DerNeXt: A Transformer-based Adversarial Robustness Enhanced Model for Palm Vein Recognition

Miaoji Zheng[1,†], Yong Zhong[1,†], Fen Liu[1,*], Qiyuan Sun[1],
Tufeng Xian[1], Weidong Wu[1], Zhiliang Zhang[1], Haitao Huang[2]
[1]School of Computer Science and Artificial Intelligence, Foshan University, Foshan, China
[2]Guangdong Chankong Big Data Technology Company, Foshan, China
2112453056@stu.fosu.edu.cn, zhongyong@fosu.edu.cn, liufen@fosu.edu.cn, 944376274@qq.com,
2112453044@stu.fosu.edu.cn, 2112453039@stu.fosu.edu.cn, 2112453043@stu.fosu.edu.cn, 569446411@qq.com

***Abstract*—Palm vein biometrics has gained considerable attention due to its advantages of being difficult to forge, highly private, and well-accepted by users. With the success of deep neural networks in visual tasks, deep learning-based palm vein recognition models have also achieved remarkable performance. However, like many visual models, existing deep learning-based palm vein recognition systems are highly vulnerable to adversarial attacks. Attackers can cause a sharp decline in model performance by adding carefully designed subtle perturbations to input images. To mitigate this issue, we propose a novel adversarial robustness-enhanced palm vein recognition model named DerNeXt. In this study, we firstly introduce the TrsanNext architecture to the field of palm vein recognition. To further enhance the stability of the input distribution to the channel mixer in the above architecture, we propose NGLU, a normalized gated linear unit. We also design DEM, a defensive efficient attention module, placed at the end of the network to ensure the quality stability and defensive capability of palm vein features. Finally, we employ an orthogonal classifier with a dense orthogonal weight matrix to refine decision boundaries, combined with a newly designed orthogonal discriminant loss to enhance adversarial robustness while maintaining high identification accuracy. We evaluate DerNeXt on three public palm vein datasets in terms of identification accuracy, equal error rate, and white-box attack experiments. The results demonstrate that DerNeXt achieves state-of-the-art recognition performance while exhibiting superior defense capability against adversarial attacks compared to other baseline models.**

***Index Terms*—Deep learning, palmvein recognition, adversarial attacks, transformer**

I. Introduction

Palm vein recognition technology has attracted extensive attention in the biometrics field. In recent years, many studies have made significant contributions to improving the accuracy of deep learning-based palm vein recognition. For example, by designing and optimizing network architectures [1]; expanding the region of interest (ROI) to capture more useful palm vein features [2]. To facilitate deployment on devices with limited computational power, many researchers have also made prominent contributions to model lightweighting. Examples include a novel low-cost smartphone-based palm vein recognition system that uses only three-channel red, green, and blue (RGB) images [3]; and a lightweight generative self-supervised contrastive learning (GSCL) method to address the scarcity of vein data [4].

Although existing methods can achieve lightweight designs while maintaining excellent recognition performance, they typically focus solely on palm vein feature learning and matching. Studies have shown that deep learning models are vulnerable to carefully crafted adversarial examples [5]. Among existing research on defense measures against adversarial attacks, using an external framework to preprocess images is a mainstream and effective approach. For instance, Li et al. [6] proposed a novel defense framework named VeinGuard, composed of a local transformer-based generative adversarial network (GAN) and a purifier, designed to use an additional defense framework to counter adversarial palm vein image attacks. Subsequently, Qin et al. [7] proposed MsMemoryGAN using a multi-scale autoencoder with a memory module, aiming to filter out perturbations in adversarial samples before recognition. Fei et al. [8] proposed a reconstruction network consisting of three sub-networks for palm vein feature exploration, feature optimization, and image reconstruction, intended for palm vein image protection. While these methods are effective, they introduce considerable pre-processing computational overhead—often exceeding that of the recognizer itself—which runs counter to the trend of deploying lightweight palm vein recognition systems.

