IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 12 February 2023, accepted 28 February 2023, date of publication 3 March 2023, date of current version 8 March 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3252009

RESEARCH ARTICLE

Vision Transformers for Vein Biometric Recognition

RAUL GARCIA-MARTIN AND RAUL SANCHEZ-REILLO, (Senior Member, IEEE)
Electronic Technology Department, University Carlos III of Madrid, 28911 Leganés, Spain
Corresponding author: Raul Garcia-Martin (raulgarc@ing.uc3m.es)

ABSTRACT In October 2020, Google researchers present a promising Deep Learning architecture paradigm for Computer Vision that outperforms the already standard Convolutional Neural Networks (CNNs) on multiple image recognition state-of-the-art datasets: Vision Transformers (ViTs). Based on the self-attention concept inherited from Natural Language Processing (NLP), this new structure surpasses the CNN image classification task on ImageNet, CIFAR-100, and VTAB, among others, when it is fine-tuned (Transfer Leaning) after a previous pre-training on larger datasets. In this work, we confirm this theory and move one step further over the CNN structures applied for Vascular Biometric Recognition (VBR): to the best of our knowledge, we introduce for the first time multiple pure pre-trained and fine-tuned Vision Transformers in this evolving biometric modality to address the challenge of the limited number of samples in VBR datasets. For this purpose, the ViTs have been trained to extract unique image features on the ImageNet-1k and ImageNet-21k and then fine-tuned for the four main existing VBR variants, i.e., finger, palm, hand dorsal, and wrist vein areas. Fourteen existing vascular datasets have been used to perform the vein identification task in the four previously mentioned modalities, based on the True-Positive Identification Rate (TPIR) and 75-25% train-test sets obtaining the following results: HKPU (99.52%), and FV-USM (99.1%); Vera (99.39%), and CASIA (96.00%); Bosphorus (99.86%); PUT-wrist (99.67%), and UC3M-CV1+CV2 (99.67%). Furthermore, we introduce UC3M-CV3: a hygienic contactless wrist database collected on smartphones and consisting of 4800 images from 100 different subjects. The promising results show the Vision Transformer's versatility in VBR under Transfer Learning and reinforce this new Neural Network architecture paradigm.

INDEX TERMS Vision transformers, vein biometric recognition, deep learning, convolutional neural networks, finger veins, transfer learning, machine learning, artificial intelligence, biometrics on mobile devices, contactless wrist vascular database, hand palm identification.

I. INTRODUCTION

Transformers have become the leading state-of-the-art Deep Learning solution for Natural Language Processing (NLP). The goal of NLP is to provide machines with the ability to understand spoken and written human language. This successful pure attention-based solution was proposed for machine translation by Vaswani et al. [1] to solve the memory constraint issue of the Recurrent Neural Networks (RNNs) that compute the text sequentially. Transformers have solved this issue, becoming the first option for the different NLP tasks.

In October 2020, in an attempt to explore new Neural Network architectures in Computer Vision (CV), Dosovitskiy et al. [2] proposed and adapted, for the first time, the pure Transformer attention-based structure for image recognition tasks. These new Neural Networks are known as Vision Transformers (ViTs).

Despite the fact that ViTs, in contrast with NLP Transformers, are not designed to understand and relate sequences of words (tokens), they have demonstrated the versatility of the Transformer concept by obtaining excellent performances

The associate editor coordinating the review of this manuscript and approving it for publication was Carmelo Militello.

This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

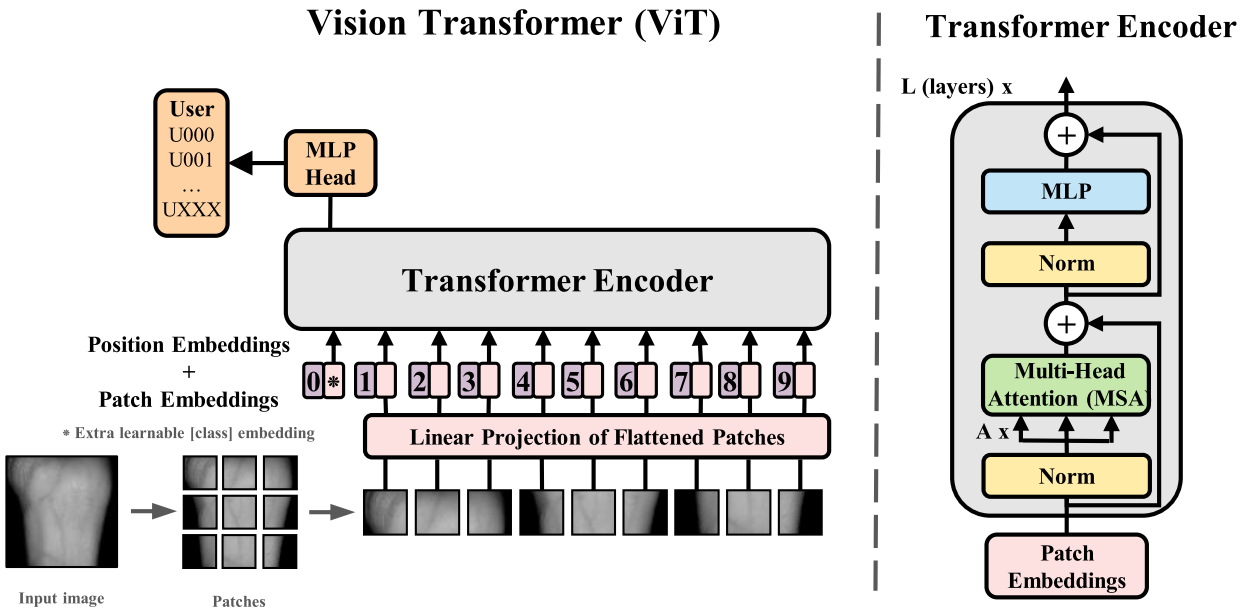

FIGURE 1. Original pure Vision Transformer (ViT) model [2] adapted for Vascular Biometric Recognition by applying Transfer Learning using dual pre-training (on ImageNet-21k and ImageNet-1k datasets) and fine-tuning (on fourteen finger, palm, dorsal, and wrist vein databases: HKPU, UTFVP, FV-USM, PLUSVein-Contactless Finger Vein Database, and NUPT-FPV; CASIA Multi-Spectral, PUT, VERA Palmvein, and PLUSVein-Contactless Palm Vein Database; Bosphorus Hand Vein Database; and PUT, UC3M-CV2, UC3M-CV1+CV2 and UC3M-CV3).

on CV tasks. ViTs have outperformed previous CNN results on ImageNet-1k [3] (ImageNet Large Scale Visual Recognition Challenge, ILSVRC-2012 [4]) ImageNet-21k [4], CIFAR-100 [5], and VTAB [6] after a Transfer Learning (TL) process consisting of a pre-training on a large-scale dataset, JFT-300M [7], and a fine-tuned procedure. Compared to state-of-the-art CNN solutions, the promising results shown by the ViT architectures in CV image recognition tasks inspire, among other factors, the presentation of the current work.

A. MOTIVATION

We present Vision Transformers for Vascular Biometric Recognition in an attempt to promote and keep VBR updated as one of the most promising biometric solutions in terms of security and respect for user privacy. The current study is the first to apply Transfer Learning to Vascular Biometric Recognition (VBR) by utilizing pre-training and fine-tuning on Vision Transformers (ViTs) [2] to showcase the superior performance of ViTs over this modality. In order to validate this concept [2], whose results exceed the state-of-the-art CNN solutions, and to demonstrate the ViTs image recognition versatility over other image recognition tasks, the presented self-attention models have been pre-trained on the midsized ImageNet-1k and ImageNet-21k datasets and have been fine-tuned on fourteen smaller VBR databases. We proposed this solution following the idea stated in [2] of pre-training in a more extensive dataset, the non-public Google JTF-300M [7] in [2], than the one used for fine-tuning, the public ImageNet dataset in [2].

As a final reason behind this study, we would like to reinforce the latest Transfer Learning methods as part of the key Deep Learning solutions for image recognition when a small-size dataset is used, as is the case in the VBR field (less than 100 samples per class). Furthermore, it is important to point out that we have already shown the excellent results of TL methods for the VBR field in [8], and they have also been extensively investigated for other image recognition purposes. We would also like to inspire other researchers focused on other imaging-based biometric modalities, such as Facial and Iris recognition, to follow TL solutions and, specifically, TL Vision Transformers architectures.

B. CONTRIBUTIONS

In this study, for the first time to the limits of our knowledge, Vision Transformers have been widely and successfully applied in Vascular Biometric Recognition by presenting a Transfer Learning process consisting of pre-training on public image recognition datasets and fine-tuning on fourteen state-of-the-art databases of the four main VBR variants.

We independently pre-train four different ViTs (pure version in Fig. 1) on ImageNet-1k and ImageNet-21k and fine-tune on the following state-of-the-art vein datasets for finger, palm, hand dorsal, and wrist: HKPU, UTFVP, FV-USM, PLUSVein-Contactless Finger Vein Database, and NUPT-FPV; CASIA Multi-Spectral, PUT, VERA Palmvein, and PLUSVein-Contactless Palm Vein Database; Bosphorus Hand Vein Database; and PUT, UC3M-CV1, and UC3M-CV2.

Moreover, to contribute to the Vascular Biometric Recognition field, we collect and introduce a new contactless wrist vein dataset, UC3M-CV3 (University Carlos III of Madrid-Contactless Version 3), providing a biometric hygienic modality for these COVID-19 demanding times.

The 4800 infrared images of this database have been acquired from 100 people (200 independent wrists) using two non-modified smartphones, as we already showed in [8] and [9] with UC3M-CV2. In addition, we introduce a CV automatic acquisition method based on the Haar Cascade algorithm [10] called Wrist Vascular Haar Cascade (WVHC) to simplify the collection process of the 4800 images of UC3M-CV3.

In light of the promising results discussed according to the ISO/IEC 19795-1 standard [11] and verified over several vein datasets, we can confirm that ViT Neural Networks should be considered as a new successful solution in the state-of-the-art of VBR when they are trained following a Transfer Learning procedure. Most of the TL CNN solutions found in the VBR state of the art, including our previous research [8], have been outperformed.

C. RELATED WORK

ViT architectures have started to be explored since their pure introduction in 2020 [1]. These Neural Network structures, based on attention mechanisms, are being applied and studied in different CV tasks such as image classification [12] and object detection [13]. Additionally, they can be found in more challenging missions like image segmentation (semantic, instance, and panoptic) [14], [15], [16] or even image generation with the introduction of ViTGAN [17].

Following the image classification target, pre-trained ViTs have demonstrated their promising performance in some biomedical research (COVID-19 detection, pulmonary nodule characterization, and cancer detection, among others) [18], [19], and a reduced number of biometrics studies [20], [21]. To the best of our knowledge, the proposed study is not only a pioneering Vascular ViT work but also represents one of the earliest biometric recognition solutions.

Before analyzing the current Deep Learning (DL) algorithms for VBR, the primary state-of-the-art vein recognition datasets have been reviewed in the next section for the four main variants: finger, palm, hand dorsal, and wrist.

1) DATASETS

Even though we have focused our efforts on the algorithm and computational performance, Table 1 reports not only the main dataset features, i.e., modality, number of subjects, samples, sessions, images, and year but also two important factors of the research VBR hardware efforts and advances: system miniaturization-integration and comfort-hygiene usability. Both key aspects nowadays are considered in this work and have been summarized as contactless usability and processing hardware in the last column of Table 1.

For a deeper study of all existing vein datasets, we strongly recommend the Handbook of Vascular Biometrics [22], where most of these datasets (among others) are summarized, including even the least-explored VBR modality: retinal vascular recognition.

1) Finger vein datasets: This vein recognition modality is the most relevant and explored method in research with numerous public [23], [24], [25], [26], [27], [28] and privately-distributed databases [29] acquired since (even before) 2010, as shown in the first four rows of Table 1.
This VBR variant is commonly based on near-infrared light transmission, unlike the rest of the VBR modalities that follow the light reflection principle. Near-infrared light transmission requires a more complex system shape to prevent all contact between user and sensor. That is probably why most state-of-the-art datasets were collected with physical contact.
However, various recent studies have surpassed this constraint by presenting non-contact solutions [30], [31]. Furthermore, in March 2021, Hitachi introduced its commercial contactless finger vein device based on light reflection: ''Finger Vein Biometric Authentication Unit (C-1)'' [32].
From Table 1, it is important to mention that even though state-of-the-art vascular recognition techniques are deeply immersed in the Deep Learning paradigm, which should be based on massive datasets, the number of images per class does not seem to increase considerably in the finger vein world. However, new public datasets such as [28] try to change this trend by including 20 captures per finger vein class and combining them with the corresponding 20 fingerprint classes.
Most of the state-of-the-art datasets and all of those presented in Table 1 were acquired using a laptop or PC, which requires a greater focus on the acquisition hardware than the processing hardware development and the full system integration.
To test the proposal ViT architectures, we have employed the public datasets HKPU [23], UTFVP [25], FV-USM [26], PLUSVein-Contactless Hand Vein Database [27], and NUPT-FPV [28].
2) Palm vein datasets: As Table 1 shows, palm VBR (based on near-infrared light reflection) has demonstrated constant contactless usability with VERA PalmVein [33] and PLUSVein-Contactless Hand Vein Database [27] since 2015. This is in sharp contrast with finger VBR, due to the hardware constraints previously cited.
The number of subjects is similar to finger vein datasets, varying approximately between 50 and 100 individuals but with fewer independent recognition areas for each. A higher number of individual users/areas per subject (usually 6 fingers instead of 2 hands) could explain why palm vein datasets typically contain more images/samples per class: it is less tedious for each volunteer to be enrolled by a palm than 3 fingers. A recent study published in 2021 [34] provides 47 samples per class (and even in different contactless hand positions) in concordance with the current DL times.

