Multi-Spectral Palmprints Joint Attack and Defense With Adversarial Examples Learning

Qi Zhu, Yuze Zhou, Lunke Fei, *Senior Member, IEEE*, Daoqiang Zhang, *Senior Member, IEEE*, and David Zhang, *Life Fellow, IEEE*

***Abstract*— As an emerging biometric technology, multi-spectral palmprint recognition has attracted increasing attention in security due to its high accuracy and ease of use. Compared to single spectral case, multi-spectral palmprint model is more susceptible to the attack of adversarial examples. However, the previous adversarial example attack approaches cannot generate the most aggressive adversarial examples for multi-spectral palmprint recognition. In addition, most of them are dependent on the explicit architecture or need time-consuming queries about the network to be attacked, which significantly limits their application in the field of security. To solve the above problems, in this paper, we proposed the multi-spectral palmprints joint attack and defense framework based on multi-view adversarial examples learning. First, we respectively capture the multi-view deep common feature space for the different spectra and the discriminative feature space across the different subjects. Second, we introduce perturbation in the deep common space to achieve adversarial multi-spectral palmprints with gradient propagation. In addition, we pursue the manifold of the difference space and use it to suppress the discriminability of the recognition model with adversarial region theory. Finally, the generated adversarial examples are fed into the training model to enhance the robustness of the recognition algorithm. The experimental results on multi-spectral palmprint dataset demonstrate that the proposed multi-view joint attack approach is superior to the state-of-the-art adversarial example attack methods in attack accuracy and transferability. Moreover, the defense strategy with the adversarial examples by our method can significantly promote the robustness of multi-spectral palmprint recognition methods.**

***Index Terms*— Biometrics, multi-spectral palmprint recognition, adversarial example, security, manifold learning.**

Manuscript received 27 October 2022; revised 23 February 2023; accepted 24 February 2023. Date of publication 8 March 2023; date of current version 16 March 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62076129, Grant 62136004, Grant 62176066, and Grant 61902183; in part by the National Key Research and Development Program of China under Grant 2018YFC2001600 and Grant 2018YFC2001602; and in part by the Key Research and Development Plan of Jiangsu Province under Grant BE2022842. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Valeria Loscri. *(Corresponding authors: Daoqiang Zhang; David Zhang.)*

Qi Zhu, Yuze Zhou, and Daoqiang Zhang are with the College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China (e-mail: zhuqi@nuaa.edu.cn; zhouyuze@nuaa.edu.cn; dqzhang@nuaa.edu.cn).

Lunke Fei is with the School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China (e-mail: flksxm@126.com).

David Zhang is with the School of Science and Engineering, The Chinese University of Hong Kong, Shenzhen 518172, China (e-mail: davidzhang@cuhk.edu.cn).

Digital Object Identifier 10.1109/TIFS.2023.3254432

I. Introduction

With the development of information technology, biometric recognition technology [1], [2] plays an important role in the field of security. Among the biometric traits, multi-spectral palmprint [3], [4], [5] has attracted more and more attention, because it contains rich texture and features of the palm under multiple frequency bands for accurate identity authentication. The multi-spectral palmprint recognition model is more susceptible to the attack in actual application scenarios when it integrates more spectra in the fusion recognition model [6]. As one of the most common forms of attack, the adversarial example with a small and intentional perturbation are intentionally designed to cause the model to make mistakes in its predictions. However, few works investigate the robustness of biometric recognition under adversarial examples, including multi-spectral palmprint recognition.

Most of the studies on multi-spectral palmprint recognition focus on effectively extracting features and fusing data from different views to improve recognition accuracy, such as RootSIFT-based feature fusion [7] and hierarchical approach [8], but they ignore the model robustness against adversarial examples. In recent years, although deep neural networks, such as deep canonical correlation analysis [9] and deep discriminative representation [10], have been applied into multi-spectral palmprint recognition, some studies [11], [12], [13], [14] have shown that the deep neural networks are also vulnerable to adversarial examples [15], [16]. With the improvement of biometric system security requirements, it is urgent to study the robustness of multi-spectral palmprint recognition model.

Adversarial attacks are usually classified as white-box attack and black-box attack. For white-box attacks, the attackers have access to all information about the target model, including the architecture, parameters, gradients, and training procedure. The attacker can make full use of the model information to generate adversarial examples. Some white-box attacks have been proposed to attack the recognition system. For example, fast gradient sign method (FGSM) is designed to quickly find a perturbation for the input to reduce classification confidence [13]. Project Gradient Descent (PGD) [17] is an iterative version of FGSM, and it uses FGSM multiple times within a supremum norm bound for the total input perturbation. Carlini and Wagner attack (CW) [18] aims to

find the perceptually-minimal input perturbation by restricting their l_2-norms. Wang et al. [19] proposed an invisible adversarial attack, which synthesizes adversarial examples that are visually indistinguishable. Despite a high attack success rate, white-box attack relies heavily on the explicit information of the model to be attacked, which is difficult to obtain in real scenarios.

For black-box attacks, the attackers have no access to the target model's parameters or training set. For example, Zeroth order optimization (ZOO) [20] approximates the gradients of the objective function with respect to the input using finite-difference numerical estimates. Boundary attack [21] learns the adversarial example by iteratively perturbing another "seed" image belonging to a different class towards the decision boundary. Zhang et al. capture the structural feature space and then generate perturbations by adversarial region [22]. One-pixel attack (OPA) is also a semi-black-box method that uses differential evolution [23], [24] to find adversarial perturbations [25]. However, it is still a challenging problem to improve the success rate of black-box attack without or with partial information of the model to be attacked.

Although above white-box and black-box attack algorithms have been proposed, they are not appropriate to effectively deal with the increasingly common multi-view fusion recognition problems, especially for multi-spectral palmprint recognition. Because existing attack methods are based on the single view data, they cannot take advantage of the complementary information between different spectra, which is not conducive to learn the most aggressive adversarial examples for multi-spectral palmprint recognition. In addition, in real-world applications, the fusion recognition architecture or its output usually cannot be obtained before attacking, but most of the existing adversarial attack methods depend on above information of the recognition model, which results in limited application scope and attack efficiency.

To solve the problems above, in this paper, we proposed the multi-spectral palmprint joints attack and defense strategy based on multi-view adversarial examples learning. First, we capture the deep common representation space of the multi-spectral palmprint images by generalized canonical correlation analysis (DGCCA) [26]. A perturbation that can significantly suppress the correlation of the features in deep common representation space is introduced by gradient propagation, which is called correlation perturbation. Second, we capture the discriminative manifold space of data and achieve another perturbation along the orthogonal mapping direction in the adversarial region of the input. We call it discriminative perturbations. Then, the adversarial example for each spectrum can be generated by synthesizing the above two types of perturbations, which simultaneously disturbs the fusion and discriminability of the multi-spectral palmprint recognition model. Finally, the generated adversarial examples are fed into the model training to enhance the robustness of the recognition algorithms. Figure 1 shows the main difference between the conventional single-view attack and our proposed multi-view joint attack for multi-spectral palmprint recognition. For a single-view attack, the perturbations are calculated separately for each spectrum and the complementary information between different spectra is ignored. For the proposed multi-view attack, the joint perturbations are learned with the latent correlation structure across different palmprint spectra, which is important to the robustness of the fusion recognition model. In brief, the main contributions of this paper are summarized as follows:

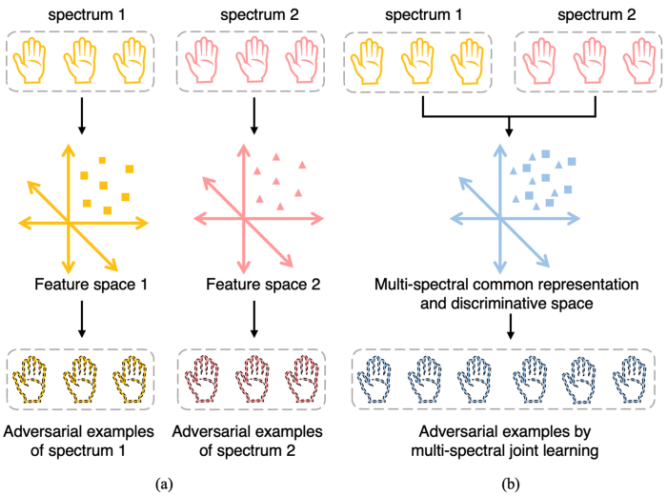

Fig. 1. The comparison of the schematics of the widely used single-view attack shown in (a) and the proposed multi-view attack shown in (b).

(1) To the best of our knowledge, it is the first work to investigate the robustness of the multi-spectral palmprint recognition algorithm, including the multi-view joint attack and defense strategy with adversarial examples learning.

(2) The proposed adversarial example generation method simultaneously disturbs the fusion and discriminability of the multi-spectral palmprint features to achieve effective attack.

(3) Our method has high applicability in attacking different multi-spectral palmprint recognition algorithms, because it is independent of classification models and does not require the time-consuming query step.

(4) The experimental results on multi-spectral palmprint dataset show that the proposed method outperforms the state-of-the-art attack methods. Adopting the adversarial examples by our method in training model can significantly promote the robustness of the multi-spectral palmprint recognition algorithms.

The rest of this paper is organized as follows. We briefly review the related work in Section II. In Section III, we present the proposed multi-view adversarial examples learning method in detail. The comparative experiments on multi-spectral palmprint dataset are presented in Section IV, with the conclusion drawn in Section V.

II. Related Work

A. Deep Canonical Correlation Analysis

Canonical correlation analysis (CCA) is a statistics-based fusion technique used to find the linear projections of two views that are most correlated, shown in Figure 2 (a). Suppose there are two feature matrixes of two views, i.e., $X_1 \in R^{m\times d_1}$ and $X_2 \in R^{m\times d_2}$. CCA pursues the linear projections of the

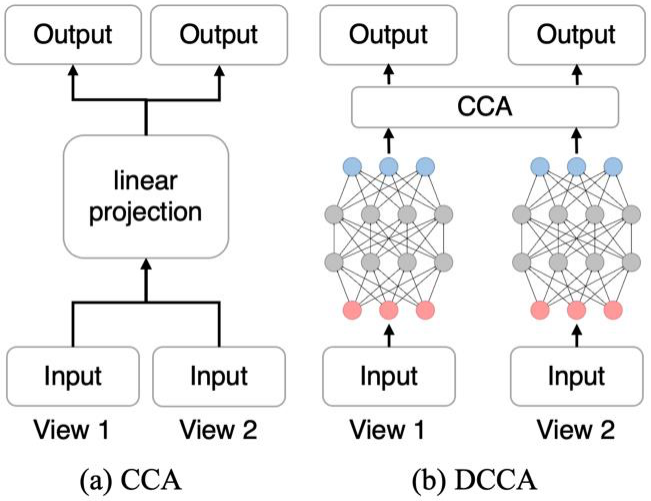

Fig. 2. The schematic of CCA and DCCA.

two views that are maximally correlated:

$$\left(Z_1^*, Z_2^*\right) = \arg\max_{(Z_1,Z_2)} corr(X_1Z_1, X_2Z_2) \tag{1}$$

where Z_1, Z_2 are the projection matrixes for the two views. However, CCA only considers the linear correlations of the views. Deep canonical correlation Analysis (DCCA) is proposed to learn nonlinear projections of two views through neural networks, shown in Figure 2 (b). The output of k-layer neural network with c_i units of x_i can be represented as:

$$h_k^i = s\left(W_k^i h_{k-1}^i + b_k^i\right) \in R^{c_i} \tag{2}$$

where $h_1^i = X_i$, $W_k^i \in R^{c_k \times d_i}$ is a matrix of weights, $b_k^i \in R^{c_i}$ is a vector of biases, and s is a nonlinear activation function. The output for x_i is defined as $f_i(x_i)$. The goal of DCCA is to jointly learn parameters for both views such that the correlation of the representations is as high as possible. We define θ_1 as the vector of all parameters for the first view, and same for θ_2. Then, $\left(\theta_1^*, \theta_2^*\right) = \arg\max_{(\theta_1,\theta_2)} corr(f_1(X_1;\theta_1), f_2(X_2;\theta_2))$. The parameters θ_1 and θ_2 are trained to optimize the total correlation of the two views using gradient-based optimization.

In addition, Deep Generalized Canonical Correlation (DGCCA) is an extension of DCCA, which can learn a nonlinear transformation for two or more views to maximize the correlation among them.

B. Difference Subspace

The difference subspace (DS) [27] aims to capture the intrinsic discriminative features of the samples from different categories. Suppose there are N-dimensional subspaces $\mathcal{P}$ and $\mathcal{Q}$ for each of the two categories. Let d_i be a difference vector between canonical vectors [28], [29] $p_i \in \mathcal{P}$ and $q_i \in \mathcal{Q}$, which form the i_{th} canonical angle α_i shown in Figure 3 (difference subspace). All difference vectors are mutually orthogonal [27]. Then, the length of each differential vector d_i is normalized to 1, and we regard the normalized vectors d_i as the orthonormal basis vectors of the difference subspace $\mathcal{D}$ spanned by $\{d_1, d_2, \ldots, d_N\}$. Difference subspace can also be defined analytically. Assume that $[\phi_1 \ldots \phi_N] \in R^{f\times N}$ and $[\psi_1 \ldots \psi_N] \in R^{f\times N}$ are orthonormal bases for subspaces $\mathcal{P}$ and $\mathcal{Q}$, respectively. The i_{th} basis vector of difference subspace $\mathcal{D}$ is equal to the normalized eigenvector x_i of $P+Q$ corresponding to the i_{th} minimum eigenvalue smaller than 1, where P and $Q \in \mathrm{R}^{f\times f}$ are the orthogonal projection matrices of the two subspaces, defined by $\sum_{i=1}^N \phi_i\phi_i^T$ and $\sum_{i=1}^N \psi_i\psi_i^T$, respectively. The difference subspace $\mathcal{D}$ is spanned by the eigenvectors of matrix $P+Q$ with eigenvalues less than 1.

The concept of the difference subspace can be generalized to the case of differences between two or more subspaces. Assuming that we have $C(C \geq 2)N$-dimensional subspaces. We define k_{th} subspaces as P_k and corresponding projection matrices as $\mathcal{P}_k$. The generalized difference subspace (GDS) [27] is defined as the subspace generated by removing the principal component subspace M of all the class subspaces from the sum subspace $\{\mathcal{P}_k\}_{k=1}^C$. From above definition, the generalized difference subspace is spanned by the N_d smallest eigenvalues of the sum matrix $G = \sum_{k=1}^C P_k$.