Our Contributions. To address this, we propose an adversarial-robust palm vein recognition model that enhances defense capability against adversarial attacks solely through targeted model architecture design. Its innovations are as follows.

- We propose—for the first time in palm vein recognition—a model architecture-oriented defense strategy, offering a novel perspective in this field.

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 62472105, in part by the Natural Science Foundation of Guangdong Province under Grants 2025A1515011385 and 2024A1515010186, Special Program for Priority Fields in Guangdong Higher Education Institutions under Grant 2025ZDZX1033 as well as Guangdong Basic and Applied Basic Research Foundation under Grant 2024A1515110038.

- We construct an adversarial-robust palm vein recognition model named DerNeXt. Based on the TransNext backbone architecture, we improve the original gated unit by proposing a Normalized Gated Linear Unit (NGLU) to enhance feature fusion and the stability of the input distribution. Furthermore, we innovatively introduce a refined Mercer kernel-based Defensive Efficient Attention Module (DEM) to ensure the stability of palm vein feature quality and strengthen feature defense capability.
- To further improve adversarial robustness, we use a robust classification layer built with a dense orthogonal weight matrix and propose an orthogonal discriminant loss. This approach optimizes feature distribution by constructing a dense orthogonal weight matrix, ensuring intra-class compactness and inter-class separation, refining decision boundaries, and thereby enhancing adversarial robustness.

Through the design described above, our proposed DerNeXt achieves identification accuracies of 98.59%, 99.67%, and 99.92% on the three public vein databases—the CASIA Multi-Spectral Palmprint Database, The Hong Kong Polytechnic University (PolyU) Multi-Spectral Palmprint Database, and the Tongji University (TongjiU) Contactless Palm Vein Database—with Equal Error Rates (EER) of 0.56%, 0.20%, and 0.12%, respectively.

II. The Palm Vein Recognition Model

In this chapter, we will provide a comprehensive introduction to the proposed DerNeXt. It utilizes TransNeXt as its backbone architecture and incorporates several innovative modifications, as illustrated in Fig. 1. This will be followed by a detailed presentation of our proposed improvements.

A. DerNeXt

Inspired by biological visual perception mechanisms, TransNeXt exhibits strong global perceptual ability, local modeling capability, and model robustness, making it highly suitable for handling the complex global structures and local details present in palm vein images, and demonstrating excellent feature decoupling capability [9]. Therefore, we select TransNeXt as the backbone architecture. However, when directly applied to palm vein recognition tasks, some information loss may occur during the decoupling process, potentially affecting recognition results. Benefiting from the versatility of TransNeXt, we can flexibly modify it to address these issues. Our proposed DerNeXt adopts a hierarchical Transformer architecture, functioning as a hierarchical visual encoder. Its overall pipeline follows the design of TransNeXt-Micro and consists of four stages. The input image I first undergoes a stride-4 patch embedding layer for downsampling. It is then sequentially passed through four stages, each built by stacking $[2, 2, 15, 2]$ DerNeXt Blocks respectively. A downsampling operation follows each stage, ultimately yielding a high-level feature representation F. The overall architecture of the model is illustrated in Fig. 1(a).

Subsequently, the refined features are fed into a classifier constructed with a dense orthogonal weight matrix. The innovative modifications in DerNeXt will be introduced in the following sections, focusing on the NGLU, the DEM, and the orthogonal weight classifier, respectively.

B. Normalized Gated Linear Unit

Although ConvGLU can effectively capture positional information from zero-padding through DWConv, its input distribution may suffer from instability [10]. To address this, we propose the NGLU. Its core improvement is the addition of a LayerNorm operation after the DWConv and the activation function. The purpose of the depthwise convolution is to provide position-aware gating signals based on local neighborhoods. The normalization operation acts as a regularizer to some extent. Therefore, applying LayerNorm before its input establishes a stable, zero-mean, unit-variance baseline for these spatial features. This can lead to higher quality and more consistent spatial features extracted by the subsequent DWConv operation, potentially generating more accurate and smoother gating masks. The operation of NGLU is defined as

$$\begin{aligned}\mathrm{NGLU}(X) = &\Big[\sigma\Big(\mathrm{DWConv}\big(\mathrm{LN}(\mathrm{Linear}_1(X))\big)\Big) \odot X\Big] \\ &\oplus \mathrm{Linear}_2(X).\end{aligned} \tag{1}$$