TABLE 1. Vascular biometric recognition (VBR) datasets.

	Dataset	Subjects	Independent users per subject	Sessions - Samples per session	Images per class	Images	Year	Contactless - Processing hardware
Finger	HKPU (vein set) [23]	105	2 (index and middle, left)	2 6	12	2520	2010	YES PC
	SDUMLA [24]	106	6 (ring, index, and middle)	1 6	6	3816	2010	NO PC
	UTFVP [25]	60	6 (ring, index, and middle)	2 2	4	1440	2012	NO PC
	FV-USM [26]	123	2 (index and middle, left)	2 6	12	5904	2013	NO PC
	MMCBNU_6000 (vein set) [29]	100	2 (index and middle, left)	1 10	10	6000	2013	NO PC
	PLUSVein-Contactless [27]	42	2 (index and middle, left)	1 5	5	1260	2018	YES PC
	NUPT-FPV (vein set) [28]	140	2 (index and middle, left)	2 10	20	16800	2022	NO PC
Palm	CASIA (vein set) [52]	100	2	2 6	12	2400	2007	LIMITED PC
	PUT Palm [38]	50	2	3 4	12	1200	2011	NO PC
	VERA PalmVein [33]	110	2	2 5	10	2200	2015	YES PC
	PLUSVein-Contactless [27]	42	2	1 5	5	420	2018	YES PC
	Liang *et al.* [34] (vein set)	105	2	3 10, 10, 27	47	9870	2021	YES PC
Hand dorsal	Bosphorus Hand Vein [35]	100	1 (left)	4 3	12	1200	2011	NO PC
	PROTECTVein [36]	40	2	1 30	30	2400	2018	NO PC + Smartphone
	FYO-Dorsal (**) [37]	160	2	2 1	2	640	2020	NO PC
Wrist	PUT Wrist [38]	50	2	3 4	12	1200	2011	NO PC
	Raghavendra *et al.* [39]	50	2	2 5	10	1000	2016	NO PC
	UC3M-CV1 [40]	50	2	2 6	12	1200	2020	YES Raspberry Pi 4
	UC3M-CV2 [9]	50	2	2 12	24	2400	2020	YES Smartphones
	UC3M-CV3 (proposed)	100	2	**2** **12**	**24**	**4800**	**2021**	**YES** **Smartphones**

(*All table) A detailed description of the fourteen datasets employed in this research is provided in Appendix A.

(**) FYO is a multimodal dataset that also includes the palm and wrist variants.

All the Table 1 datasets for palm veins are publicly available except for Liang et al. [34] (image set) and have been tested in this work.

3) Hand dorsal vein datasets: Databases using the hand dorsal are less common in the state-of-the-art VBR. The number of images per class is again higher than in finger datasets and similar to palm datasets, with around 10-30 images. To our knowledge, most of the existing vein dorsal datasets were acquired using a contact system.
   All the datasets reported in Table 1 are publicly available. However, only the Bosphorus Hand image set [35] has been used in this study because PROTECTVein [36] was not obtained, and FYO-Dorsal [37], which is a subset of the multimodal FYO cannot be used, in our opinion, for unimodal vein recognition (2 samples per dorsal class).
4) Wrist vein datasets: Finally, in this Vascular Biometric Recognition datasets summary, Table 1 shows the current wrist VBR datasets [38], [39], [40]. As the last three rows infer, we are focusing on promoting not only this VBR modality but also the contactless human-system interaction and the hardware miniaturization/integrability by even embedding a DL (CNN) model into several smartphones [8] and exploring VBR ViTs (the proposed study) for the first time in this modality.
   In this research, the public PUT database [38] has been used along with UC3M-CV1 [39], UC3M-CV2 [9], UC3M-CV1+CV2 [40], and UC3M-CV3 (proposed in this work).

2) RECOGNITION ALGORITHMS

Although Convolutional Neural Networks are not included in the main line of research explored in this work, state-of-the-art VBR algorithms are exposed in this section highlighting the two current Deep Learning trends followed so far: the already mature VBR paradigm of CNN architectures (present) and the still unexplored self-attention structures (future).

1) Convolutional Neural Networks: At the expense of traditional Machine Learning, it is well known that Deep Learning algorithms, based on Neural Networks, have conquered not only Computer Vision (by using CNNs) but also every process that requires an intelligent algorithm: from text translation using Recurrent Neural Networks (RNNs, NLP) to speech recognition using MultiLayer Perceptron (MLP) Neural Networks.
   State-of-the-art VBR solutions have been explored by moving from basic vein-designed CNNs with a few convolutional layers, through more advanced architectures, such as VGG16 and VGG19 [41] with 2 and 3 blocks of convolutional levels, to one of the most sophisticated and deepest networks, ResNet [42], which consists of more than 152 layers. ResNet established the micro-architecture concept through the residual model.
   This path, strongly influenced by the CV image recognition advances through challenges such as the ImageNet Large Scale Visual Recognition Challenge (ILSVRC) [4], was established around 2017, as Table 2 infers. In this sense, Liu et al. [43] proposed a basic CNN designed by them, consisting of 5 layers and based on the AlexNet architecture [44]. This CNN obtained a high performance (Equal Error Rate, EER = 0.80 % and True Positive Identification Rate, TPIR = 99.53 %) on the SDUMLA [24] finger vein dataset (87-13 % train-test). In this vein benchmark, Hong et al. [45] tested the VGG16 architecture by applying Transfer Learning (fine-tuning) supported by the ImageNet dataset.
   The EER achieved was 3.91 % for a 2.5-97.5 % test-train (data augmentation) closed scenario. Jhong et al. [46] also presented a VGG16-inspired solution for the palm modality using their private database. In this case, they achieved good biometric performance with TPIR = 96.54 % but with what can be considered a high training percentage of 90-10 train-test set for a dataset of 30 subjects and ten images per palm.
   Along this journey, while the image recognition architectures were going deeper, VBR works followed this trend by exploring the VGG19 architecture (older cousin of VGG16) in [8] and the also previously cited research proposed by Hong et al. [45] (worse results than VGG16).In 2019, Al-Johania et al. [47] obtained, for the dorsal Bosphorus dataset [35], a TPIR of 99.0 %, 99.25 %, and 95.51 %, respectively, for the AlexNet, VGG16, and VGG19 in an 80-20 train-test scenario. Along with [8], these are some of the few studies that widely argue that the Transfer Learning concept applied to VBR is an excellent solution to the lack of VBR data.
   Table 2 also includes two works from 2018 and 2019, [48] and [49], where ResNet50/ResNet101 and DenseNet161 [50] were implemented on the public SDUMLA finger vein dataset. Other studies, such as [51] on the CASIA [52] image set, proposed their own-designed CNNs.
   As we propose a different Neuronal Network paradigm for the VBR world based on Vision Transformers, in the next section and Table 3, our approach is closely compared to the four existing vein self-attention studies. To evaluate the results with the previous DL state-of-the-art solutions, the last row of Table 2 includes the obtained performance.
   It's noteworthy that most literature works train CNN models from scratch, as our analysis shows. However, our proposed Transfer Learning results can be compared against [45], [47] and directly against [9] by using the same datasets and train-test split. As expected, it seems the pre-trained ViTs surpass

TABLE 2. State-of-the-art deep learning solutions for VBR.

Study	Year	Vascular modality	Dataset	DL technique	Train-Test (%)	Biometric performance
Liu *et al.* [43]	2017	Finger	SDUMLA	Own-designed CNN of 5 layers based on AlexNet	87-13	EER = 0.80 % TPIR = 99.53 %
Hong *et al.* [45]	2017	Finger	Own 1 and 2 (private) SDUMLA + ImageNet	VGG16 Transfer Learning (feature extractor)	2.5-97.5 (data augmentation)	EER = 3.91 %
Kim *et al.* [48]	2018	Finger + Finger shape	SDUMLA HKPU (vein set)	ResNet50/ResNet101	50-50	EER = 2.34 % EER = 0.79 %
Song *et al.* [49]	2019	Finger	SDUMLA HKPU (vein set)	DenseNet161	50-50	EER = 2.35 % EER = 0.33 %
Jhong *et al.* [46]	2020	Palm	Own (private)	VGG16	90-10	TPIR = 96.54 %
Obayya *et al.* [51]	2020	Palm	CASIA (vein set)	Own CNN architecture	80-20	EER = 0.07 % TPIR = 99.40 %
Al-Johania *et al.* [47]	2019	Hand dorsal	Bosphorus and Dr. Badawi + ImageNet	AlexNet, VGG16, and VGG19 Transfer Learning (feature extractor and fine-tuning)	80-20	TPIR = 99.25 % TPIR = 100.00 %
Garcia-Martin *et al.* [9]	2021	Wrist	UC3M-CV1 UC3M-CV2 PUT + ImageNet	VGG16, VGG19, ResNet50 and ResNet152 Transfer Learning (feature extractor)	50-50	EER = 0.38 % TPIR = 98.67 % EER = 0.78 % TPIR = 97.67 %
Proposed	**2023**	**Finger Palm Dorsal Wrist**	**14 finger, palm, hand dorsal and wrist datasets * + ImageNet**	**Vision Transformers - Transfer Learning (fine tuning)**	**50-50**	**TPIR = 99.50 % TPIR = 98.25 % TPIR = 98.72 % TPIR = 99.08 %**

(*) Biometric results provided in this table for, respectively, PUT Wrist, UC3M-CV2, UC3M-CV1+CV2, and UC3M-CV3.

TABLE 3. State-of-the-art self-attention-based solutions for VBR.

Study	Year	Vascular modality	Dataset	DL technique	Train-Test (%)	Biometric performance
Huang *et al.* [55]	2020	Finger	SDUMLA and THU-FVFDT2	Hybrid spatial-attention CNN	83-17 50-50	TPIR = 99.53 % TPIR = 98.64 %
Huang *et al.* [57]	2021	Finger	SDUMLA, MMCBNU_6000, FV-USM, and their own FV-SCUT	Hybrid joint-attention CNN	80-20 75-25 67-33 50-50	EER = 0.35 % EER = 0.08 % EER = 0.34 % EER = 0.49 %
Lu *et al.* [58]	2021	Finger	SDUMLA, FV-USM, and MMCBNU_6000	Pure own Vision Transformers	84-16 80-20 67-33	TPIR = 93.50 % TPIR = 91.84 % TPIR = 91.75 %
Huang *et al.* [59]	2022	Finger	HKPU, SDUMLA, UTFVP, FV-USM, MMCBNU_6000, THU-FVFDT, PLUSVEIN FNGER, and SCUT-FV	Pure Vision Transformers	80-20	EER = 2.37 % EER = 1.50 % EER = 1.97 % EER = 0.44 % EER = 0.69 % EER = 3.60 % EER = 2.08 % EER = 1.56 %
Proposed	**2023**	**Finger Palm Dorsal Wrist**	**14 finger, palm, hand dorsal and wrist datasets * + ImageNet**	**Pure and hybrid Vision Transformers - Transfer Learning (fine tuning)**	**50-50**	**TPIR = 99.50 % TPIR = 98.25 % TPIR = 98.72 % TPIR = 99.08 %**

(*) Biometric results provided in this table for, respectively, PUT Wrist, UC3M-CV2, UC3M-CV1+CV2, and UC3M-CV3.

the biometric performance of all the previously cited works.

2) Self-Attention-based Networks: Since 2018, after the success of Transformers in NLP induced by the release of "Attention Is All You Need" [1] (2017), some Computer Vision works, [53] and [54], focused their efforts on combining CNN architectures with attention techniques or entirely replaced the former networks by the latter structures.
Along the same lines, regarding VBR, Huang et al. [55] introduced, for the first time to the best of our knowledge, the attention concept for vein identification. In 2020, Huang et al., presented a CNN (inspired in ResNet50 [42]) combined with a spatial attention module (based on the CNN U-Net [56]) inserted into the first layers of the entire architecture. In the following year, in the wake of the previous idea of introducing spatial information to the network, Huang et al. [57] presented a CNN-like structure along with vascular joint attention micro-architecture. In this case, the structure was used for finger vein authentication in a closed-set scenario and training from scratch (80-20 % train-test ratio). The results obtained were the following: EER 0.35 %, 0.08, 0.34 %, and 0.49 %, respectively, on SDUMLA-HMT, MMCBNU_6000, FV-USM, and their self-built FV-SCUT.
But it was not until October 2020 that Dosovitskiy et al. [2] introduced a pure self-attention Computer Vision algorithm in "An Image is Worth 16×16 Words: Transformers for Image Recognition at Scale": the Vision Transformer.
In 2021, after this revolutionary approach and following the promising results shown in different CV tasks, Lu et al. [58] proposed, for the first time as far as we know, a Neuronal Network consisting of a Vision Transformer architecture. They trained four ViTs from scratch in a closed scenario on the public finger vein datasets FV-USM, SDUMLA, and MMCBNU_6000, obtaining TPIRs of 93.5 %, 91.75 %, and 91.84 %, respectively, for a train-test percentage of 84-16 %, 67-33 %, and 80-20 %. These ViT results, which do not seem to improve on the previous state-of-the-art CNN performance but open a new VBR paradigm, could be enhanced using TL techniques, as we present in an innovative fashion in the following sections of this work.
In May 2022, Huang et al. [59] widely demonstrated excellent results for the first time by applying ViTs on several public databases to finger vein authentication in this case.
After these two ViT from-scratch approaches, in this work, we pioneered a Vision Transformer solution not only for finger vein identification and authentication but also for the palm, dorsal, and wrist modalities combating the lack of vein data by using Transfer Learning (pre-training in a different image classification benchmark + fine-tuning).