III. Proposed Method

In this section, we will introduce the proposed multi-view joint attack method, which generates perturbations by reducing the correlation between samples and disturbing discriminative information in samples. The framework of this method is shown in Figure 3.

A. Adversarial Perturbations for Reducing Correlation

In this paper, we propose a multi-view joint attack method for multispectral palmprint recognition. For the first step, we learn the nonlinear map to maximize the correlation among different views. We try to generate correlation perturbations that can minimize the correlation information between projected feature matrixes, and these perturbations will eventually affect the classification results.

In the condition of two views, DCCA is used to capture the correlated feature space. Now we define two feature matrixes of two views as $X_1 \in R^{m\times d_1}$ and $X_2 \in R^{m\times d_2}$, and their corresponding projected matrixes are expressed as $f_1(X_1;\theta_1)$ and $f_2(X_1;\theta_2)$, where θ_1 and θ_2 are the parameters of deep networks of each view. The goal is to jointly learn parameters for both views' deep networks such that $corr(f_1(X_1;\theta_1), f_2(X_1;\theta_2)$ is as high as possible, and then we have:

$$\left(\theta_1^*, \theta_2^*\right) = \arg\max_{(\theta_1,\theta_2)} corr\left(f_1\left(X_1;\theta_1\right), f_2\left(X_2;\theta_2\right)\right) \tag{3}$$

Let $Y_1 = f_1(X_1;\theta_1) \in R^{m\times d'}$, $Y_2 = f_2(X_2;\theta_2) \in R^{m\times d'}$ be the feature matrices, whose columns are the top-level representations produced by the deep models on the two views, for a training set of size m. We centralize data matrices by $\bar{Y}_1 = Y_1 - \frac{1}{m}Y_1$, $\bar{Y}_2 = Y_2 - \frac{1}{m}Y_2$, and we define $\sum_{12} = \frac{1}{m-1}\bar{Y}_1\bar{Y}_2^T$, $\sum_{12} = \frac{1}{m-1}\bar{Y}_1\bar{Y}_1^T + r_1 I$, and $\sum_{22} = \frac{1}{m-1}\bar{Y}_2\bar{Y}_2^T + r_2 I$ for regularization constants r_1, r_2. Assume that r_1, $r_2 > 0$ so that Σ_{11} and Σ_{22} are positive definite. Total correlation of the top k components of Y_1 and Y_2 is the sum of the top k values

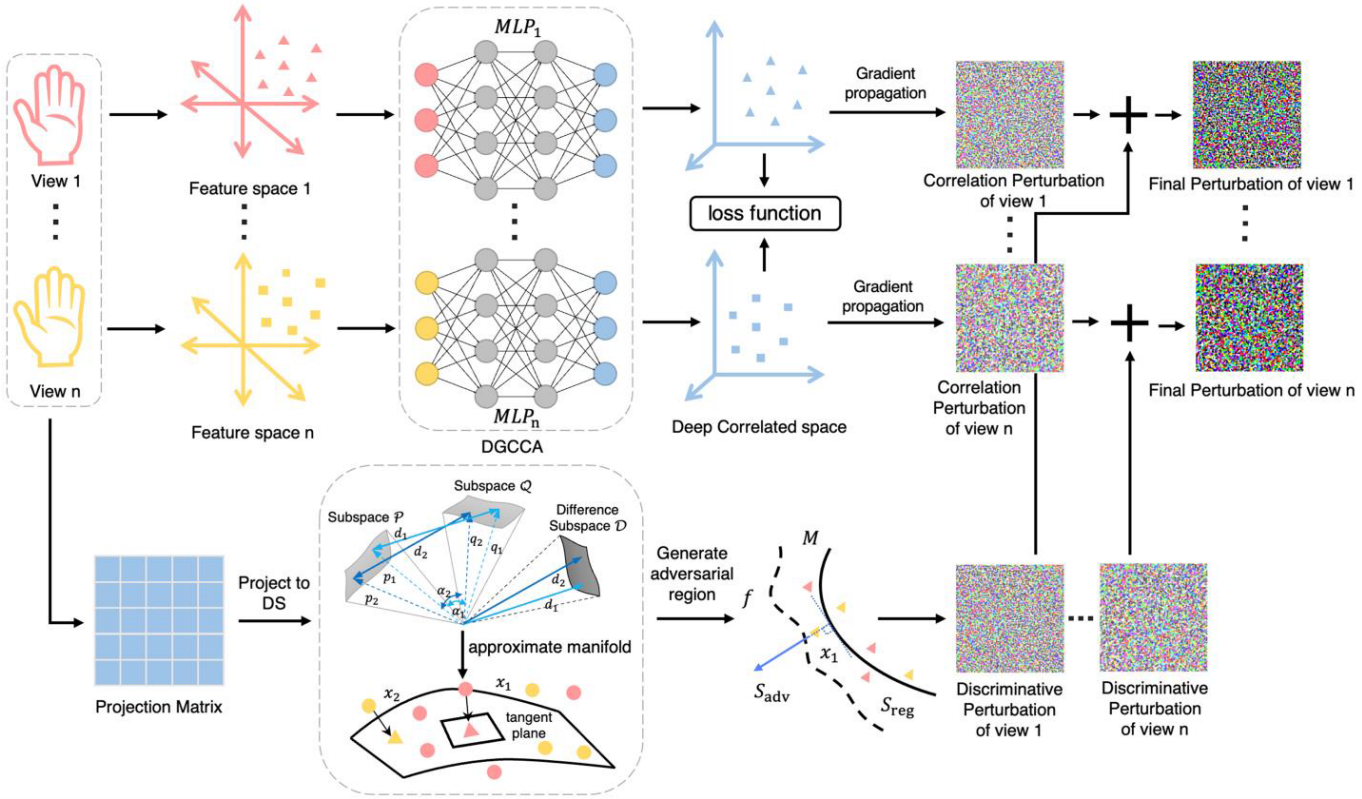

Fig. 3. The framework of our proposed attack method. As the inputs, view 1 to view n represent the palmprints under n different spectra. We use DGCCA to capture multi-view deep common feature of input views. Then we get correlation perturbations for view 1 to view n by calculating the gradient of loss by Eq. (7), which can be used to reduce the correlation among features after projection. Simultaneously, we approximate the discriminative manifold by mapping the original data into the difference subspace and generate discriminative perturbation according to the theory of adversarial region. Finally, these two types of perturbations are synthesized to obtain the most aggressive adversarial examples for multi-spectral palmprint recognition.

of the matrix $U = \Sigma_{11}^{-\frac{1}{2}} \Sigma_{12} \Sigma_{22}^{-\frac{1}{2}}$. When we take $k = d'$, the total correlation is the matrix trace-norm of U: $corr\,(Y_1, Y_2) = \|U\|_{tr} = tr\left(U^T U\right)^{\frac{1}{2}}$. We reduce the total correlation between the projected feature matrices by adding some perturbations to the original image. We could use PGD to attack DCCA models and get perturbations. PGD is a popular steepest descent attack, and it is used to generate adversarial perturbations efficiently. In each iteration, the perturbation is updated in the steepest descent direction. In order to facilitate optimization, $-corr\,(Y_1, Y_2)$ is used as the loss function of DCCA model during training. The gradients of view 1 and view 2 are represented as $\nabla_{X_1}(-corr\,(Y_1, Y_2))$ and $\nabla_{X_2}(-corr\,(Y_1, Y_2))$, respectively. In a non-targeted setting, we have an iterative formulation to generate adversarial example of view i:

$$X_i^{t+1} = Clip_{x_i,\epsilon}\left(X_i^t + \alpha \cdot \text{sign}\left(\nabla_{X_i^t}\left(-corr\left(H_1^t, H_2^t\right)\right)\right)\right) \tag{4}$$

where $X_i^0 = X_i$, $Clip$ denotes the function that projects its argument to the surface of X_i's ε-neighbor ball $B_\varepsilon(x) : \{x' : \|x' - x\|_\infty \leq \varepsilon\}$, t indicates the iteration number and α is the step size of an iteration.