As illustrated in Fig. 1(b), the architecture of the NGLU bears resemblance to the ConvGLU structure. The principal distinction resides in the gating pathway, wherein the input X initially undergoes transformation via the linear layer Linear_1, succeeded by Layer Normalization (LN). The processed features then traverse through DWConv operation, culminating in the activation function σ.

C. Defensive Efficient Attention Module

Deep network features encapsulate high-level semantic information and exhibit a direct correlation with the final classification decision, rendering them vulnerable to adversarial attacks. To mitigate this vulnerability, we propose the DEM, integrated within the fourth stage of the DerNeXt architecture. As depicted in Fig. 1(c), the DEM processes the input feature map F'—generated by the preceding three stages—with dimensions (h, w, c) corresponding to height, width, and channel size, respectively. The feature F' initially traverses a PreAttentionMLP layer, which expands the channel dimensionality by a factor of three. This expanded representation is subsequently partitioned into three distinct attention computation pathways: query (Q), key (K), and value (V), facilitated through patch embedding. The core innovation of our approach lies in the construction of a Mercer kernel-based operation [11], founded upon the Mercer theorem. Specifically, we project the Q and K components using a hyperbolic cosine-based mapping function, formally defined as (using Q as an example)

$$\phi(Q) = \left(1, \frac{(Q-\bar{Q})^2}{\sqrt{2!}}, \frac{(Q-\bar{Q})^4}{\sqrt{4!}}, \cdots, \frac{(Q-\bar{Q})^{2i}}{\sqrt{(2i)!}}\right), \tag{2}$$