II. VISION TRANSFORMERS (ViTs)

The following two sections detail the original ViT architecture and the differences between Transfer Learning on CNNs and ViTs.

A. ARCHITECTURE

Fig. 1 shows the model overview of the pure ViT presented in [2] and adapted for VBR in the current work. First of all, the input image of H x W size is divided into 2D square patches (x_p) of P x P size and C channels. Therefore the resulting number of these square patches is $N = \frac{H \cdot W}{P^2}$:

$$\mathbf{x}_p \in \mathbb{R}^{N \times (P^2 C)}, \tag{1}$$

where C is the number of channels in the input image. For example, as Fig. 1 exemplified, for a 240 × 240 square image (H x W) with a patch size of 80 × 80 pixels (P x P), there are 9 patches ($N = \frac{H \cdot W}{P^2} = \frac{240 \cdot 240}{80^2} = 9$) of size 80 × 80 x 3 (C = 3, RGB image). Then these patches are flattened into a 1D array obtaining a ($P^2 \cdot C$) x 1 vector. This process replicates the 1D sequence of token embeddings in NLP Transformers.

The patches are linearly projected (**E** in (2)), obtaining (N+1) x D vectors (patch embeddings, $p_e^N = x_p^N \cdot E$) by following (2).

$$\mathbf{z}_0 = \begin{bmatrix} \mathbf{x}_{class} \\ \mathbf{x}_p^1 \mathbf{E} \\ \mathbf{x}_p^2 \mathbf{E} \\ \vdots \\ \mathbf{x}_p^N \mathbf{E} \end{bmatrix} + \mathbf{E}_{pos} = \begin{bmatrix} \mathbf{x}_{class} \\ \mathbf{p}_e^1 \\ \mathbf{p}_e^2 \\ \vdots \\ \mathbf{p}_e^N \end{bmatrix} + \mathbf{E}_{pos} \tag{2}$$

where $\mathbf{E} \in \mathbb{R}^{P^2 C \times D}$ and $\mathbf{E}_{pos} \in \mathbb{R}^{(N+1) \times D}$

D is the depth or hidden dimension of the Transformer. In addition, the position of each patch is included ($\mathbf{E_{pos}}$) to preserve the positional information, and x_{class} ($x_{class}=z_0^0$) is the learnable parameter included in place 0. This sequence of patch embeddings is the input of the Transformer encoder.

As Fig. 1 shows, the Transformer encoders (or Transformer layers L) are implemented by adding a Multihead Self-Attention (MSA) and a MultiLayer Perceptron (MLP) block after a normalization layer. A residual connection is included between each normalization layer input and the output of the MSA and MLP blocks (Fig. 1 right), as it is respectively described by (3) and (4):

$$\mathbf{z}'_l = \text{MSA}(\text{LN}(\mathbf{z}_{l-1})) + \mathbf{z}_{l-1}, \quad l = 1 \ldots L \tag{3}$$

where l is the layer ViT index and LN a normalization layer,

$$\mathbf{z}_l = \text{MLP}(\text{LN}(\mathbf{z}'_l)) + \mathbf{z}'_l, \quad l = 1 \ldots L \tag{4}$$

where l is the layer ViT index and LN a normalization layer.

We recommend the original ViT paper [2] for more details about the MSA layer.

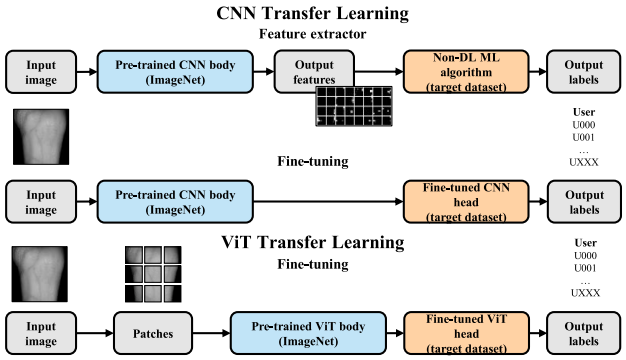

FIGURE 2. CNN transfer learning vs. current ViT transfer learning.

Finally, z_0^L (coming from the learnable embedding x_{class}) is the output of the ViT that is normalized (5) and converted into a final decision class label by an MLP block.

$$\begin{aligned} y &= LN(\mathbf{z}_L^0) \\ &\text{where } \mathbf{z}_L^0 \text{is the identification decision and} \\ &\text{the output of the ViT the in last layer L} \end{aligned} \tag{5}$$

This decision head is implemented by an MLP during the ViT pre-training and by a linear layer during the fine-tuning. In the hybrid ViT version, instead of dividing the input image into several patches that pass through the Transformer, the output of a CNN (a ResNet50, for example), acting as a feature extractor, is the input of the ViT.

B. CNN TRANSFER LEARNING VS. VIT TRANSFER LEARNING

1) CNN TRANSFER LEARNING

As Fig. 2 summarizes and according to what is widely explained in [60], Transfer Learning is usually applied in the CNN world following two variants:

1) Feature extractor: a CNN is pre-trained in a medium-large dataset to extract unique features that are then classified for the required recognition task with a traditional non-DL Machine Learning method such as Logistic Regression (Fig. 2).
2) Fine-tuning: a CNN is pre-trained in a medium-large dataset to extract unique features. First, the head of the CNN and then the entire network is retrained (fine-tuning) in the required recognition task.

2) VIT TRANSFER LEARNING

Vision Transformers have been explored by using the TL fine-tuning technique: the ViT is pre-trained on medium-large datasets and retrained on the desired task by replacing the classification head with a zero-initialized head. Commonly ViTs are fine-tuned using a dataset with higher-resolution images.

III. VASCULAR BIOMETRIC RECOGNITION DATABASES

In this section, the pre-training datasets ImageNet-1k and ImageNet-21k (image classification tasks) and the proposed contactless wrist dataset, UC3M-CV3, directly acquired using two smartphones, are detailed before exploring the ViT model that we are presenting. The details of the 13 remaining state-of-the-art vein datasets used in this work for finger (HKPU, UTFVP, FV-USM, PLUSVein-Contactless Finger Vein Database, and NUPT-FPV), palm (CASIA Multi-Spectral, PUT, VERA Palmvein, and PLUSVein-Contactless Palm Vein Database), hand dorsal (Bosphorus Hand Vein Database), and wrist (PUT, UC3M-CV1, and UC3M-CV2) are collected in Appendix A.

A. IMAGENET

We have used this well-known middle-size benchmark to pre-train the proposed ViTs. ImageNet was revealed in 2009 by Deng et al. [3] and is one of the most popular datasets in the CV image recognition world. The pre-training has been divided into two training processes. Firstly, we use the 21-k version of ImageNet. This dataset contains over 14 million RGB images with more than 21,000 distinct object categories ranging from any kind of animal to all types of vehicles. Secondly, to follow a fine-tuning refinement process that concludes with vein recognition training, we pre-train again using, in this case, ImageNet-1k, a subset of ImageNet consisting of 1.3 million images and 1000 object classes.

The pre-trained process on these datasets allows the ViT models to learn a broad range of generic imaging features, which are then refined during the fine-tuning distilling procedure to obtain unique and discriminative vascular characteristics of each individual. The public nature of these state-of-the-art benchmarks, combined with the described TL procedure, make this data not only suitable for VBR but also for other biometric recognition modalities.

B. FINE-TUNE: UC3M-CV3 - PROPOSED CONTACTLESS WRIST VEIN DATASET ON SMARTPHONES

Following the line of our two previous contactless databases presented in [9] and [40], UC3M-CV1 (UC3M-Contactless Version 1) and UC3M-CV2 (UC3M-Contactless Version 2), we expand the latter by introducing UC3M-CV3, which doubles the number of subjects. This new dataset consists of 4800 vein images captured from both wrists of 100 volunteers in two separate sessions per each individual (12 samples per wrist and session). The interval between sessions was between 2 and 4 weeks. The 24 samples for each wrist were automatically acquired in 24 different user attempts. We fine-tune our double pre-trained ImageNet ViTs on this benchmark and the mentioned VBR datasets described in Appendix A.

UC3M-CV3 was collected during 2021 with the two same smartphones Xiaomi© Pocophone F1 and Xiaomi© Mi 8 and their integrated facial recognition NIR cameras used in the precursive UC3M-CV2. In this case, to improve the user collection experience and automate the capturing process, we introduce an automatic wrist tracking algorithm based on the well-known Haar Cascades [10] method: Wrist Vascular Haar Cascade (WVHC). This algorithm was trained with

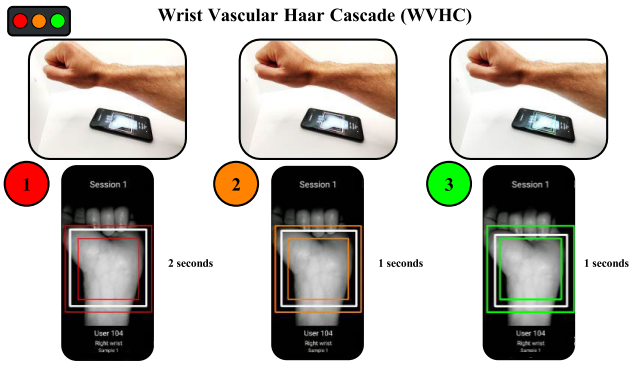

FIGURE 3. Wrist Vascular Haar Cascade (WVHC) algorithm for UC3M-CV3 data collection. Sample sequence following a traffic light sequence: red, amber, and green.

the 1200 images from UC3M-CV2 as the positive set and 1000 negative images that include empty background scenes and SMILES [61] images (grayscale pictures with smiling and not smiling faces) in a 19-81 % composition. This combination, along with the rectangle guiding algorithm included in WVHC and shown in Fig 3., prevents the possible wrist tracking false positives caused by background and volunteer faces features.

The users placed the WVHC tracking rectangle for each capture between the two other guiding rectangles, as Fig 3. shows. The volunteers were guided through three traffic light transitions or states:

1) Red: the two guiding rectangles light up red when users place the WVHC tracking rectangle between them (Fig. 3). The user must move the wrist so that the white tracking rectangle is positioned between the guiding interior and exterior squares and hold this position for 2 seconds to progress to the next stage. If not, the sequence starts again.
2) Amber: the two guiding rectangles light up green when the user keeps the WVHC tracking rectangle between them (Fig. 3). The user should hold the wrist still for 1 second to advance to the next estate. If not, the sequence starts again.
3) Green: After one more second of preserving the wrist in the same position, the sample is captured. In this state, the guiding lines illuminate green (Fig. 3).

This algorithm sequence provides feedback to the user on how to place the wrist correctly and ensures certain repeatability across samples from all users, as Fig. 3 shows. The Wrist Vascular Haar Cascade was trained in 16 states for approximately 2 hours using a 24 × 24 pixel window size acquiring 162,336 unique features.

To obtain more details about the capturing hardware, we recommend consulting [9].

The greyscale images (8 bit/pixel monochromatic images with values from 0, black, to 255, white) with 640 C–480 resolution (VGA) has been stored in JPEG/JFIF compressed format. Table 1 and (6) summarize the values of this database (24 images per class, 200 classes):

$$100\ subjects\ \times\ 2\ wrists\ \times\ 12\ samples \times\ 2\ sessions\ =\ 4800\ images \tag{6}$$

TABLE 4. Details of the pure vascular ViTs implemented.

Model	Layers (L)	Hidden dimension (D)	MLP size	Heads (A)	Parameters (M.)	Patch size (P)	Number of patches (N)
ViT-Small	12	192	1536	3	21.7	16	196
ViT-Base	12	768	3072	12	87.5	32	49
ViT-Large	24	1024	4096	16	303.4	16	196

TABLE 5. Details of the hybrid vascular ViT implemented.