After adding correlation perturbation to the original image, the correlation coefficient between $f_1(X_1', \theta_1)$ and $f_2\left(X_2', \theta_2\right)$ is reduced, which suppresses the feature fusion ability of the model to be attacked.

For $n (n > 2)$ views, we capture the correlated feature space by DGCCA. We define the input matrix with m samples for i_{th} view as $X_i \in R^{m \times d_i}$, and corresponding projected matrix is expressed as $Y_i = f_i(X_i, \theta_i)$. We have the scaled empirical covariance matrix $Y_i Y_i^T \in R^{d_i' \times d_i'}$ and as the corresponding projection matrix $P_i = Y_i^T (Y_i Y_i^T)^{-1} Y_i \in R^{m*m}$, noting that P_i is symmetric and idempotent. Furthermore, we define the sum of projection matrices as $M = \sum_{i=1}^n P_i$. Thus, the objective function of DGCCA can be expressed as:

$$\min_{G, U_i} \sum_{i=1}^n \left\| G - U_i^T f_i\,(X_i, \theta_i) \right\|_F^2 = rn - Tr\left(GMG^T\right) \tag{5}$$

where G is the shared representation of different views, whose rows are equal to the top r eigenvectors of M and $U_i = (Y_i Y_i^T)^{-1} Y_i G^T$. Minimizing the DGCCA objective is equivalent to maximizing $Tr\left(GMG^T\right)$, which is the sum of eigenvalues of M, i.e., $\sum_{i=1}^r \lambda_i\,(M)$, and the loss function of DGCCA can be expressed as follows:

$$Loss = -\sum_{i=1}^r \lambda_i\,(M) \tag{6}$$

The gradient for view i can be expressed as follow:

$$\frac{\partial L}{\partial Y_i} = 2U_i G - 2U_i U_i^T Y_i \tag{7}$$

The next step is similar with the case of two views, we could use PGD method to attack DGCCA models and get correlation

perturbations for each view, which reduces the correlation of the features across different spectra and may lead incorrect classification.

B. Difference Subspace of Multi-Spectral Palmprints

Here, we first genialize the difference subspace [27] from single view data to multi-view data fusion. Then, we capture difference subspace of the samples and disturb the discriminative information in multi-spectral palmprint images by introducing perturbation. Given n views, we define $\mathcal{P}_i$ as the subspace of view i for class 1, and $\mathcal{Q}_i$ as the subspace of view i for class 2. The number of basis of subspace above are N. Assume that $P_i = [\phi_1^i \dots \phi_N^i] \in R^{f\times N}$ expresses orthonormal basis for subspaces $\mathcal{P}_i$, and $Q_i = [\varphi_1^i \dots \varphi_N^i] \in R^{f\times N}$ expresses orthonormal basis for subspaces $\mathcal{Q}_i$. The corresponding sum matrix G for multi-view data can be computed as follows:

$$G = \sum_{i=1}^{n}\sum_{j=1}^{N}\phi_j^i\phi_j^{i^T} + \sum_{i=1}^{n}\sum_{j=1}^{N}\varphi_j^i\varphi_j^{i^T} \tag{8}$$

Theorem 1: The difference subspace of multi-view data is spanned by the N_d normalized eigenvectors of sum matrix G and the corresponding eigenvalues of those N_d eigenvectors are less than 1.

Proof: Given two projection matrices A and B, we have $\lambda_k = (\lambda_k' - 1)^2$, where λ_k' and λ_k are the k_{th} largest eigenvalue of matrix $A + B$ and that of matrix AB, respectively [27]. The eigenvalues λ_k' of multi-view sum matrix G is equal to $1-\sqrt{\lambda_k}$ when $\lambda_k' < 1$, and the eigenvector v_k of the multi-view sum matrix G can be written as follows:

$$Gv_k = \left(1 - \sqrt{\lambda_k}\right) v_k \tag{9}$$

Then we substitute the expression for G into Eq. (9) and simplify it, we obtain:

$$\left(\sum_{i=1}^{n} P_i^T P_i + \sum_{i=1}^{n} Q_i^T Q_i\right) v_k = \left(1 - \sqrt{\lambda_k}\right) v_k \tag{10}$$

We can use a linear combination of basis vectors from multi-view subspaces to represent v_k, i.e., $v_k = \sum_{j=1}^{N}(a_j p_j + b_i q_i)$, where $p_i \in (\mathcal{P}_1, \mathcal{P}_2 \dots \mathcal{P}_n)$ and $q_i \in (\mathcal{Q}_1, \mathcal{Q}_2 \dots \mathcal{Q}_n)$. Eq. (10) can be rewritten as:

$$\left(\sum_{i=1}^{n} P_i^T P_i + \sum_{i=1}^{n} Q_i^T Q_i\right)\sum_{j=1}^{N}(a_j p_j + b_j q_j) = \left(1 - \sqrt{\lambda_k}\right)\sum_{j=1}^{N}(a_j p_j + b_j q_j) \tag{11}$$

From [30], Eq. (11) has relationships as follows: $\sqrt{\lambda}_j = \eta_j$, $(\sum_{i=1}^{n} P_i^T P_i)q_j = \eta_j p_j$, $(\sum_{i=1}^{n} Q_i^T Q_i)p_j = \eta_j q_j$, $(\sum_{i=1}^{n} P_i^T P_i)p_j = p_j$ and $(\sum_{i=1}^{n} Q_i^T Q_i)q_j = q_j$. Thus, Eq. (11) can be represented as:

$$\sum_{j=1}^{N}\left(\left(b_j\eta_j + a_j\eta_k\right)p_j + \left(a_j\eta_j + b_j\eta_k\right)q_j\right) = 0 \tag{12}$$

where we can obtain equation that:

$$b_j\eta_j + a_j\eta_k = a_j\eta_j + b_j\eta_k = 0 \tag{13}$$

since p_j, q_j are linearly independent and from Eq. (13) we get that $\left(\eta_j + \eta_{j+1}\right)\left(a_{j+1} + b_{j+1}\right) = 0$ if $j \neq k$. Here, $a_{j+1} = -b_{j+1}$, because $\eta_j + \eta_{j+1} \neq 0$. Then we substitute the equation into Eq. (13) and find that $a_{j+1}\left(\eta_{j+1} - \eta_j\right) = b_{j+1}\left(\eta_{j+1} - \eta_j\right) = 0$. Since $\eta_{j+1} \neq \eta_j$, we get $a_{j+1} = b_{j+1} = 0 (j \neq k)$. In addition, we can obtain that $b_k\eta_k + a_k\eta_k = 0$ by substituting $j = k$ into Eq. (13). Hence, we get $a_k + b_k = 0$ because $\eta_k \neq 0$. Finally, the eigenvector v_k can be rewritten as $v_k = p_k - q_k$. It proves that the difference subspace of multi-view data is spanned by the normalized eigenvector of corresponding sum matrix.