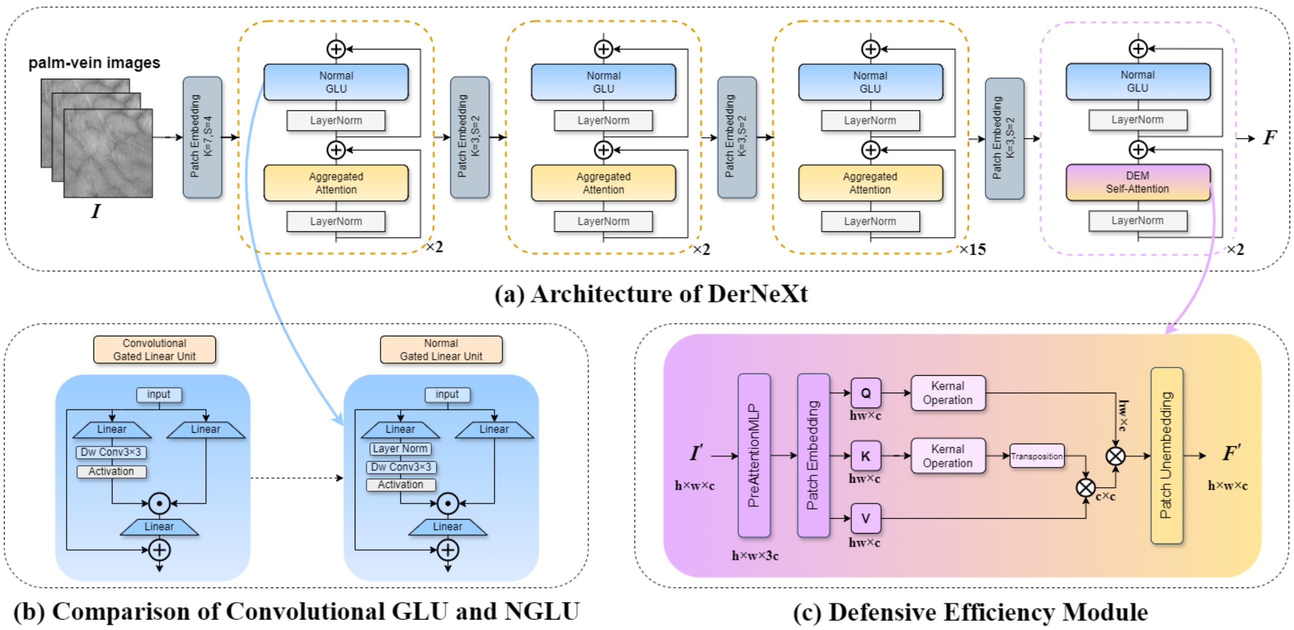

Fig. 1. Details of the proposed DerNeXt.

where $\bar{Q}$ represents the mean value of Q, and i denotes the expansion order. The corresponding kernel operation can be defined via Taylor expansion as

$$K_{(Q,K)} = \sum_{i=0}^{\infty} \frac{[(Q-\bar{Q})(K-\bar{K})]^{2i}}{(2i)!}. \tag{3}$$

By employing the kernel operation, we can not only enhance feature robustness but also improve computational efficiency. As noted in [11], by exchanging the order of multiplication—first computing the product of the projected Q and K—the computational complexity is reduced from $O(N^2)$ to $O(N)$, significantly boosting computational efficiency. Therefore, the self-attention matrix is calculated as

$$\mathrm{Att}(Q,K,V) = \mathrm{softmax}\left(\frac{\phi(K)^T \cdot V}{\sqrt{C}}\right)\phi(Q), \tag{4}$$

where C is a scaling factor and T denotes the transpose operation. Finally, the output features are obtained via patch unembedding.

D. Orthogonal Classifier

To enable DerNeXt to maintain high accuracy while significantly enhancing its intrinsic robustness against adversarial perturbations, we adopt the concept of the orthogonally robust classifier proposed by Xu et al. [12]. The core idea of this method is to guide the feature extraction network to learn features that are compact within classes and dispersed between classes by predefining a weight matrix with a specific structure for the classification layer. Consequently, good robustness can be achieved even under standard training.

We discard the conventional approach of randomly initializing the classification layer weights and training them throughout. Instead, during the model initialization phase, we explicitly construct a fixed classification layer weight matrix $W \in \mathbb{R}^{P\times K}$ whose column vectors are mutually orthogonal and of equal length, based on the final output feature dimension P and the number of classes K. The procedure is presented in Algorithm 1.

Algorithm 1 The Construction of a Dense Orthogonal Classifier (adapted from [12])

Require: The number of classes K, the length s, and the feature dimension P.
1: Set $T = \log_2 P$ and $M^{(0)} = 2^{-T/2}s$.
2: **for** $t = 1$ to T **do**
3: Apply formula (3.8) to recursively construct a series of $M^{(t)}$.
4: **end for**
5: Choose the first K columns of $M^{(T)}$ to build the weight matrix W.
Ensure: The weight matrix W for the classifier.

We assume that the feature dimension satisfies $P = 2^T$, where T is a given integer. The vector length s is left as a hyperparameter to be given.In this work, the length of the weight vector s is set to 10. Set $M^{(0)} = 2^{-T/2}s$. For $t = 1, \cdots, T$, we recursively define

$$M^{(t)} = \begin{bmatrix} M^{(t-1)} & -M^{(t-1)} \\ M^{(t-1)} & M^{(t-1)} \end{bmatrix}. \tag{5}$$

The matrix W generated by the algorithm exhibits two crucial mathematical properties.

- **Equal Length and Orthogonality** All column vectors $\mathbf{w}_i$ of W satisfy $\|\mathbf{w}_i\|_2 = s$, and for any $i \neq j$, we have $\mathbf{w}_i^T \mathbf{w}_j = 0$
- **Density** All elements of the matrix W share the same absolute value $2^{-T/2}s$. It is a fully dense matrix, avoiding the structural redundancy and potential underfitting issues that may arise from the sparsity of the weight matrix in methods such as [13], [14].