Model	Number of residual bottlenecks	Parameters (Millions)	Patch size (P)
ResNet50 + ViT-Large	[3, 4, 6, 3]	328.1	1

IV. VISION TRANSFORMERS (ViT) FOR VASCULAR BIOMETRIC RECOGNITION

A. ARCHITECTURE

In this study, three pure Vision Transformer architectures have been implemented and tested for Vascular Biometric Recognition: ViT Small (ViT-S) [62], ViT Base (ViT-B) [2], and ViT Large (ViT-L) [2], and one hybrid structure, combining the output of the ResNet50 CNN and the Vit-L (R50 + ViT-L) [2]. All parameters of the pure ViT, including the number of layers or ViTs blocks (L), the hidden dimension (D), the number of self-attention heads (A), the number of parameters (in Millions), the patch size, and the number of patches (N) are shown in Table 4.

Table 5 summarizes the values for the hybrid ViT composed of the ResNet50 architecture and the ViT-L32.

We have implemented ViT-S/16, ViT-B/32, ViT-L/16, and ResNet50 + ViT-L/32 for the fourteen vein datasets. The size of the MLP block (output of each L layer of the ViT encoder) is 4 times (4·D) the hidden size (D), except for the small ViT version (ViT-S), which is 8 times (8·D).

The number of patches (N) for ViT-S/16, ViT-B/32, and ViT-L/16 is 196, 49, and 196 because all the images employed have been rescaled to 224 × 224 ($N = \frac{H \cdot W}{P^2}$). The hybrid model is composed of the ResNet50 that consists of 4 stages or blocks of 3, 4, 6, and 3 bottlenecks. The output feature map of the CNN is the input to the ViT-L/32.

It is important to note that hybrid ViT models have been previously and successfully employed in the literature for

other Computer Vision recognition tasks combining the local feature extraction abilities of CNNs with the high-level relationships capture of ViTs. However, due to the training complexity, apart from the 4 pure ViTs previously detailed in this work, only the ResNet50 + ViT-L/32 has been successfully fine-tuned.

B. PRE-TRAINING

As was previously indicated, the 4 different ViTs have been pre-trained, firstly on ImageNet-21k, which contains more than 14 million images (21,000 classes), and secondly on ImageNet-1K (1,000 classes and 1.3 million mages). In this process, all images have been secured from being reused, as it is described in [62]. All models, pre-trained by S. Paul, can be found in [63] for TensorFlow© implementations.

1) IMAGENET-21K

For this pre-training procedure, the models were trained [63] during 300 epochs with the Adam (Adaptive Moment Estimation) optimizer (with $\beta_1 = 0.9$ and $\beta_2 = 0.999$) and a batch size of 4096 images.

A linear learning-rate warmup followed by a cosine decay (10,000 total steps) has been applied (example in Fig. 4). The base learning rate was 0.001.

The resolution of all input images was 224 × 224.

2) IMAGENET-1K

After the ImageNet-21k, the models were retrained [63] with the Stochastic Gradient Decent (SGD) optimizer (momentum $\gamma = 0.9$) using a batch size of 512.

A linear learning-rate warmup of 500 steps followed by a cosine decay of 19.500 steps has been applied (example in Fig. 4). The base learning rate was 0.01 and 0.03.

All input images were 224 × 224.

C. FINE-TUNING

The Transfer Learning process has been achieved after a ViTs training refinement process. This fine-tuning procedure to identify each user of the fourteen vein datasets demonstrates the powerful performance and versatility of Vision Transformers.

All the models have been fine-tuned during 50-80 epochs with a Stochastic Gradient Decent (SGD) optimizer and a base learning rate of 0.03-0.06 followed by a cosine decay after a previous linear warmup (except for some hybrid ViT models) that always starts at 0.006 (example in Fig. 4). The linear warmup has been performed in all cases during 10 steps. The batch size oscillates between 8 and 64 images.

As previously mentioned, the input image size of all ViTs has been set to 224 × 224 x 3 in accordance with the ImageNet pre-training. All images have been rescaled to this size. A Region Of Interest (ROI) extraction was only carried out for the FV-USM fine-tuning process. In this specific case, the images were centrally cropped from 640 × 480 to 20 × 400 in an effort to minimize the impact of background noise

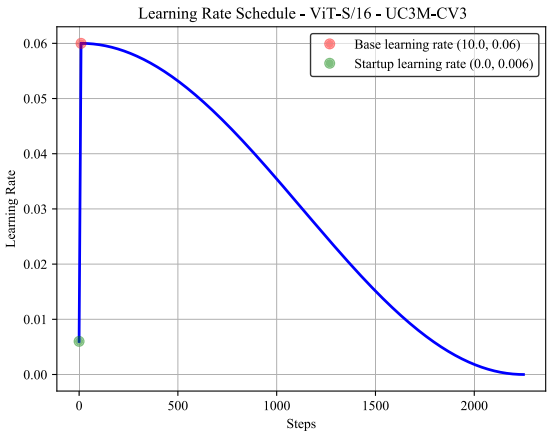

FIGURE 4. Learning rate schedule: linear warmup + cosine decay for the ViT-S/16 model on UC3M-CV3.

before being resized to 224 × 224. This step was likely demanded due to the training complexity introduced by the limited number of samples per class.

Equation (7) exemplifies the learning rate schedule for the example of Fig. 4: ViT-S/16 model on UC3M-CV3 (50-50 % train-test ratio):

$$Total\_steps = \frac{Number\_of\_training\_images \cdot Epochs}{Batch\_size} = \frac{2400 \cdot 60}{64} = 2250 \quad (7)$$

All data, whenever possible, has been split into 50-50 % and 75-25 % train-test subsets.

The ViT models have been fine-tuned using Python 3.7.9 programming language, TensorFlow© 2.0 (2.6.0 version) open-source library, and its Keras© (2.6.0 version) API.

The processing has been carried out using the NVIDIA® GeForcer® RTX 2080 Ti (11 GB GDDR6 memory) GPU and the 9th Generation Intel® Core™ i9k (64-bit, 16 GB of RAM, 3.6 GHz) CPU of the Dell© Alienware Aurora R8 computer with Windows 10 Home OS.

The Cuda® parallel computing platform (version 11.2) has been used to perform the CPU-GPU communication to parallelize the computations. Table 6 summarizes the number of epochs, the base learning rate, and the batch size for the tested datasets of the finger, palm, dorsal, and wrist modalities.

Table 6 shows that for most of the ViT-S/16 and ViT-B/32 structures, the number of epochs, the base learning rate, and the batch size is 60, 0.06, and 64. For ViT-L/16, these values are 60, 0.03, and 8. We assume that the ViT-L architecture requires this base learning and batch size reduction due to its larger size (more L layers and larger D dimension). Therefore the training process is softer. For the same reason, the ViT-R50+L/32 models have been optimally trained with an initial 0.006 learning rate schedule without following the previous linear warmup. It is also interesting to emphasize that ViT-S/16 and ViT-B/32 have been trained over the CASIA dataset during 80 epochs. In this case, and as discussed in the experiments' session, we hypothesize that the lack of data

(only 6 independent samples per user) requires long training and probably means an inadequate model generalization ability.

V. EXPERIMENTS AND RESULTS

In this work, the biometric and offline computing performance for the four explored ViT structures have been obtained following the ISO/IEC 19795-1 [11] for vascular biometric identification.Therefore, the proposed ViT structures are evaluated based on two criteria: accuracy of individual identification in each database and speed of each attempt, evaluating the time required to achieve this performance.

A. BIOMETRIC PERFORMANCE

To verify the identification performance the Cumulative Match Characteristic (CMC) plots shown in Fig. 5 to 8 represent the True-Positive Identification Rate (TPIR) over the returned rank (R). The Failure-To-Enrol Rate (FTER) and the Failure-To-Acquire Rate (FTAR) are unknown for all the datasets.

All CMC curves show the rank 1 value (R = 1) highlighted with a start, as the legend indicates. Therefore, according to the reduced size of the employed databases (from 84 to 840 unique users), we consider rank 1 as the only significant value to obtain appropriate conclusions.

1) FINGER

Fig. 5 shows the CMC curves for the three finger vein datasets in which we obtain remarkable results: HKPU V1.0, FV-USM, and NUPT-FPV.

It is worth noting that, some palm vein datasets, the lack of samples per class made it impossible to achieve successful ViTs biometric performance, even after applying crop, rotation resize, and flip data augmentation. We identified this problem in the UTFVP and PLUSVein-Contactless Palm datasets. They only consist of, respectively, 4 and 5 per class. We consider inaccurate/erroneous practices to obtain robust authentication/identification algorithms with these reduced numbers of images per user/class in the Deep Learning paradigm. After experimenting with the four ViT structures and the fourteen vein datasets, we reinforce this claim by not obtaining relevant results over the UTFVP and PLUSVEIN-Contactless finger datasets. In the FV-USM case, the dataset has been split into 60-40 % (7-5 images per class) and 75-25 % (9-3 images per class) train-test subsets to be able to train the ViTs, instead of the desired 50-50 % (6-6 images) and 75-25 % (9-3 images). With the proposed architectures and this vein recognition variant, at least 7 images per class were necessary for training, as we confirm with the palm vein CMC curves (6 images in this case).

2) PALM

Fig. 6 includes the CMC plots for the four palm vein datasets: CASIA Multi-Spectral, PUT Palm, VERA Palmvein, and PLUSVein-Contactless Palm Vein Database.

Following the previous results obtained on the FV-USM, UTFVP and PLUSVEIN-Contactless finger datasets, the CASIA and PLUSVein-Contactless Palm databases have been respectively split into 67-33 % (8-4 images per class) and 83-17 % (10-2 images per class), and 60-40 % (6-4 images) and 80-20 % (8-2 images) train-test sets. All the multispectral correspondence samples have been used in the same subsets, i.e., the sample of one user at 850 nm and the same sample at 940 nm (950 nm in PLUSVein-Contactless Palm dataset) are used just for train or validation. For this reason, we do not split the data into 75-25 % (odd number of samples) train-test subsets but 83-17 % and 80-20 %.

As it has been anticipated, at least 6-7 images per user have been required to train the proposed vein ViTs properly. Definitely, these values depend on the image features. It also should be considered that testing the models with 2 images is not precise enough. Furthermore, the models over all datasets should test in real-time to correctly verify the performance for a final real application. In this sense, the computational time performance is shown in the next section.

3) DORSAL

Fig. 7 provides the CMC curves for the unique dorsal vein database used: Bosphorus Hand Vein Database (vein set). This dataset consisting of 12 low-quality dorsal samples for the 100 independent classes has been split into 50-50 % and 75-25 % train-test subsets.

As shown later, ViT-L/16 obtains the best biometric performance not only in this VBR modality but also the other three. As expected, the greater the amount of data used for training, the smaller the amount employed for testing and, inevitably, the higher TPIR values, as demonstrated by comparing the 50-50 % train-test and 75-25 % train-test curves.

4) WRIST

Fig. 8 shows the CMC curves for the four wrist vein datasets tested, PUT Wrist, UC3M-CV1+CV2, and UC3M-CV2, and the proposed non-contact UC3M-CV3 database.

Table 8 summarizes the highest ViT performance on the fourteen vein databases.

Upon evaluating Table 8 and all CMC curves, it is clear that ViT-L/16 is the best-performing vascular ViT, achieving the highest biometric classification outcomes for all the tested datasets except for the NUPT-FPV (vein set) database, although the difference is not excessive (TPIR75-25 @ ViT-B/32 = 98.48 % vs. TPIR75-25 @ ViT-L/16 = 97.69 %). These results are consistent with the findings of the original ViTs article, where larger ViT architectures are associated with better performance. It is truly remarkable that ViT-L/16 accomplished a TPIR of more than 98 % for all datasets except CASIA (TPIR50-50 = 96.00 %) and Bosphorus (TPIR50-50 = 97.67 %), in the 50-50 % train-test scenarios. This value increases up to over 99 % except for CASIA (TPIR75-25 = 97.75 %) and UC3M-CV2 (TPIR75-25 = 98.83 %) with the 75-25 % train-test subsets.

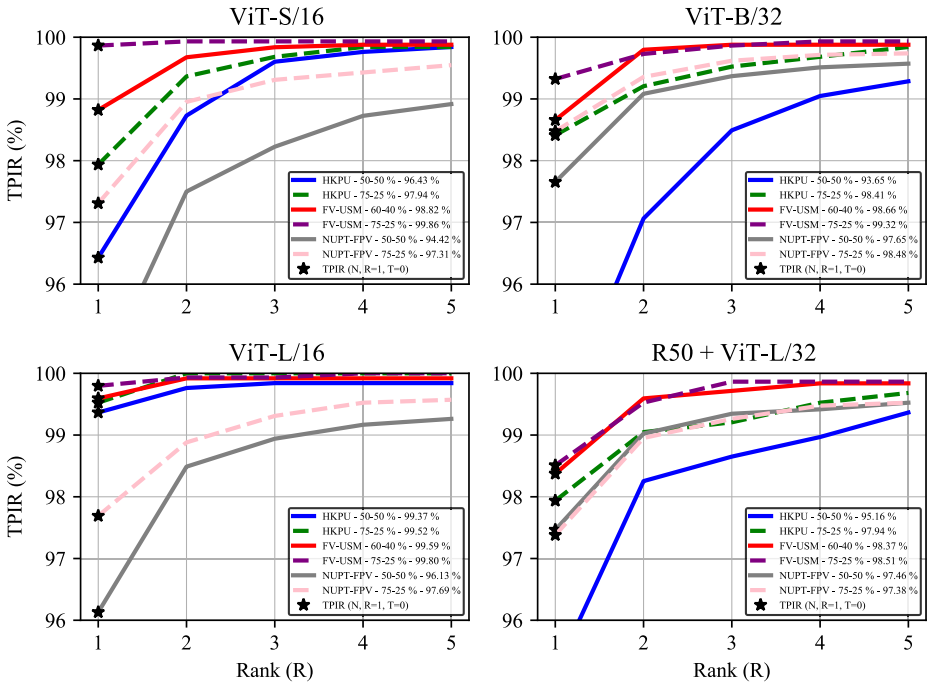

FIGURE 5. Biometric performance: CMC curves for the four proposed transfer learning ViT architectures (ViT-S/16, ViT-B/32, ViT-L/16, and ResNet50 + ViT-L/32) pre-trained on ImageNet and fine-tuned over the finger vein datasets split in 50-50 % and 75-25 % train-test sets. The continuous and line-line curves in different colors show the ViTs performance on the HKPU V1.0, FV-USM, and NUPT-FPV datasets. The TPIR at rank 1, TPIR (N, R = 1, T = 0), for each curve is provided in the legend.