C. Adversarial Perturbations for Disturbing Discriminative Information

For general multi-spectral palmprint recognition tasks, we usually cannot obtain the explicit information of the target model. To overcome this problem, we develop a target-free adversarial attack method that is independent of the classification models. We can generate adversarial examples according to the definition of the adversarial region.

Adversarial region is based on the adversarial phenomenon [22] in data manifold. A well-trained model is invariant to distortions of the input manifold [31], [32]. Assume that x is the input data and $g(x)$ is the output of the target model, we can create such distortion by making $\frac{\partial g(x)}{\partial x}$ orthogonal to the manifold tangent vectors at point x. We now define a dataset D in a d-dimensional space, and M is the corresponding manifold. Suppose $x_1 \in D$ and x_1^* is the projection point of x_1 on manifold M. The adversarial region of x_1 is defined as

$$S_{AR}(x_1) = \left\{ x \,\middle|\, x = x_1 + \frac{\varepsilon\left(x_1 - x_1^*\right)}{\left\|x_1 - x_1^*\right\|_2}, \varepsilon \in [\varepsilon_{min}, \varepsilon_{max}] \right\} \tag{14}$$

where ε_{min} ensures the adversarial example deviate far enough from the tangent plane to cause misclassification, and ε_{max} ensures that the perturbation on x is hard to perceive. Actually, the adversarial region of D defines a region, and points in this region move along the direction orthogonal to the tangent plane of manifold M.

In addition, adversarial regions impose a potential threat to all machine learning models. We define two different models that have different decision hyperplanes f_1 and f_2. Figure 4 shows an example of $S_{AR}(x_1)$ in 2D space. The region is divided into adversarial region and regular region based on these hyperplanes. If adversarial regions of two models have an intersection, the adversarial examples in the intersection of two adversarial regions can transfer between two models. In other words, this suggests that adversarial examples generated according to adversarial region have strong transferability.

Suppose that $x \in R^D$ is a training example from dataset Q and that x^* is the projection point of x in manifold M. Based

on the definition of adversarial region, the discriminative perturbation can be generated as follows:

$$x_{per} = \varepsilon \frac{x - x^*}{\|x - x^*\|_2} \tag{15}$$

where ε controls the magnitude of the perturbation. Since the manifold M is difficult to construct explicitly, we approximate it by projecting the multi-spectral palmprint data to difference subspace, which can capture the difference feature among different palmprints effectively.

Algorithm 1 Algorithm of the Proposed Multi-View Joint Attack

Input: Dataset D with n views, subspace $\mathcal{P}_k^v$ for class k of view v, sample $x^v \in D$ of view v, component index q, the magnitude of perturbation ε, and an increasing function $f(i)$ me

Output: Adversarial example set D_{adv}

1: Generate the perturbation p_1^v by using Eq.(4)
2: Construct generalized difference subspace
Calculate projection matrix P_k^v of subspace $\mathcal{P}_k^v$ by using Eq.(15)
Calculate sum matrix G by using Eq. (17)
Generate GDS spanned by M eigenvectors of G
3: Project x to the GDS:
$y^v = GDS\_transform(x^v)$
4: Scale y^v with $f(i)$:
If $i > q$: $y_i^{v*} = f(i)\, y_i^v$
Else: $y_i^{v*} = y_i^v$
5: Generate the perturbation p_2^v:
$x^{v*} = GDS\_inverse\ (y^{v*})$, $p_2^v = \varepsilon \frac{x^{v*} - x^v}{\|x^{v*} - x^v\|_2}$
6: Generate adversarial examples:
$x_{adv}^v = x^v + a p_1^v + b p_2^v$
7: **return** D_{adv}

Assuming that we have $C(\geq 2)N$-dimensional subspaces in f-dimensional vector space. The subspace of each class is denoted as $\mathcal{P}_k\,(k = 1, \ldots, C)$, and each subspace $\mathcal{P}_k$ is formed from orthonormal basis like $[\phi_1^k \ldots \phi_N^k] \in R^{f \times N}$. The corresponding projection matrix can be expressed as:

$$P_k = \sum_{i=1}^{N_k} \phi_{ik} \phi_i^{k^T} \tag{16}$$

The difference subspace $\mathcal{D}_c$ can be defined as the subspace produced by removing the principal component subspace of all the classes from the sum subspace. From this definition, D_c is spanned by N_d eigenvectors, $\{\mathrm{d}_i\}_{i=N\times C-N_d+1}^{N\times C}$ corresponding to the N_d smallest eigenvalues, and principal component subspace M_c is spanned by the eigenvectors $\{\mathrm{d}_i\}_{i=1}^{N\times C-N_d}$ of the multi-view sum matrix G of k classes, which is expressed as:

$$G = \sum_{k=1}^{C} P_k = \sum_{k=1}^{C} \sum_{i=1}^{N} \phi_i^k \phi_i^{k^T} \tag{17}$$

Now we obtain a transformation matrix $Z \in R^{N_d \times N_d}$ that consists of N_d eigenvectors, which span the GDS. We transform

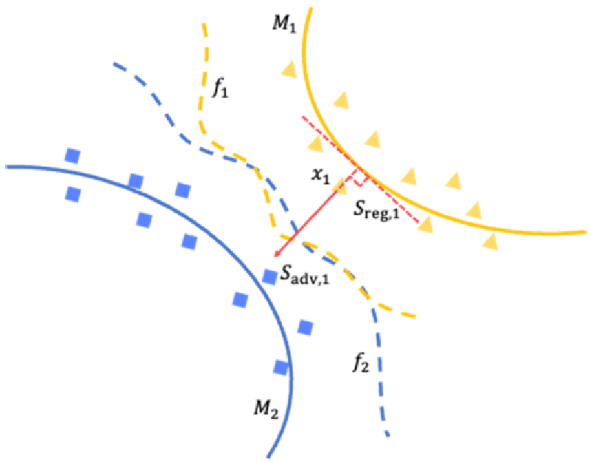

Fig. 4. The schematic of adversarial region. There are manifolds M_1, M_2 and some points of these two manifolds. f_1, f_2 are the decision boundaries by two different classifiers. $S_{reg,1}$ and $S_{adv,1}$ are the regular region and adversarial region obtained by M_1.

x to y, such that $y = Z^T x$ and $y^* = Z^T x^*$, where $y \in R^{N_d}$ and $y^* \in R^{N_d}$. We assume that the N_d eigenvectors, which span GDS, can approximately construct the data manifold and $y_i = y_i^*$ for $i \in \{1, 2, \ldots, N_d\}$. Here, y_i and y_i^* refer to the i_{th} entries of vectors y and y^*, respectively. According to the Eq. (15), discriminative perturbation of x can be generated as follows:

$$\begin{aligned} x_{\text{per}} &= \varepsilon \frac{(x - x^*)}{\|x - x^*\|_2} \\ &= \sum_{i=1}^{k} \left[\left(1 + \frac{\varepsilon}{\|x - x^*\|_2} \right) y_i - \frac{\varepsilon}{\|x - x^*\|_2} y_i^* \right] z_i \\ &\quad + \sum_{i=k+1}^{N_d} \left[\left(1 + \frac{\varepsilon}{\|x - x^*\|_2} \right) y_i - \frac{\varepsilon}{\|x - x^*\|_2} y_i^* \right] z_i \\ &\quad - \sum_{i=1}^{N_d} y_i z_i \\ &= \sum_{i=k+1}^{N_d} \frac{\varepsilon}{\|x - x^*\|_2} y_i z_i \left(1 - \frac{y_i^*}{y_i} \right) \end{aligned} \tag{18}$$

where $\|x - x^*\|_2$ is a constant, and ε controls the magnitude of the perturbation. We use $f(i) = 1 - \frac{y_i^*}{y_i}$ to denote the weight of the small multiplicative distortion of component z_i. The value of $f(i)$ changes according to the index i, and $|f(i)|$ is an increasing function. Finally, we derive the algorithm as follows:

IV. Experiments

In this section, we applied the proposed method into the task of attacking multi-spectral palmprint recognition model, and conducted extensive experiments to demonstrate the effectiveness of our proposed method. We introduced the setting of our experiment and then validate the performance of our proposed method in different bands settings. We also compared our method with some state-of-the-art white box and black box attack methods. In addition,

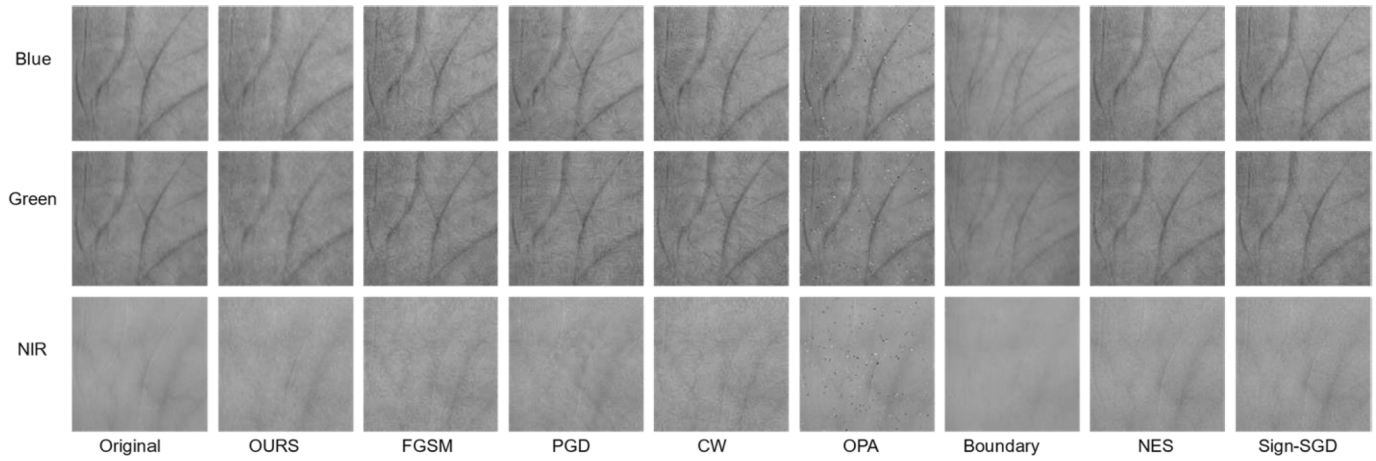

Fig. 5. The comparison of the original palmprint under different spectral bands and their adversarial examples generated by different methods (SSIM = 0.75).

TABLE I
DETAILS OF POLYU MULTI-SPECTRAL PALMPRINT DATASET

Spectrum	Blue	Green	NIR
Individual number	250	250	250
Palm number	500	500	500
Image number per palm	12	12	12
Palmprint image number	6000	6000	6000
Image size(pixels)	128×128	128×128	128×128
Vein structure feature	Weak	Weak	Strong
Palm line feature	Clear	Clear	Unclear
Example			

TABLE II
THE COMPARISON OF ATTACK SUCCESS RATE (%) UNDER DIFFERENT DEFENSES STRATEGIES AND TIME CONSUMPTION OF DIFFERENT ATTACK METHODS (SSIM = 0.95)

	Method	No Defense	JPEG	PD	Time(sec)
white box	FGSM	99.72	89.33	89.40	**2.12**
	PGD	**100.0**	**93.26**	**93.97**	74.23
	CW	99.82	90.54	92.15	18.31
black box	OPA	71.42	28.31	40.55	9976.84
	NES	63.77	49.78	58.55	1952.24
	Sign-SGD	71.12	54.92	60.92	1638.67
	Boundary	83.57	62.44	55.00	1876.24
	Ours	**86.15**	**84.83**	**84.46**	**28.61**

we fed the generated adversarial examples into the model training and test the robustness of the proposed method. The experiment is implemented using Python on a PC with Windows 10, I5-12400 CPU, RTX-3080 GPU and 32-GB random access memory. The code of our proposed method is available at: https://github.com/zhouyyyuze/Multi-Spectral-Palmprint-Attack.

A. Dataset and Experimental Settings

In the experiment, a multi-spectral palmprint dataset from the Hong Kong Polytechnic University (PolyU) [33] was selected. The palmprint images in this dataset are taken from different spectral bands: Blue, Green and NIR. Details of PolyU Multi-spectral Palmprint Dataset are summarized in Table I.

In the experiment, we use attack success rate (%) as the metric to evaluate the effectiveness of attacks, which is the proportion of successful attacks among the total number of test images, and it is defined as $\frac{1}{N}\sum_{n=1}^{N}[f(x) \neq f(x')]$. In terms of visual quality assessment of attacks, we chose the structural similarity (SSIM) [34] as a measurement of image similarity. Given a clean image x and its corresponding adversarial example x_{adv}, the SSIM (x, x_{adv}) measures the similarity between x and x_{adv}. The higher the similarity between x and x_{adv} is, the SSIM (x, x_{adv}) is larger. In addition, we use transfer rate [35] as the metric to evaluate the transferability of attacks. The transfer rate is the proportion of adversarial examples of a model that are likely to attack the other model. We set parameters by grid search in the experiment.

B. Comparison With Other Attack Methods

To better prove the effectiveness of our proposed attack on multi-spectral palmprint recognition model, we will compare our proposed attack with other white-box attacks including FGSM, PGD, CW, and black-box attacks including OPA, natural evolution strategies attack (NES) [36], sign-based stochastic gradient descent attack (Sign-SGD) [37] and boundary attack. We apply these methods into attacking the multi-spectral palmprint recognition model, and the success rate, computational time, and transfer rate are reported. The target palmprint recognition model fuses palmprints under three spectral bands of blue, green and nir. We randomly used 4800 images for each view to train the multi-view adversarial examples learning algorithm, and the remaining images are treated as test set. We repeat the experiments ten times and report the average results. The defense methods include JPEG