Once the orthogonal classifier weight matrix W is constructed, we integrate it into the model. During the subsequent training process, the weights W of the classification layer are frozen and no longer participate in gradient updates. The training objective is only for the parameters θ of the feature encoder $f(\cdot;\theta)$, aiming to minimize the following intra-class compactness loss

$$\mathcal{L}_{\text{Intra}} = \frac{1}{N} \sum_{i=1}^{N} \left\| f\left(x_i\right) - w_{y_i} \right\|_2^2, \tag{6}$$

where (x_i, y_i) denotes the training sample and its ground-truth label, and W_{y_i} is the column vector corresponding to class y_i in the weight matrix W. This training paradigm forces feature representations to stay away from decision boundaries, thereby endowing DerNeXt with a stronger inherent resistance to subtle adversarial perturbations. Although the fixed orthogonal classifier guides the encoder to learn a feature distribution with intrinsic robustness, solely optimizing the distance from features to the fixed class centers may be insufficient for maximizing the discriminativeness of the classification decision boundaries. To address this, we introduce the cross-entropy loss in a collaborative, rather than substitutive, manner. The cross-entropy loss focuses on precisely delineating decision boundaries to ensure high classification accuracy, while the intra-class compactness loss is responsible for shaping a robust geometric structure of the feature space. The two losses complement each other, jointly guiding the model's learning process, which results in the proposed orthogonal discriminant loss. The total loss function is formulated as

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{CE}} + \lambda \cdot \mathcal{L}_{\text{intra}}, \tag{7}$$

where $\mathcal{L}_{\text{CE}}$ is the standard cross-entropy loss, $\mathcal{L}_{\text{intra}}$ is the intra-class compactness loss, and the hyperparameter λ balances the importance of these two terms.

III. Experiments

To comprehensively evaluate the performance of the proposed DerNeXt, we conduct extensive experiments from multiple perspectives. This section begins by introducing the datasets and experimental settings. Subsequently, we compare the identification performance of DerNeXt with several state-of-the-art methods on three public palm vein datasets. To validate the adversarial robustness, we further assess the model's defensive capability under various white-box attacks. Finally, ablation studies are performed to analyze the contribution of each proposed component.

A. Datasets

To evaluate the performance of DerNeXt, we conducted extensive experiments on three public vein databases: the CASIA Multi-Spectral Palmprint Database [15] from the Institute of Automation, Chinese Academy of Sciences; the PolyU Multi-Spectral Palmprint Database [16] from The Hong Kong Polytechnic University; and the TongjiU Contactless Palm Vein Database [17] from Tongji University. For the original images, we first determined key palm points using distance transform and boundary tracing. The palm orientation was then corrected by calculating the angle between the line connecting two key points and the vertical direction. Finally, a square region containing the core part of the palm was cropped as the extracted palm vein ROI.

B. Experimental Settings

The three databases were split into training and test sets to simulate a real-world scenario: for each subject, the first collection batch was used for training and the second for testing. In CASIA, this resulted in 200 classes (100 subjects, 2 palms each) with 600 training and 600 test images. PolyU comprises 500 classes (250 subjects) with 3,000 images per set. TongjiU contains 600 classes (300 subjects) with 6,000 images per set. To mitigate data scarcity, image augmentation was applied, expanding to approximately 100 images per class.

Following transfer learning practice, all models were pre-trained on CIFAR-100 for initialization. We used SGDR for both pre-training and fine-tuning. During pre-training, the learning rate was set to 0.01 (min 0.0005), weight decay to 0.05, batch size to 32, and training lasted for 200 epochs. For fine-tuning, the learning rate was 0.001 (min 0.000005) with other hyperparameters kept the same, trained for 100 epochs. All experiments were conducted using PyTorch on an NVIDIA RTX 3090 GPU. For quantitative evaluation, five advanced deep learning-based recognizers—EfficientViT [18], MobileNetV4 [19], RepViT [20], MSVT [1], and MPSNet [3]—were employed for comparative study. Simultaneously, to verify the robustness of the proposed method against adversarial attacks, several classic white-box attack techniques, including FGSM [5] , RAND+FGSM [3] , and PGD [21], were utilized. These methods were used to generate adversarial samples on the test images of the three datasets, which were then fed into the trained classifiers for comparative evaluation.