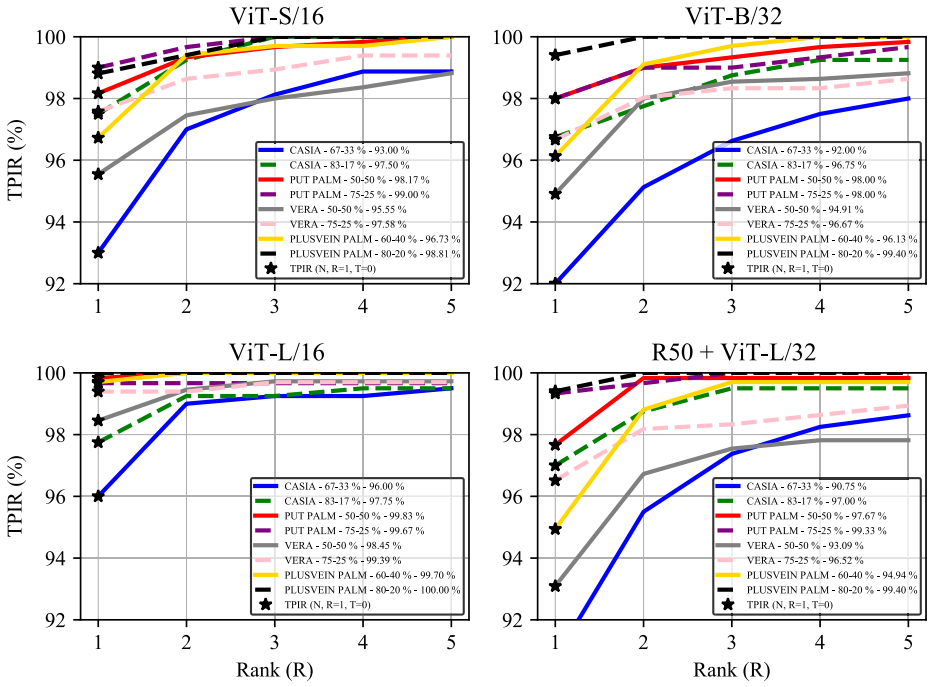

FIGURE 6. Biometric performance: CMC curves for the four proposed transfer learning ViT architectures (ViT-S/16, ViT-B/32, ViT-L/16, and ResNet50 + ViT-L/32) pre-trained on ImageNet and fine-tuned over the palm vein datasets split in 50-50 % and 75-25 % train-test sets. The continuous and line-line curves in different colors show the ViTs performance on the CASIA Multi-Spectral, PUT Palm, VERA Palmvein, and PLUSVein-Contactless Palm Vein datasets. The TPIR at rank 1, TPIR (N, R = 1, T = 0), for each curve is provided in the legend.

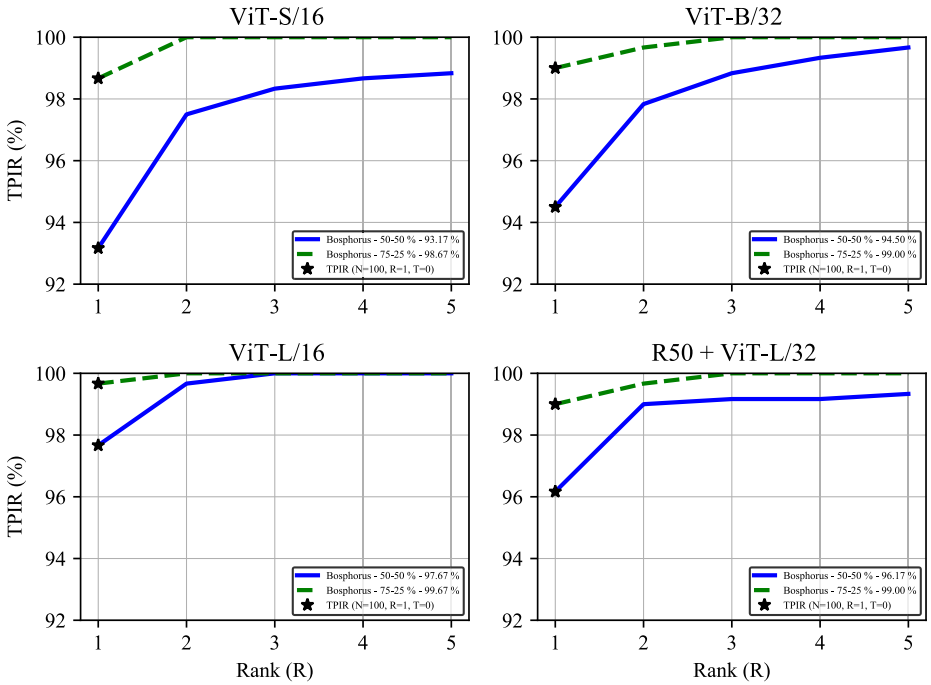

FIGURE 7. Biometric performance: CMC curves for the four proposed transfer learning ViT architectures (ViT-S/16, ViT-B/32, ViT-L/16, and ResNet50 + ViT-L/32) pre-trained on ImageNet and fine-tuned over the dorsal vein dataset split in 50-50 % and 75-25 % train-test sets. The blue continuous and green line-line curves in different colors show the ViTs performance on the Bosphorus Hand Vein Database (vein set). The TPIR at rank 1, TPIR (N = 100, R = 1, T = 0), for each curve is provided in the legend.

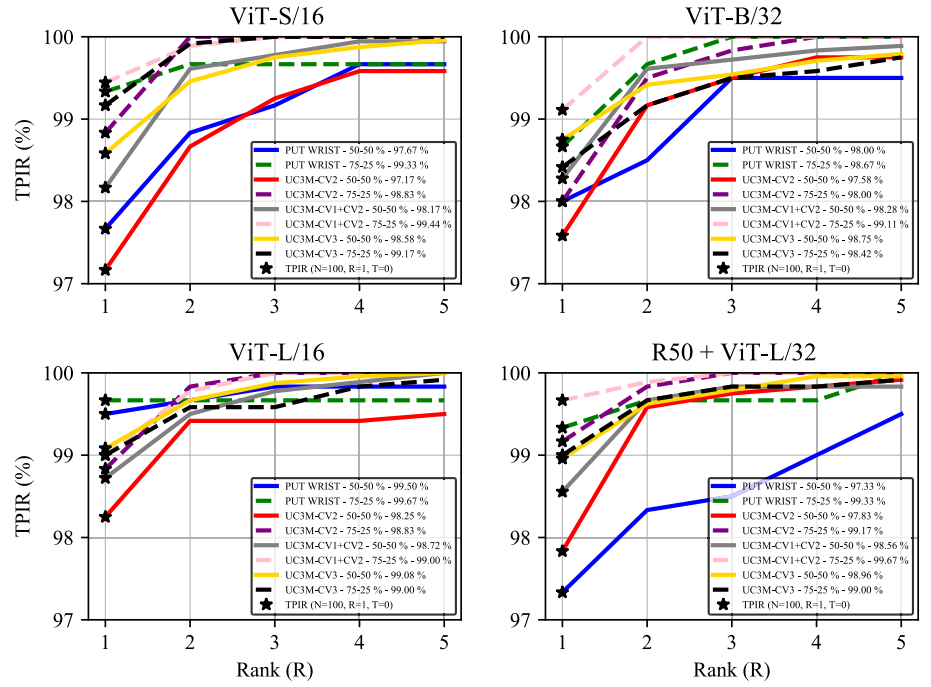

FIGURE 8. Biometric performance: CMC curves for the four proposed transfer learning ViT architectures (ViT-S/16, ViT-B/32, ViT-L/16, and ResNet50 + ViT-L/32) pre-trained on ImageNet and fine-tuned over the wrist vein datasets split in 50-50 % and 75-25 % train-test sets. The continuous and line-line curves in different colors show the ViTs performance on the PUT Wrist, UC3M-CV1+CV2, UC3M-CV2, and UC3M-CV3 datasets. The TPIR at rank 1,TPIR (N=100, R = 1, T = 0), for each curve is provided in the legend.

TABLE 6. Vein fine-tuning hyperparameters details for the finger and palm vascular datasets.

Dataset	Model	Epochs	Base learning rate	Batch size	Train-Test (%)	Train-Test (images)	Training time (min)
HKPU (vein set)	ViT-S/16	60	0.06	64	50-50 75-25	2400-2400 3600-1200	13.54 17.33
	ViT-B/32	60	0.06	64	50-50 75-25		18.87 22.52
	ViT-L/16	60	0.03	8	50-50 75-25		113.60 156.27
	ViT-R50+L/32	65	0.03	20	50-50 75-25		51.89 71.27
FV-USM	ViT-S/16	60	0.06	64	60-40 75-25	600-600 900-300	32.32 37.88
	ViT-B/32	60	0.06	32	60-40 75-25		48.70 56.59
	ViT-L/16	60	0.03	8	60-40 75-25		307.38 349.42
	ViT-R50+L/32	65	0.03	20	60-40 75-25		130.35 152.31
NUPT-FPV (vein set)	ViT-S/16	60	0.06	64	50-50 75-25	600-600 900-300	82.96 104.62
	ViT-B/32	70	0.06	64	50-50 75-25		133.08 164.24
	ViT-L/16	60	0.03	8	50-50 75-25		758.85 1037.87
	ViT-R50+L/32	65	0.03	20	50-50 75-25		327.12 443.20
CASIA (vein set)	ViT-S/16	80	0.06	32	67-33 83-17	600-600 900-300	19.97 23.38
	ViT-B/32	80	0.06	32	67-33 83-17		29.51 34.86
	ViT-L/16	60	0.03	8	67-33 83-17		140.37 158.20
	ViT-R50+L/32	50	0.006	20	67-33 83-17		46.93 54.69
PUT Palm	ViT-S/16	60	0.06	64	50-50 75-25	600-600 900-300	10.43 11.07
	ViT-B/32	60	0.06	64	50-50 75-25		11.63 13.18
	ViT-L/16	60	0.03	8	50-50 75-25		60.02 75.86
	ViT-R50+L/32	50	0.006	20	50-50 75-25		22.52 29.97
VERA	ViT-S/16	60	0.06	64	50-50 75-25	600-600 900-300	12.30 14.85
	ViT-B/32	60	0.06	32	50-50 75-25		16.52 19.28
	ViT-L/16	60	0.03	8	50-50 75-25		103.54 128.67
	ViT-R50+L/32	50	0.006	20	50-50 75-25		35.74 44.67
PLUSVein-Contactless Palm	ViT-S/16	60	0.06	64	60-40 80-20	600-600 900-300	6.51 7.38
	ViT-B/32	60	0.06	64	60-40 80-20		8.08 9.11
	ViT-L/16	60	0.03	8	60-40 80-20		45.18 54.89
	ViT-R50+L/32	50	0.006	20	60-40 80-20		19.43 23.09

TABLE 7. Vein fine-tuning hyperparameters details for the dorsal hand and wrist vascular datasets.

