineering from the Shenzhen University in 2015, the M.Sc. degree in Electrical Engineering from the City University of Hong Kong in 2019. He was with Pantheon Lab as a Data Scientist from 2019 to 2021, and was with SmartMore as a Research and Development Engineer from 2021 to 2023. He is currently working as a Researcher Assistant with City University of Hong Kong. His research interests include computer vision and deep learning.

Wing-Yin Yu (S'21) received the B.Eng. degree in Information Engineering from City University of Hong Kong, in 2019. He is currently pursuing the Ph.D. degree at Department of Electrical Engineering at City University of Hong Kong. His research interests are deep learning and computer vision.

Yu-Zhi Zhao (S'19) received the B.Eng. degree in Electronic and Information Engineering from Huazhong University of Science and Technology (HUST), 2018 and the Ph.D. degree in Electronic Engineering from City University of Hong Kong (CityU) in 2023. His research interests include low-level vision, computational photography, generative models, and representation learning. He serves as a peer-reviewer in international conferences and journals such as CVPR, ICCV, ECCV, WACV, ACCV, ICASSP, ICIP, and several IEEE Transactions.