We employed identification accuracy and Equal Error Rate (EER) to evaluate the proposed method. Furthermore, the identification accuracy of the proposed method on adversarial samples is also reported in Table II to reflect the adversarial robustness of the recognition model. The identification accuracy and EER of various methods on the three datasets are shown in Table I.

C. Identification Performance of DerNeXt

As shown in the Table I, compared with the other five methods, the proposed DerNeXt achieves the highest recognition accuracy and the lowest EER on each dataset. This strong performance can be explained by the following reasons.

- DerNeXt adopts TransNeXt as its backbone, an architecture that combines the global perception capability of Transformers with the local modeling strength of CNNs. This hybrid design is well-suited for palm vein images, which contain both global vascular topology and fine local textures.
- The proposed NGLU provides high-quality gating signals, which can more precisely activate key features related to palm vein textures while suppressing interference from background noise and non-critical regions in the samples. This enables the model to perform feature fusion stably and robustly.

TABLE I
IDENTIFICATION ACCURACY AND EER FOR DIFFERENT METHODS

Dataset	Methods	Accuracy (%)	EER (%)
CASIA	MobileNetV4-Conv-L	97.66	1.27
	MPSNet	98.07	1.24
	RepViT-M2.3	98.44	1.02
	EfficientViT-M5	98.44	0.94
	MSVT	98.12	0.89
	DerNeXt	**98.59**	**0.56**
PolyU	MobileNetV4-Conv-L	99.33	0.34
	MPSNet	98.33	0.47
	RepViT-M2.3	99.42	0.33
	EfficientViT-M5	99.17	0.33
	MSVT	99.50	0.27
	DerNeXt	**99.67**	**0.20**
TongjiU	MobileNetV4-Conv-L	99.83	0.20
	MPSNet	99.79	0.35
	RepViT-M2.3	99.75	0.21
	EfficientViT-M5	99.33	0.92
	MSVT	99.88	0.15
	DerNeXt	**99.92**	**0.12**

D. Results on White-Box Attacks

We employed the attack strength parameter ϵ to perform FGSM, RAND+FGSM, and PGD attacks. The experiments continued to use the five aforementioned methods for comparison. We fixed ϵ at 0.01 to generate adversarial samples, which were then fed into all methods for recognition. Table II presents the experimental results for each method on the three databases. The experimental results show that under FGSM attacks, DerNeXt significantly outperforms other models across all datasets in terms of defensive capability. This advantage becomes even more pronounced under the stronger PGD attack. The baseline models generally see their accuracy plummet to below 1% or even 0% under PGD attacks. In contrast, DerNeXt maintains recognition rates of 9.17%, 11.25%, and 14.08% on the three datasets, respectively. This superior performance can be attributed to the following factors:

- The TransNeXt architecture inherently possesses strong feature extraction capabilities. When combined with the NGLU, it enables stable and robust feature extraction, thereby reducing the burden on subsequent processing stages.
- DerNeXt integrates the DEM module at the high-level semantic stage. This module, utilizing a Mercer kernel-based attention mechanism, strengthens resistance to perturbations in deep features. This allows the model's high-level semantic representations to remain relatively intact even against strong attacks like PGD, which is key to maintaining double-digit accuracy.
- As discussed in the theoretical analysis, the orthogonal classifier and the orthogonal discriminant loss jointly shape a feature distribution characterized by "high intra-class compactness and strict inter-class separation." In the context of adversarial attacks, this means an attacker would need to generate perturbations with significantly larger magnitudes to push a sample's features out of its highly compact class cluster and across the decision boundary into another cluster.