Dataset	Model	Epochs	Base learning rate	Batch size	Train-Test (%)	Train-Test (images)	Training time (min)
Bosphorus (vein set)	ViT-S/16	60	0.06	64	50-50 75-25	600-600 900-300	7.67 9.20
	ViT-B/32	60	0.06	64	50-50 75-25		10.60 12.50
	ViT-L/16	60	0.03	8	50-50 75-25		55.18 73.89
	ViT-R50+L/32	50	0.006	20	50-50 75-25		21.10 27.64
PUT Wrist	ViT-S/16	60	0.06	64	50-50 75-25	600-600 900-300	10.42 10.90
	ViT-B/32		0.06	64	50-50 75-25		11.80 13.59
	ViT-L/16		0.03	8	50-50 75-25		57.89 76.65
	ViT-R50+L/32		0.006	20	50-50 75-25		22.54 28.36
UC3M-CV2	ViT-S/16	60	0.06	64	50-50 75-25	1200-1200 1800-600	13.35 16.56
	ViT-B/32		0.06	64	50-50 75-25		17.56 21.81
	ViT-L/16		0.03	8	50-50 75-25		111.41 145.37
	ViT-R50+L/32		0.03	20	50-50 75-25		45.86 60.37
UC3M-CV1+ CV2	ViT-S/16	60	0.06	64	50-50 75-25	1800-1800 2700-900	18.59 23.66
	ViT-B/32		0.06	64	50-50 75-25		25.40 31.47
	ViT-L/16		0.03	8	50-50 75-25		166.39 226.55
	ViT-R50+L/32		0.03	20	50-50 75-25		66.75 88.25
UC3M-CV3	ViT-S/16	60	0.06	64	50-50 75-25	2400-2400 3600-1200	24.80 31.21
	ViT-B/32		0.06	64	50-50 75-25		34.49 41.42
	ViT-L/16		0.03	8	50-50 75-25		223.80 297.47
	ViT-R50+L/32		0.03	20	50-50 75-25		87.15 115.74

When analyzing the wrist vein results, ViT-L/16 achieves TPIR higher than 98 % (50-50 %) for all four sets. It is also notable that the two CMC curves for the train-test scenario are quite similar in the UC3M-CV3 case per each ViT architecture. Given that the capture hardware for UC3M-CV2 and UC3M-CV3 is the same (i.e., two unmodified smartphones: Xiaomi© Pocophone F1 and Xiaomi© Mi 8), we hypothesize that the high-quality data automatically acquired by the novel WVHC algorithm without operator intervention is what makes the difference. Furthermore, the ViT performance, particularly in the 50-50 % split, is better for UC3M-CV3 than UC3M-CV2 even though the former consists of twice as many users, 200. Last but not least, it is also noteworthy that the UC3M-CV1+CV reaches the top TPIR in the four ViTs except the ViT-L/16. Once again, we believe that data is the critical factor because UC3M-CV1+CV2 is the database with the most samples per class (36) coming from two different capture devices, which likely leads not only to higher experimental results but also a potentially better generalization ability.

B. COMPUTATIONAL TIME PERFORMANCE

1) FINE-TUNING TIME

Table 6 shows the ViT architecture fine-tuning times for each modality and dataset. As shown, the column for the required training time illustrates that, as expected from architectures based on input patches, smaller patch sizes (P) result in a greater number of patches (N), leading to longer training

TABLE 8. Biometric performance: Highest TPIRs.

	Dataset	Class samples	Model	$TPIR_{50-50}$ (%)	$TPIR_{75-25}$ (%)
Finger	HKPU (vein set)	12	ViT-L/16	99.37	99.52
	UTFVP	4	ViT-S/16 ViT-B/32 ViT-L/16 ViT-R50+ L/32	-	-
	FV-USM	12	ViT-L/16	99.59 (* 60-40 % train-test)	99.86
	PLUSVein-Contactless	5	ViT-S/16 ViT-B/32 ViT-L/16 ViT-R50+ L/32	-	-
	NUPT-FPV (vein set)	20	ViT-B/32	97.65	98.48
Palm	CASIA (vein set)	12	ViT-L/16	96.00 (* 67-33 % train-test)	97.75 * (83-17 % train-test)
	PUT Palm	12	ViT-L/16	99.83	99.67
	VERA	10	ViT-L/16	98.45	99.39
	PLUSVein-Contactless	10	ViT-B/32	99.70 (* 60-40 % train-test)	100.00 (* 80-20 % train-test)
Dorsal	Bosphorus (vein set)	12	ViT-L/16	97.67	99.67
Wrist	PUT Wrist	12	ViT-L/16	99.50	99.67
	UC3M-CV2	24	ViT-L/16	98.25	98.83
	UC3M-CV1+CV2	36	ViT-L/16	98.72	99.00
	UC3M-CV3	24	ViT-L/16 (and ViT-S/16)	99.08	99.00

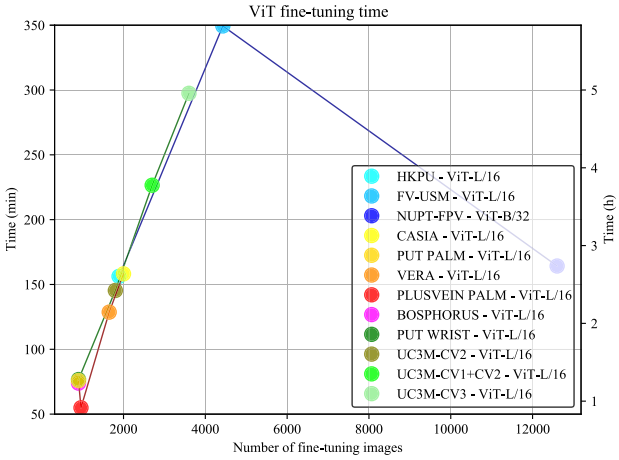

FIGURE 9. Computational time performance: Training (fine-tuning) time for the best ViT biometric performing architectures, i.e., ViT-L/16 in all cases except NUPT-FPV (ViT-B/32). The 75 % of each database gives the number of fine-tuning images, which comes from the 75-25 % train-test split. The blue, red, and green curves link the times for each VBR modality: finger, palm, and wrist (dorsal hand is only given by the Bosphorus dataset). The time is represented in minutes on the left y-axis and hours on the right y-axis. A detailed comparison of all trained ViTs can be found in Appendix B (Fig. 10).

TABLE 9. Offline ViT computational time for the UC3M-CV3 dataset.

Model	Train-Test (%)	Train-Test (images)	Total transaction average time for 10 attempts * (ms)	Unit transaction time (ms)
ViT-S/16	50-50	2400-2400	2351.41	235.14
ViT-B/32			2152.09	215.21
ViT-L/16			2428.05	242.80
ViT-R50+L/32			2323.19	232.32

(*) 10 attempts means we test the training set 10 times.

times. Therefore, the original finding from the ViT article is confirmed and can be verified for the four VBR modalities in Appendix B (Fig. 10)).

In this sense, Fig. 9 presents the fine-tuning time for the top biometric performance achieved by the ViTs in all cases with 75-25 % train-test scenarios, as per Table 8. The training time is plotted against the number of images used per dataset. The green line in the graph, representing the wrist data, shows that training time increases linearly as the number of training images grows. Nevertheless, this linear relationship is not evident in the finger (blue line) and palm (red line) variants. In the case of the finger, the NUPT-FPV training had a shorter time compared to HKPU, despite using 12,600 images for training rather than 1,890. As previously noted in Table 7, this discrepancy is attributed to the fact that the best TPIR was obtained with the ViT-B/32 model instead of ViT-L/16.

The PUT Palm data illustrates this divergence in the palm vein case. The longer fine-tuning time observed in this case compared to PLUSVein-Contactless Palm Vein (900 vs. 960) can be attributed to the difference in images format - 24 bit/pixel BMP as opposed to 8 bit/pixel PNG format, since in terms of resolution and as has been previously indicated, all input images have been rescaled to 224 × 224. It makes sense that PUT Wrist and PUT Palm obtained similar fine-tuning times, respectively, and in this case, 75.86 min (1.26 h) and 76.65 s (1.28 h), with the exact number of images (900) and image format (24-bit/pixel BMP).

The linear behavior shown in Fig. 9 helps to predict this key factor – the required fine-tuning time – for real-world applications that involve thousands of users, such as access control to stadiums, buildings, or other massive events. It is important to note that, depending on the final application, 17.3 hours obtained by ViT-L/16 could be considered a significant amount of time for NUPT-FPV, which consists of 140 users and 12600 training images.

2) INFERRING TIME

The offline computational time workload for direct (rank 1) identification transactions has been measured by inferring the testing batches (previously used to validate the models) in each ViT model in the UC3M-CV3 dataset. Table 9 presents the arithmetic average of the inference time of ten full-train-set attempts, i.e., we tested ten times the 900 training images.

VI. CONCLUSION

This work extensively demonstrates the high biometric performance of Vision Transformer across the four primary vein recognition modalities. Four different ViT architectures have been successfully trained over twelve state-of-the-art vein databases. In contrast to the trend in Vascular Biometric Recognition and to overcome the shortage of massive vein datasets, the models have not been trained from scratch but otherwise fine-tuned after an extended pre-training in an entirely different benchmark task.

Furthermore, a new wrist contactless VBR dataset with 100 users, 200 independent classes, and 4800 images has been introduced in this sense: UC3M-CV3. In order to stimulate the integration and development of these biometric modalities, this new database was entirely acquired using just two non-modified commercial smartphones during two separate sessions. In addition, the novel three-state algorithm proposed for automatic image acquisition/enrolment, Wrist Vascular Haar Cascade (WVHC), shows promising improvements in biometric performance and user sample capture comfort. This computationally efficient traditional Machine Learning algorithm was trained using the legacy UC3M-CV2 database.

The three pure ViT implemented, ViT-S/16, ViT-B/32, and ViT-L/16, generally perform better than the hybrid version, R50 + ViT-L/32, despite the ViTs creators' belief that convolutional inductive bias works well for small datasets. However, after analyzing the biometric identification performance according to the ISO/IEC 19795-1 standard [11], it becomes evident that ViT-L/16 outperforms the rest, as demonstrated by the CMC curves and rank-1 table for the twelve datasets.

DL-based solutions require extensive data for model training, and Vision Transformers is no exception. Furthermore, they need more images than the previous state-of-the-art CNNs algorithms. In this sense, the proposed ViTs were not successfully trained on the UTFVP (4 samples/class) and PLUSVein-Contactless finger vein datasets (5 samples/class). Additionally, after following a severe train-test ratio of 75-25 % and 50-50 % across all databases, the ViT only performed correctly on the FV-USM (12 samples/class), CASIA (12 samples/class), and PLUSVein-Contactless Palm Vein dataset (5 samples/class) for 60-40% - 75-25 %, 67-33 % - 83-17 %, and 60-40 % - 80-20 % train-test scenarios. This behavior results from the shortage of data, particularly in the unsuccessful training on UTFVP and PLUSVein-Contactless finger vein. Therefore, in order to obtain more realistic scenarios for new VBR algorithms, we encourage researchers to create larger datasets and not rely on evaluations based on testing models with only 3, 2, or even 1 image.

As demonstrated by the fine-tuning computational time analysis and the inherent behavior of ViTs concerning their input patch size, smaller patch sizes (P) result in a greater number of patches (N) and, therefore, more iterations, leading to longer training time. We presented the fine-tuning time for each ViT and dataset across the training images to demonstrate the viability or infeasibility of the proposed solution in real applications with thousands of users.

The offline computational time or inference time obtained evidence reduced unit transaction times, which means the applicability of VBR ViT in the four variants for real-time applications such us HPC embedded environments [64].

We hope this work contributes to the Vascular Biometric Recongion research community and industry to drive further development and adoption of this technology by pioneering implemented not only vein TL Vision Transformers but also DL models applied over the 4 main vein biometric modalities.

APPENDIX A
VBR DATABASES

A. FINGER

1) HKPU V1.0 (VEIN SET)

The Hong Kong Polytechnic University Finger Image Database (Version 1.0) was acquired in 2010 using a contactless device. The finger vein subset (near-infrared images) consists of 2520 images collected from the index and middle finger of the left hand of 105 subjects. Per finger, 6 samples were obtained in 2 different sessions using a near-infrared (NIR) camera and 850 nm LED (Light Emitting Diode) light (unknown devices) on transmission configuration.

The greyscale images (8 bit/pixel monochromatic images with values from 0, black, to 255, white) with 513 C– 256 resolution were stored in BMP format. Table 1 and (8) summarize the values of this database (12 images per class, 205 classes):

$$105\ subjects\ \times\ 2\ fingers\ \times\ 1\ hand\ (left) \times\ 6\ samples\ \times\ 2\ sessions\ =\ 2520\ images \quad (8)$$

2) UTFVP

This image set, the University of Twente Finger Vascular Pattern database, collected in 2012, consists of 1440 finger vein images. In total, 2 samples per session from the ring, index, and middle finger of both hands from 60 volunteers were acquired using a contact device composed of a BCi5 monochrome CMOS camera, B+W 093 infrared filter, and 8 SFH4550 near-infrared LEDs configured in transmission mode (mirror included).

The greyscale (8 bit/pixel) images with 672 C– 380 resolution were stored in Portable Network Graphics (PNG) format. Table 1 and (9) summarize the values of this database (4 images per class, 360 classes):

$$60\ subjects\ \times\ 3\ fingers\ \times\ 2\ hands \times\ 2\ samples\ \times\ 2\ sessions\ =\ 1440\ images \quad (9)$$

3) FV-USM

The University of Sains Malaysia built the FV-USM dataset with 5904 unique images from 123 volunteers in 2013. In 2 different sessions, 6 samples of the index and middle fingers of both hands were captured using a contact device comprising of a Sony© PSEye camera with an IR passing filter (unknown) and 3 LEDs (unknown) in contact and transmission mode.