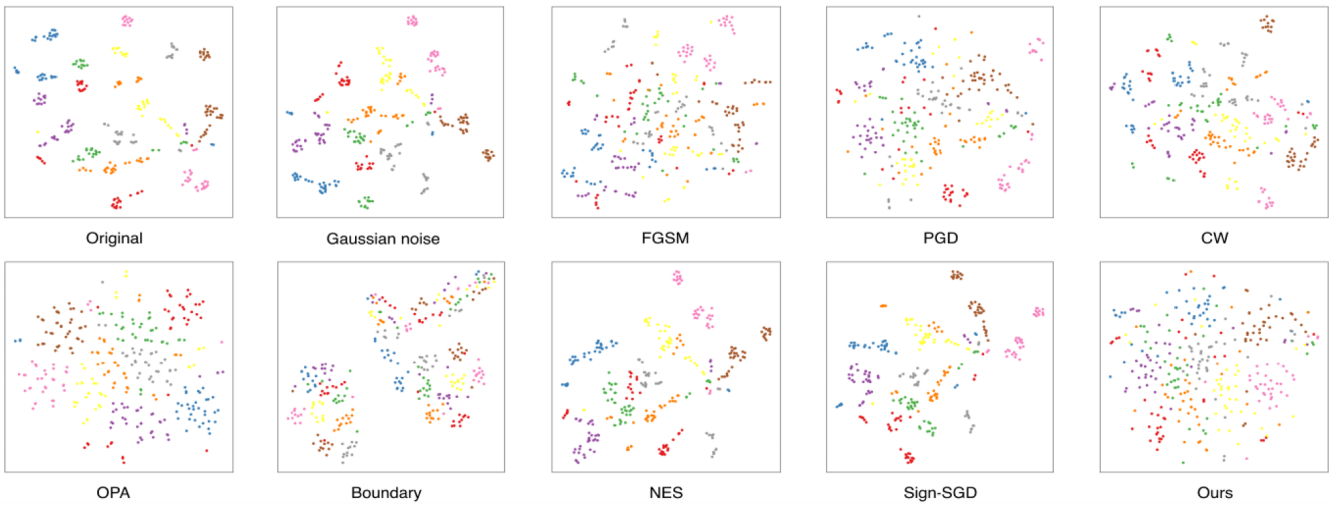

Fig. 6. The T-SNE visualization of the original data and adversarial examples generated by different methods (SSIM = 0.90).

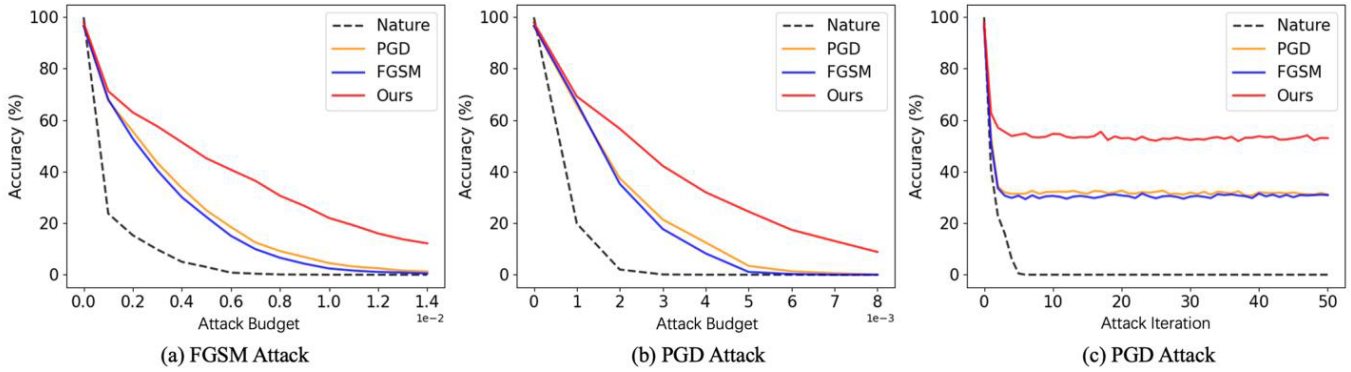

Fig. 7. The accuracy (%) for natural and robust models with adversarial examples generated by different attacks.

compression [38] and pixel deflection [39]. We summarize the results in Table II.

As Table II shows, among all the black attack methods, our proposed method has the highest attack success rate. Under JPEG compression and pixel deflection defenses, the adversarial examples generated by our method are more robust to current defense methods compared with other attacks. The boundary attack achieved a high success rate without defense, but its success rate decreased significantly under defense methods. Although the white-box attacks including FGSM, PGD, CW achieved high success rates of more than 95%, it is worth noting that we do not need access to the architecture or gradients of the target model, which means that our method has much wider applicability in real-world attack task. In addition, those black-box attack methods are too time-consuming, because they require a large number of queries for generating adversarial examples. Our proposed method can generate adversarial examples efficiently, because it does not need any queries to the target model, which makes our method have strong model applicability.

Figure 5 shows the adversarial examples generated using FGSM, PGD, CW, OPA, Boundary, NES, Sign-SGD and Ours when SSIM = 0.75. Compared with other methods, the perturbations added to the clean images by our method are more imperceptible. In other words, under the same intensity of interference noise, the adversarial example generated by our method is the hardest to be detected by the user of the recognition model to be attacked.

To better show the influence of adversarial examples on features, we visualize the features of the generated adversarial examples through T-SNE [40], which projects features onto a low dimensional space to visualize the distribution of samples. As shown in Figure 6, each color represents a category. The more chaotic the categories are, the worse the separability is, which also means that the attack method is more effective. Compared with other method, the adversarial examples generated by our method are more chaotic, which indicates that adversarial examples generated by our method are more likely to result in misclassification.

In addition, we also focus on the comparison of transferability of those methods. We first generate adversarial examples of DGCCA model by these methods, and then test the transferability on target models with SVM, ResNet34, VGG16 and Competitive Code (CompCode) [41]. In this experiment, we measure the transferability using the criteria of transfer rate used in [35], and summarize the results in Table III.

From Table III, we find the adversarial examples generated by our method have stronger transferability than the

TABLE III
THE TRANSFER RATE (%) OF THE ADVERSARIAL EXAMPLES GENERATED BY DIFFERENT METHODS

SSIM	Method	SVM	ResNet34	VGG16	CompCode
0.95	FGSM (white)	71.42	39.25	49.82	31.54
	PGD (white)	87.55	41.92	50.17	34.09
	CW (white)	74.28	40.75	57.04	43.45
	OPA (black)	22.67	81.75	74.34	78.24
	NES (black)	38.41	67.11	53.82	29.84
	Sign-SGD (black)	30.73	36.65	38.78	20.49
	Boundary(black)	81.48	60.92	83.21	69.81
	Ours (black)	**95.12**	**98.64**	**97.20**	**81.62**
0.90	FGSM (white)	79.81	75.39	88.52	71.08
	PGD (white)	91.32	53.64	87.48	76.61
	CW (white)	84.05	62.67	85.39	65.14
	OPA (black)	25.59	90.46	82.69	**91.92**
	NES (black)	51.22	74.72	58.95	34.52
	Sign-SGD (black)	40.92	47.61	49.56	25.59
	Boundary(black)	92.20	86.02	93.47	88.52
	Ours (black)	**97.13**	**99.52**	**98.73**	88.71
0.85	FGSM (white)	98.93	91.05	94.17	79.16
	PGD (white)	**99.71**	83.76	94.43	88.94
	CW (white)	98.49	88.20	94.00	83.84
	OPA (black)	30.91	95.06	89.21	92.77
	NES (black)	67.73	82.84	67.65	42.60
	Sign-SGD (black)	56.24	65.18	62.08	34.09
	Boundary (black)	94.25	91.12	96.95	94.47
	Ours (black)	99.29	**99.63**	**99.35**	**94.92**

comparison methods, including FGSM, PGD, CW, OPA, NES, Sign-SGD and Boundary, on different target models with different values of SSIM.