E. Ablation Experiments

To validate the effectiveness of DerNeXt, we conduct ablation studies on three palm vein datasets. Results in Table III show that sequentially adding the NGLU, DEM, and orthogonal classifier to the TransNeXt backbone progressively enhances adversarial robustness, with clear synergistic effects. Under strong PGD attacks, the baseline achieves only 1.87% average accuracy. Adding NGLU alone yields limited gain (2.62%) but stabilizes training. Incorporating DEM boosts accuracy to 7.69%, demonstrating its effectiveness in high-level semantic defense. The full DerNeXt model achieves the best performance, with 11.50% accuracy under PGD and 43.62% under RAND+FGSM attacks. The orthogonal classifier geometrically hardens decision boundaries, while NGLU and DEM respectively stabilize features and defend semantics, forming a multi-layer defense system whose overall gain exceeds the sum of individual contributions.

IV. CONCLUSION

We propose DerNeXt, an adversarial robustness-enhanced palm vein recognition model. Built on the TransNeXt backbone, it integrates the Normalized Gated Linear Unit (NGLU), the Defensive Efficient Attention Module (DEM), and an orthogonal classifier to form a multi-layer defense system. Future work will explore class-incremental learning to enhance the model's adaptability in open-world settings where new classes continuously emerge.

REFERENCES

[1] H. Qin, C. Gong, Y. Li, et al., "Label enhancement-based multiscale transformer for palm-vein recognition," *IEEE Trans. Instrum. Meas.*, vol. 72, pp. 1–17, 2023.

[2] D. Luo, J. Huang, W. Yang, et al., "RSNet: Region-specific network for contactless palm vein authentication," *IEEE Trans. Inf. Forensics Security*, 2025.

[3] S. J. Horng, D. T. Vu, T. V. Nguyen, et al., "Recognizing palm vein in smartphones using RGB images," *IEEE Trans. Ind. Informat.*, vol. 18, no. 9, pp. 5992–6002, 2021.

[4] W. F. Ou, L. M. Po, X. F. Huang, et al., "GSCL: Generative self-supervised contrastive learning for vein-based biometric verification," *IEEE Trans. Biom., Behav., Identity Sci.*, vol. 6, no. 2, pp. 230–244, 2024.

TABLE II
IDENTIFICATION ACCURACY(%) OF VARIOUS METHODS UNDER FGSM, RAND+FGSM, AND PGD WHITE-BOX ATTACKS

Dataset	Methods	No Attack	FGSM ($\varepsilon = 0.01$)	RAND+FGSM ($\varepsilon = 0.01$, $\alpha = 0.005$)	PGD ($\varepsilon = 0.01$)
CASIA	MobileNetV4-Conv-L	97.66	7.12	13.55	0
	MPSNet	98.07	9.41	13.49	0.20
	RepViT-M2.3	98.44	10.16	9.90	0.16
	EfficientViT-M5	98.42	8.46	13.33	0.21
	MSVT	98.12	12.96	16.12	1.34
	DerNeXt	**98.59**	**28.97**	**30.26**	**9.17**
PolyU	MobileNetV4-Conv-L	99.33	9.14	8.13	0.47
	MPSNet	98.33	13.32	12.44	0.14
	RepViT-M2.3	99.42	3.66	10.28	1.36
	EfficientViT-M5	99.17	6.62	7.23	0.01
	MSVT	99.50	13.65	14.36	2.62
	DerNeXt	**99.67**	**30.94**	**39.97**	**11.25**
TongjiU	MobileNetV4-Conv-L	99.83	12.87	16.11	0.59
	MPSNet	99.79	16.64	18.89	1.13
	RepViT-M2.3	99.75	14.43	7.33	0.77
	EfficientViT-M5	99.33	19.19	16.34	0.31
	MSVT	99.88	22.44	15.66	1.10
	DerNeXt	**99.92**	**36.19**	**60.62**	**14.08**

TABLE III
IDENTIFICATION ACCURACY (%) OF ABLATION EXPERIMENTS ON THREE DATABASES UNDER FGSM, RAND+FGSM, AND PGD WHITE-BOX ATTACKS