The greyscale images (8 bit/pixel monochromatic images) with 640 C– 480 resolution (VGA) were stored in JPEG (Joint Photographic Experts Group)/JFIF (JPEG File Interchange Format) compressed format. Table 1 and (10) summarize the values of this database (12 images per class, 492 classes):

$$123\ subjects\ \times\ 2\ fingers\ \times\ 2\ hands \\ \times\ 6\ samples\ \times\ 2\ sessions\ =\ 5904\ images \quad (10)$$

4) PLUSVEIN-CONTACTLESS FINGER VEIN DATABASE

The PLUSVein-Contactless Finger Vein dataset is a subset of the PLUSVein-Contactless Finger and Hand Vein Database. It was collected by Uhl et al. [27] in 2018 along with the palm vein subset, which we also tested in this study. The proposed contactless finger-palm system was designed to capture this database using a novelty reflection and transmission technique by an IDS Imaging UI-ML3240-NIR with quantum efficiency in the NIR spectrum, a MIDOPT FIL LP780/27 NIR-pass filter, five 808 nm NIR laser diodes (transmission), sixteen 850 nm Osram© SFH 4550 LEDs (reflection), and sixteen 950 nm Vishay Semiconductors CQY LEDs (reflection).

In total, in a unique session, 1260 finger vein images were captured from 42 subjects and their ring, index, and middle finger of both hands. The greyscale images (8 bit/pixel) with 1280 C– 1024 resolution (Super XGA) were stored in PNG format. Table 1 and (11) summarize the values of this database (5 images per class, 252 classes):

$$42\ subjects\ \times\ 3\ fingers\ \times\ 2\ hands \\ \times\ 5\ samples\ \times\ 1\ session\ =\ 1260\ images \quad (11)$$

5) NUPT-FPV (VEIN SET)

In May 2022, Ren et al. [28] presented a multimodal fingerprint and finger vein dataset. The multimodal NUPT-FPV acquisition system combines an optical sensor and a NIR camera with 850 nm LED transmission light (unknown devices).

The finger vein subset used in the proposed work consists of 16,800 images from the ring, index, and middle fingers of both hands of the 140 participants. In the 2 sessions, 10 images were captured per finger.

The greyscale images (8 bit/pixel) with 300 C– 450 resolution were stored in BMP format. Table 1 and (11) summarize the values of this database (20 images per class, 840 classes):

$$140\ subjects\ \times\ 3\ fingers\ \times\ 2\ hands \\ \times\ 10\ samples\ \times\ 2\ sessions\ =\ 16800\ images \quad (12)$$

B. PALM

1) CASIA MULTI-SPECTRAL (VEIN SET)

The CASIA Multi-Spectral Palmprint Database (Version 1.0) [52] was acquired in 2007 using a contact device. The palm vein subset (NIR images) consists of 2400 images collected from both hands of 100 subjects. Per palm, 6 samples were obtained in 2 different sessions using a NIR CCD (Charge-Coupled Device) camera and 850 and 940 nm LED light (unknown devices) on reflection configuration.

The greyscale images (8 bit/pixel) with 768 C– 576 resolution were stored in JPEG/JFIF compressed format. Table 1 and (13) summarize the values of this database (12 images per class, 200 classes):

$$100\ subjects\ \times\ 2\ palms\ \times\ 6\ samples \\ \times\ 2\ sessions\ =\ 2400\ images \quad (13)$$

2) PUT PALM VEIN

The Poznan University of Technology (PUT) Palm Vein database is a subset of the PUT database. It was collected in 2011 along with the wrist vein subset, which we also tested in this study.

The contact palm system, proposed by Kabacinski and Kowalski [38], consisting of a USB camera and 850 nm LED light (unknown devices), captured 1200 infrared images from 50 users in 3 sessions.

The greyscale (24 bit/pixel) 1024 C– 768 images were stored in BMP format. Table 1 and (14) summarize the values of this database (12 images per class, 100 classes):

$$50\ subjects\ \times\ 2\ palms\ \times\ 4\ samples \\ \times\ 3\ sessions\ =\ 1200\ images \quad (14)$$

3) VERA PALMVEIN

This palm vein dataset consists of 2200 images from 110 volunteers. The corresponding publication [33], indicates that 5 samples per session (2 sessions) were contactless acquired for both palms of each user.

The capturing hardware comprises an Imaging Source© camera (an unknown device with a CCD Sony©) ICX618 sensor), and 940 nm LED illumination (unknown device). An HC-SR04 ultrasound sensor was included to provide feedback on proper distancing between the user's palm and the camera.

The greyscale (24 bit/pixel) 480 C– 640 (VGA) images were stored in PNG format. Table 1 and (15) summarize the values of this database (10 images per class, 220 classes):

$$110\ subjects\ \times\ 2\ palms\ \times\ 5\ samples \\ \times\ 2\ sessions\ =\ 2200\ images \quad (15)$$

4) PLUSVEIN-CONTACTLESS HAND VEIN DATABASE

The PLUSVein-Contactless Hand Vein subset of the PLUSVein-Contactless Finger and Hand Vein databases (previously detailed) includes 840 images captured from both palms of 42 users in 1 session.

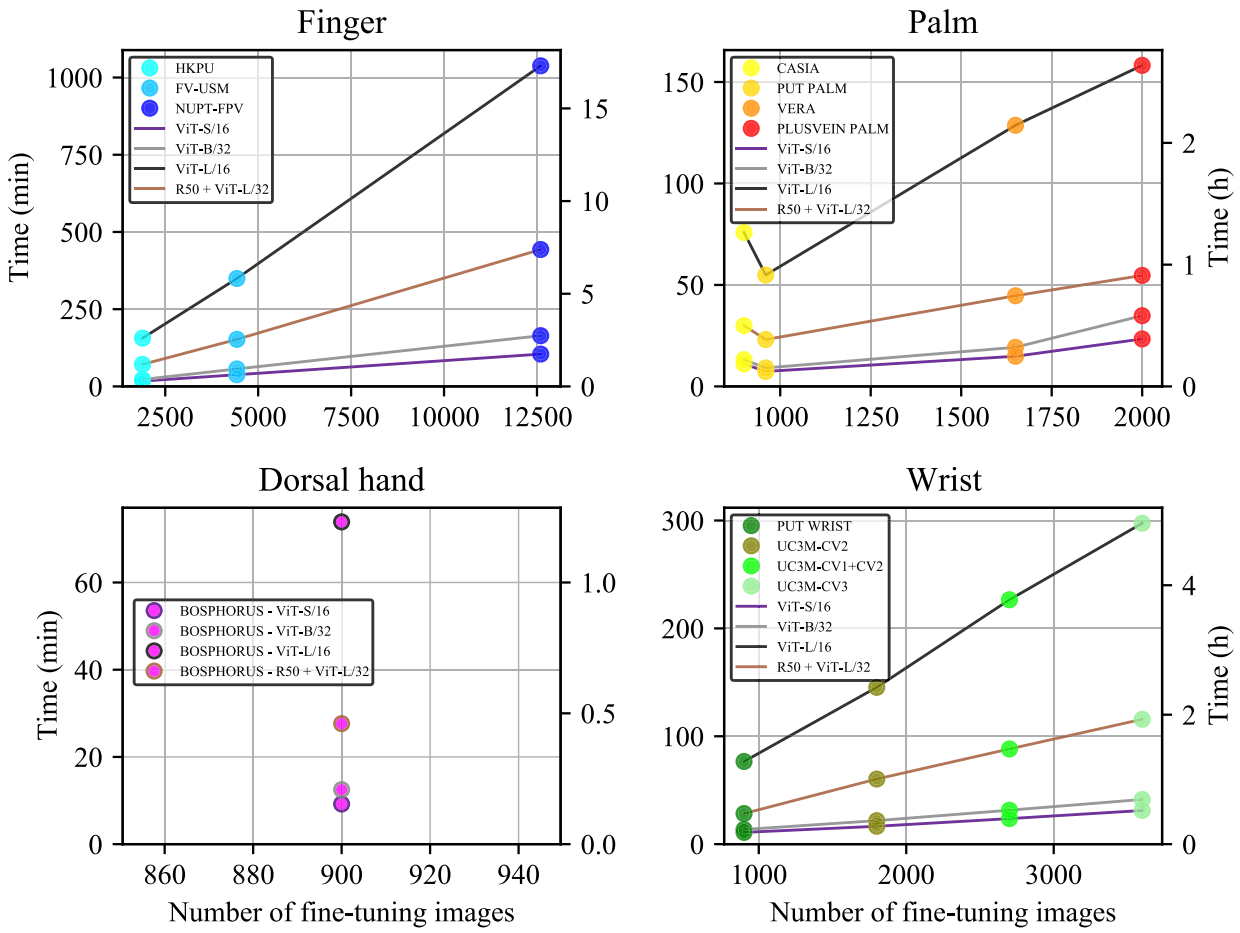

FIGURE 10. Computational time performance: Training (fine-tuning) time for the four ViTs and the twelve VBR datasets. The number of fine-tuning images is given by each database's 75-25 % train-test split. The time is represented in minutes on the left y-axis and hours on the right y-axis.

The greyscale (8 bit/pixel) 1280 C– 1024 images (Super XGA) were stored in PNG format. Table 1 and (16) summarize the values of this database (10 images per class, 84 classes):

$$42\ subjects\ \times\ 2\ palms\ \times\ 10\ samples \times\ 1\ sessions\ =\ 840\ images \tag{16}$$

C. HAND DORSAL

1) BOSPHORUS HAND VEIN DATABASE

This public dorsal dataset is a subset of the Bosphorus Hand Database, which includes hand geometry and hand vein data. It was collected by Yükselusing et al. [35] in 2010 using a NIR camera (WAT-902H2 ULTIMATE) and two unknown NIR light sources set in reflection mode

For the collection, 100 volunteers provided 12 samples of their left hand in 1 session but under 4 different conditions: 3 images under normal conditions, 3 images after carrying a 3 Kg bag for one minute, 3 images after closing and opening the fist using an elastic ball for one minute, and 3 images after placing an ice pack in the hand dorsal for one minute.

In total, 1200 greyscale images (24 bit/pixel) with 300 × 240 resolution were stored in BMP format. Table 1 and (17) summarize the values of this database (12 images per class, 100 classes):

$$100\ subjects\ \times\ 1\ hand\ dorsal\ (left)\ \times\ 12\ samples \times\ 1\ sessions\ =\ 1200\ images \tag{17}$$

D. WRIST

1) PUT WRIST VEIN

This public wrist subset of the PUT database (previously detailed in the palm modality) consists of 1200 images obtained from both wrists of 50 users in 3 different sessions with a contact system.

The greyscale (24 bit/pixel) 1024 C– 768 images were stored in BMP format. Table 1 and (18) summarize the values of this database (12 images per class, 100 classes):

$$50\ subjects\ \times\ 2\ wrists\ \times\ 4\ samples \times\ 3\ sessions\ =\ 1200\ images \tag{18}$$

2) UC3M-CV1

We acquired this wrist dataset in 2020 [40] using a small contactless device consisting of a modified Logitech® HD Webcam C525, a designed Printed Circuit Board (PCB) with

eight Osram© SFH 4715 A LEDs, and a Raspberry® Pi 4 Model B as processing and storage unit.

In the database collection process, 50 volunteers provide in two sessions both wrists to obtain 1200 images. In this study, UC3M-CV1 has been combined with the following dataset, UC3M-CV2.

The greyscale (8 bit/pixel) 480 C– 640 images were stored in JPG format. Table 1 and (19) summarize the values of this database (12 images per class, 100 classes):

$$50\ subjects\ \times\ 2\ wrists\ \times\ 6\ samples \times\ 2\ sessions\ =\ 1200\ images \quad (19)$$

3) UC3M-CV2

Also, in 2020 we collected a new subset extension, UC3M-CV2, published in [9]. For the same 50 volunteers, 2400 new images were captured using two smartphones in a contactless way.

The greyscale (8 bit/pixel) 640 C– 480 images (VGA) were stored in JPG format. Table 1 and (20) summarize the values of this database (24 images per class, 100 classes):

$$50\ subjects\ \times\ 2\ wrists\ \times\ 12\ samples \times\ 2\ sessions\ =\ 2400\ images \quad (20)$$

APPENDIX B
TRAINING (FINE-TUNING) TIME FOR THE FOUR ViTs AND THE TWELVE VBR DATASETS

Fig. 10 shows the training (fine-tuning) time for the four ViTs architectures, the four VBR variants, and the twelve successfully trained datasets.

ACKNOWLEDGMENT

The authors would like to thank Aritra Roy-Gosthipaty for his implementation inputs. Raul Garcia-Martin truly appreciates the grammar review and moral support provided by Maribeth Hoath-Perez. This work would not have been possible without the initial inspiring idea of Elena Figueira-Rodríguez. In addition, Raul Garcia-Martin would like to be grateful for the unconditional support provided by Laura Martínez-Pastor.

REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2017, *arXiv:1706.03762*.

[2] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[3] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Miami, FL, USA, Jun. 2009, pp. 248–255.

[4] *ImageNet Large Scale Visual Recognition Challenge (ILSVRC)*. Accessed: Jan. 12, 2023. [Online]. Available: https://www.image-net.org/

[5] *The CIFAR-100 Dataset*. Accessed: Jan. 12, 2023. [Online]. Available: https://www.cs.toronto.edu/ kriz/cifar.html

[6] X. Zhai, J. Puigcerver, A. Kolesnikov, P. Ruyssen, C. Riquelme, M. Lucic, J. Djolonga, A. S. Pinto, M. Neumann, A. Dosovitskiy, L. Beyer, O. Bachem, M. Tschannen, M. Michalski, O. Bousquet, S. Gelly, and N. Houlsby, "A large-scale study of representation learning with the visual task adaptation benchmark," 2019, *arXiv:1910.04867*.

[7] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," 2015, *arXiv:1503.02531*.

[8] R. Garcia-Martin and R. Sanchez-Reillo, "Deep learning for vein biometric recognition on a smartphone," *IEEE Access*, vol. 9, pp. 98812–98832, 2021, doi: 10.1109/ACCESS.2021.3095666.

[9] R. Garcia-Martin and R. Sanchez-Reillo, "Vein biometric recognition on a smartphone," *IEEE Access*, vol. 8, pp. 104801–104813, 2020, doi: 10.1109/ACCESS.2020.3000044.

[10] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, Kauai, HI, USA, Dec. 2001, p. 1, doi: 10.1109/CVPR.2001.990517.

[11] *Information Technology Biometric Performance Testing and Reporting Part 1: Principles and Framework*, Standard ISO/IEC 19795-1, American National Standards Institute, 2021.

[12] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," 2020, *arXiv:2012.12877*.

[13] Y. Li, H. Mao, R. Girshick, and K. He, "Exploring plain vision transformer backbones for object detection," 2022, *arXiv:2203.16527*.

[14] R. Ranftl, A. Bochkovskiy, and V. Koltun, "Vision transformers for dense prediction," 2021, *arXiv:2103.13413*.

[15] R. Strudel, R. Garcia, I. Laptev, and C. Schmid, "Segmenter: Transformer for semantic segmentation," 2021, *arXiv:2105.05633*.

[16] F. Li, H. Zhang, H. Xu, S. Liu, L. Zhang, L. M. Ni, and H.-Y. Shum, "Mask DINO: Towards a unified transformer-based framework for object detection and segmentation," 2022, *arXiv:2206.02777*.

[17] K. Lee, H. Chang, L. Jiang, H. Zhang, Z. Tu, and C. Liu, "ViTGAN: Training GANs with vision transformers," 2021, *arXiv:2107.04589*.

[18] A. Parvaiz, M. A. Khalid, R. Zafar, H. Ameer, M. Ali, and M. M. Fraz, "Vision transformers in medical computer vision—A contemplative retrospection," 2022, *arXiv:2203.15269*.

[19] H. Chen, C. Li, G. Wang, X. Li, M. M. Rahaman, H. Sun, W. Hu, Y. Li, W. Liu, C. Sun, S. Ai, and M. Grzegorzek, "GasHis-transformer: A multi-scale visual transformer approach for gastric histopathological image detection," *Pattern Recognit.*, vol. 130, Oct. 2022, Art. no. 108827, doi: 10.1016/j.patcog.2022.108827.

[20] A. George and S. Marcel, "On the effectiveness of vision transformers for zero-shot face anti-spoofing," 2020, *arXiv:2011.08019*.

[21] B. M. Alejo, "Unconstrained ear recognition using transformers," *Jordanian J. Comput. Inf. Technol.*, vol. 7, no. 4, pp. 326–336, Dec. 2021, doi: 10.5455/jjcit.71-1627981530.

[22] A. Uhl, C. Busch, S. Marcel, and R. Veldhuis, *Handbook of Vascular Biometric*. Cham, Switzerland: Springer, 2020.

[23] A. Kumar and Y. Zhou, "Human identification using finger images," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 2228–2244, Apr. 2012, doi: 10.1109/TIP.2011.2171697.

[24] Y. Yin, L. Liu, and X. Sun, "SDUMLA-HMT: A multimodal biometric database," in *Biometric Recognition*. Springer, 2011, pp. 260–268, doi: 10.1007/978-3-642-25449-9_33.

[25] P. Tome, M. Vanoni, and S. Marcel, "On the vulnerability of palm vein recognition to spoofing attacks," in *Proc. Int. Conf. Biometrics Special Interest Group (BIOSIG)*, Darmstadt, Germany, May 2015, pp. 1–10.

[26] 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, Jun. 2014, doi: 10.1016/j.eswa.2013.11.033.

[27] C. Kauba, B. Prommegger, and A. Uhl, "Combined fully contact-less finger and hand vein capturing device with a corresponding data set," *Sensors*, vol. 19, no. 22, pp. 1–25, 2019, doi: 10.3390/s19225014.

[28] 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 Security*, vol. 17, pp. 2030–2043, 2022, doi: 10.1109/TIFS.2022.3175599.

[29] Y. Lu, S. J. Xie, S. Yoon, Z. Wang, and D. S. Park, "An available database for the research of finger vein recognition," in *Proc. 6th Int. Congr. Image Signal Process. (CISP)*, Hangzhou, China, Dec. 2013, pp. 410–415, doi: 10.1109/CISP.2013.6744030.

[30] 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 Security*, vol. 15, pp. 2641–2654, 2020, doi: 10.1109/TIFS.2020.2971144.

[31] C. Kauba, B. Prommegger, and A. Uhl, "Combined fully contactless finger and hand vein capturing device with a corresponding dataset," *Sensors*, vol. 19, no. 22, p. 5014, Nov. 2019, doi: 10.3390/s19225014.
[32] Social Innovation. *Hitachi Releases a Contactless Finger Vein Authentication Unit to Answer Demands for New Normal*. Accessed: Jan. 12, 2023. [Online]. Available: https://social-innovation.hitachi/en/article/touchless-finger-vein/
[33] P. Tome and S. Marcel, "On the vulnerability of palm vein recognition to spoofing attacks," in *Proc. Int. Conf. Biometrics*, Phuket, Thailand, May 2015, pp. 319–325, doi: 10.1109/ICB.2015.7139056.
[34] X. Liang, D. Zhang, G. Lu, Z. Guo, and N. Luo, "A novel multicamera system for high-speed touchless palm recognition," *IEEE Trans. Syst., Man, Cybern., Syst.*, vol. 51, no. 3, pp. 1534–1548, Mar. 2021, doi: 10.1109/TSMC.2019.2898684.
[35] A. Yuksel, L. Akarun, and B. Sankur, "Hand vein biometry based on geometry and appearance methods," *IET Comput. Vis.*, vol. 5, no. 6, pp. 398–406, 2011, Accessed: Jan. 12, 2023. [Online]. Available: https://www.semanticscholar.org/paper/Hand-vein-biometry-based-on-geometry-and-appearance-Yuksel-Akarun/318d91a7c3856390e529fa82540712b8c6c46ff9
[36] C. Kauba and A. Uhl, "Shedding light on the veins–reflected light or transillumination in hand-vein recognition," in *Proc. Int. Conf. Biometrics (ICB)*, Gold Coast, QLD, Australia, Feb. 2018, pp. 283–290, doi: 10.1109/ICB2018.2018.00050.
[37] O. Toygar, F. O. Babalola, and Y. Bitirim, "FYO: A novel multimodal vein database with palmar, dorsal and wrist biometrics," *IEEE Access*, vol. 8, pp. 82461–82470, 2020.
[38] R. Kabacinski and K. Kowalski, "Vein pattern database and benchmark results," *Electron. Lett.*, vol. 47, no. 20, pp. 1127–1128, 2011.
[39] R. Raghavendra and C. Busch, "A low cost wrist vein sensor for biometric authentication," in *Proc. IEEE Int. Conf. Imag. Syst. Techn. (IST)*, Chania, Greece, Oct. 2016, pp. 201–205.
[40] R. Garcia-Martin and R. Sanchez-Reillo, "Wrist vascular biometric recognition using a portable contactless system," *Sensors*, vol. 20, no. 5, p. 1469, Mar. 2020.
[41] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, *arXiv:1409.1556*.
[42] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Dec. 2016, pp. 770–778.
[43] W. Liu, W. Li, L. Sun, L. Zhang, and P. Chen, "Finger vein recognition based on deep learning," in *Proc. 12th IEEE Conf. Ind. Electron. Appl. (ICIEA)*, Siem Reap, Cambodia, Jun. 2017, pp. 205–210, doi: 10.1109/ICIEA.2017.8282842.
[44] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 2, pp. 84–90, Jun. 2012, doi: 10.1145/3065386.
[45] H. Hong, M. Lee, and K. Park, "Convolutional neural network-based finger-vein recognition using NIR image sensors," *Sensors*, vol. 17, no. 6, p. 1297, Jun. 2017, doi: 10.3390/s17061297.
[46] S.-Y. Jhong, P.-Y. Tseng, N. Siriphockpirom, C.-H. Hsia, M.-S. Huang, K.-L. Hua, and Y.-Y. Chen, "An automated biometric identification system using CNN-based palm vein recognition," in *Proc. Int. Conf. Adv. Robot. Intell. Syst. (ARIS)*, Aug. 2020, pp. 1–6, doi: 10.1109/ARIS50834.2020.9205778.
[47] 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. 91–178, 2019.
[48] 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, p. 2296, Jul. 2018.
[49] 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.
[50] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Honolulu, HI, USA, Jul. 2017, pp. 2261–2269.
[51] M. I. Obayya, M. El-Ghandour, and F. Alrowais, "Contactless palm vein authentication using deep learning with Bayesian optimization," *IEEE Access*, vol. 9, pp. 1940–1957, 2021.
[52] *CASIA Multi-Spectral Palmprint Database*. Accessed: Jan. 12, 2023. [Online]. Available: http://biometrics.idealtest.org/
[53] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," 2020, *arXiv:2005.12872*.
[54] H. Wang, Y. Zhu, B. Green, H. Adam, A. Yuille, and L.-C. Chen, "Axial-DeepLab: Stand-alone axial-attention for panoptic segmentation," 2020, *arXiv:2003.07853*.
[55] Z. Huang and C. Guo, "Robust finger vein recognition based on deep CNN with spatial attention and bias field correction," in *Proc. 12th Int. Conf. Adv. Comput. Intell. (ICACI)*, Dali, China, Aug. 2020, pp. 614–619, doi: 10.1109/ICACI49185.2020.9177758.
[56] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," 2015, *arXiv:1505.04597*.
[57] J. Huang, M. Tu, W. Yang, and W. Kang, "Joint attention network for finger vein authentication," *IEEE Trans. Instrum. Meas.*, vol. 70, pp. 1–11, 2021, doi: 10.1109/TIM.2021.3109978.
[58] H. Lu, Y. Li, C. Zhao, W. Liu, Y. Li, and N. Ma, "A novel finger-vein recognition approach based on vision transformer," in *Proc. Int. Conf. Frontiers Electron., Inf. Comput. Technol.*, May 2021, pp. 1–6, doi: 10.1145/3474198.3478217.
[59] J. Huang, W. Luo, W. Yang, A. Zheng, F. Lian, and W. Kang, "FVT: Finger vein transformer for authentication," *IEEE Trans. Instrum. Meas.*, vol. 71, pp. 1–13, 2022, doi: 10.1109/TIM.2022.3173276.
[60] A. Rosebrock. *Deep Learning For Computer Vision*. PyImageSearch. Accessed: Jan. 12, 2023. [Online]. Available: https://www.pyimagesearch.com/2018/09/24/opencv-face-recognition/
[61] Hromi. (Oct. 31, 2022). *Hromi/SMILEsmileD*. GitHub. Accessed: Jan. 12, 2023. [Online]. Available: https://github.com/hromi/SMILEsmileD
[62] A. Steiner, A. Kolesnikov, X. Zhai, R. Wightman, J. Uszkoreit, and L. Beyer, "How to train your ViT? Data, augmentation, and regularization in vision transformers," 2021, *arXiv:2106.10270*.
[63] *TensorFlow Hub*. Accessed: Jan. 12, 2023. [Online]. Available: https://tfhub.dev/sayakpaul/collections/vision
[64] C. Militello, V. Conti, S. Vitabile, and F. Sorbello, "Embedded access points for trusted data and resources access in HPC systems," *J. Supercomput.*, vol. 55, no. 1, pp. 4–27, Jan. 2010.

RAUL GARCIA-MARTIN received the bachelor's degree in industrial electronics and automation engineering and the master's degree in electronics systems and applications engineering from the University Carlos III of Madrid (UC3M), in 2016 and 2018, respectively. At the same time, he acquired experience in industrial applications as a software and hardware developer in several companies. He is currently working on his Ph.D. thesis with the University Group for Identification Technologies (GUTI, UC3M), researching vascular biometric recognition and collaborating with the Human–Computer Interaction Engineering Group, Massachusetts Institute of Technology (MIT), in Augmented Reality applications.

RAUL SANCHEZ-REILLO (Senior Member, IEEE) received the Ph.D. degree in telecommunication engineering from the Universidad Politecnica de Madrid, in 2000. He is currently a Full Professor with the University Carlos III of Madrid (UC3M). He is also the Head of the University Group for Identification Technologies (GUTI). His research and development group is involved in project development related to identification technologies, either by the user of secure elements (such as smartcards) and/or by using biometrics. In addition to research and development activities, he has also managed projects concerning a broad range of applications, from social security services till financial payment methods. He has taken part in European projects, being the WP leader in most of them. In 2009, he founded IDTestingLab, an evaluation laboratory for identification products. He is also a member of SC17, SC27, and SC37 standardization committees, holding the international secretary of SC17 WG11 and SC37 WG2.