Dataset	Methods	No Attack	FGSM ($\varepsilon = 0.01$)	RAND+FGSM ($\varepsilon = 0.01$, $\alpha = 0.005$)	PGD ($\varepsilon = 0.01$)
CASIA	TransNeXt	98.33	10.34	14.58	1.33
	TransNeXt+NGLU	98.56	9.62	16.08	1.26
	TransNeXt+NGLU+DEM	98.54	19.66	24.61	4.67
	DerNeXt	98.59	28.97	30.26	9.17
PolyU	TransNeXt	99.14	14.21	13.55	1.99
	TransNeXt+NGLU	99.17	15.01	13.49	2.61
	TransNeXt+NGLU+DEM	99.52	21.94	9.90	7.73
	DerNeXt	99.67	30.94	39.97	11.25
TongjiU	TransNeXt	99.67	20.66	19.03	2.30
	TransNeXt+NGLU	99.72	20.62	22.48	3.98
	TransNeXt+NGLU+DEM	99.90	30.34	48.66	10.66
	DerNeXt	99.92	36.19	60.62	14.08

[5] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," in *Proc. Int. Conf. Learn. Represent.*, San Diego, CA, USA, 2015, pp. 1–10.

[6] Y. Li, S. Ruan, H. Qin, et al., "Transformer based defense GAN against palm-vein adversarial attacks," *IEEE Trans. Inf. Forensics Security*, vol. 18, pp. 1509–1523, 2023.

[7] H. Qin, Y. Fu, H. Zhang, et al., "MsMemoryGAN: A multi-scale memory GAN for palm-vein adversarial purification," *arXiv preprint arXiv:2408.10694*, 2024.

[8] L. Fei, J. Yang, W. K. Wong, et al., "Palm-vein images reconstruction against adversarial attacks," in *Proc. IEEE Int. Conf. Acoust., Speech, Signal Process.*, 2025, pp. 1–5.

[9] D. Shi, "Transnext: Robust foveal visual perception for vision transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 17773–17783.

[10] M. Sandler, A. Howard, M. Zhu, et al., "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4510–4520.

[11] M. Zhang, C. Zhang, Q. Zhang, et al., "ESSAformer: Efficient transformer for hyperspectral image super-resolution," in *Proc. IEEE/CVF Int. Conf. Comput. Vis.*, 2023, pp. 23073–23084.

[12] C. Xu, X. Li, and M. Yang, "An orthogonal classifier for improving the adversarial robustness of neural networks," *Inf. Sci.*, vol. 591, pp. 251–262, 2022.

[13] T. Pang, C. Du, and J. Zhu, "Max-mahalanobis linear discriminant analysis networks," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 1–10.

[14] T. Pang, K. Xu, Y. Dong, C. Du, N. Chen, and J. Zhu, "Rethinking softmax cross-entropy loss for adversarial robustness," in *Proc. Int. Conf. Learn. Represent.*, 2020.

[15] Y. Hao, Z. Sun, T. Tan, and C. Ren, "Multispectral palm image fusion for accurate contact-free palmprint recognition," in *Proc. IEEE Int. Conf. Image Process.*, Mar. 2008, pp. 281–284.

[16] D. Zhang, Z. Guo, G. Lu, L. Zhang, and W. Zuo, "An online system of multispectral palmprint verification," *IEEE Trans. Instrum. Meas.*, vol. 59, no. 2, pp. 480–490, 2009.

[17] 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*, vol. 10, no. 4, p. 78, 2018.

[18] X. Liu, H. Peng, N. Zheng, et al., "Efficientvit: Memory efficient vision transformer with cascaded group attention," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2023, pp. 14420–14430.

[19] D. Qin, C. Leichner, M. Delakis, et al., "MobileNetV4: Universal models for the mobile ecosystem," in *Proc. Eur. Conf. Comput. Vis.*, Cham, Switzerland: Springer, 2024, pp. 78–96.

[20] A. Wang, H. Chen, Z. Lin, et al., "Repvit: Revisiting mobile cnn from vit perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2024, pp. 15909–15920.

[21] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–23.