C. The Performance of Our Method Under Different Spectral Band Combinations

We now evaluate the performance of our proposed multi-spectral palmprint joint attack based on multi-view adversarial examples learning with untargeted setting. We compared the performance of our method under two or three spectral bands. The target model includes SVM, Resnet34, VGG16 and CompCode, and we evaluate our attack with different magnitude of the perturbation. We summarize the results in Table IV.

As Table IV shown, when we relax the constraint of SSIM, the success rates of our attack increase. When SSIM $= 0.9$, there is a high success rate in those target models. In addition, we find that compared with model fusing two spectral bands, the model fusing three views is more susceptible to adversarial examples. Although the recognition accuracy of the model can be improved by integrating more palmprints under different spectral bands, the model is more vulnerable to adversarial examples. In addition, ours proposed method achieves higher success rates when target model fuses more palmprint images under different spectra. This is because our proposed method can make good use of information between different spectra.

D. Adversarial Defense Experiment

The adversarial examples derived from our method can not only be used to attack the palmprint recognition model, but integrating them into training is expected to improve the robustness of the model. In this section, we will verify the effectiveness of the proposed method in improving model robustness in adversarial training (AT) [42], [43], [44], [45]. The training process is as follows: we treat training dataset as the original images, and then our proposed attack is used to generate corresponding adversarial examples for each image. Then, an augmented dataset is formed by combining the original and adversarial examples. The enhanced adversarial training model is trained by the augmented dataset. We compared the performance of adversarial training by FGSM and PGD, and evaluated the model robustness against FGSM attack and PGD attack under different attack budgets, with the results shown in Figure 7 (a) and (b). It is observed that the performance of nature model drops quickly as the attack budget increases. The FGSM and PGD model improve the model robustness significantly across a wide range of attack budgets. The model trained by our proposed method

TABLE IV
THE ATTACK SUCCESS RATE (%) OF OUR ATTACK FOR DIFFERENT RECOGNITION MODELS WITH DIFFERENT SPECTRA

SSIM	Palmprint	DGCCA	ResNet34	VGG16	CompCode
0.95	Blue, Green	83.53	85.65	80.14	73.72
	Blue, NIR	85.23	98.03	82.36	71.81
	Green, NIR	84.83	81.76	88.23	**84.15**
	Blue, Green, NIR	**86.15**	**98.64**	**97.20**	76.27
0.90	Blue, Green	94.12	95.44	88.17	87.53
	Blue, NIR	96.35	98.84	90.53	88.82
	Green, NIR	96.12	96.23	92.72	**89.02**
	Blue, Green, NIR	**96.68**	**99.52**	**98.73**	88.71
0.85	Blue, Green	95.71	97.12	92.43	93.64
	Blue, NIR	97.52	**99.82**	94.72	92.44
	Green, NIR	98.52	99.12	95.53	93.26
	Blue, Green, NIR	**98.62**	99.63	**99.35**	**94.92**

further boosts the performance of the FGSM and PGD model by a large margin under different attack budgets. We also conduct experiments using PGD attack with different attack iterations under a fixed attack budget of 0.5, and the results are shown in Figure 7 (c). It is observed that our model consistently outperforms others across a wide range of attack iterations. In general, these results demonstrate our method can significantly promote the robustness of the multi-spectral palmprint recognition algorithms.

V. CONCLUSION

In this paper, we investigate the robustness of multi-spectral palmprint recognition algorithms and proposed the multi-spectral palmprints joint attack and defense strategy based on multi-view adversarial examples learning. Different from the previous single view attack, our proposed method can exploit the information among different views to simultaneously reduce the correlation among features after projection and disturb the discriminative information in palmprints, which can effectively suppress the representation and classification ability of the model. In addition, compared with the conventional black-box attacks, our proposed method can construct adversarial examples with higher efficiency, which makes it has strong applicability for real-world multi-view fusion application problems. The extensive experiments on multi-spectral palmprint datasets demonstrate that the proposed adversarial example attack approach is superior to the state-of-the-art adversarial example generation methods and has strong transferability. The experiment results show that fusing more multi-spectral palmprint images improves recognition accuracy, but it makes the model more vulnerable to adversarial examples. Moreover, the defense strategy with the adversarial examples by our method can significantly promote the robustness of multi-spectral palmprint recognition algorithms.

REFERENCES

[1] S. Li and B. Zhang, "Joint discriminative sparse coding for robust hand-based multimodal recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 3186–3198, 2021.

[2] A. Jain and A. Kumar, "Biometric recognition: An overview," in *Second Generation Biometrics: The Ethical, Legal and Social Context*. Dordrecht, The Netherlands: Springer, 2012, pp. 49–79.

[3] D. Zhong, X. Du, and K. Zhong, "Decade progress of palmprint recognition: A brief survey," *Neurocomputing*, vol. 328, pp. 16–28, Feb. 2019.

[4] Y. Aberni, L. Boubchir, and B. Daachi, "Multispectral palmprint recognition: A survey and comparative study," *J. Circuits, Syst. Comput.*, vol. 28, no. 7, Jun. 2019, Art. no. 1950107.

[5] M. D. Bounneche, L. Boubchir, A. Bouridane, B. Nekhoul, and A. A. Chérif, "Multi-spectral palmprint recognition based on oriented multiscale log-Gabor filters," *Neurocomputing*, vol. 205, pp. 274–286, Sep. 2016.

[6] X. Sun and S. Sun, "Adversarial robustness and attacks for multi-view deep models," *Eng. Appl. Artif. Intell.*, vol. 97, Jan. 2021, Art. no. 104085.

[7] X. Yan, W. Kang, F. Deng, and Q. Wu, "Palm vein recognition based on multi-sampling and feature-level fusion," *Neurocomputing*, vol. 151, pp. 798–807, Mar. 2015.

[8] D. Hong, W. Liu, J. Su, Z. Pan, and G. Wang, "A novel hierarchical approach for multispectral palmprint recognition," *Neurocomputing*, vol. 151, pp. 511–521, Mar. 2015.

[9] G. Andrew et al., "Deep canonical correlation analysis," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1247–1255.

[10] S. Zhao and B. Zhang, "Deep discriminative representation for generic palmprint recognition," *Pattern Recognit.*, vol. 98, Feb. 2020, Art. no. 107071.

[11] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[12] C. Szegedy et al., "Intriguing properties of neural networks," 2013, *arXiv:1312.6199*.

[13] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.

[14] N. Papernot, P. McDaniel, and I. Goodfellow, "Transferability in machine learning: From phenomena to black-box attacks using adversarial samples," 2016, *arXiv:1605.07277*.

[15] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.

[16] H. Xu et al., "Adversarial attacks and defenses in images, graphs and text: A review," *Int. J. Autom. Comput.*, vol. 17, no. 2, pp. 151–178, 2020.

[17] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2017, *arXiv:1706.06083*.

[18] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Secur. Privacy (SP)*, May 2017, pp. 39–57.

[19] Z. Wang et al., "Invisible adversarial attack against deep neural networks: An adaptive penalization approach," *IEEE Trans. Dependable Secure Comput.*, vol. 18, no. 3, pp. 1474–1488, Jul. 2019.

[20] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "ZOO: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proc. 10th ACM Workshop Artif. Intell. Secur.*, Nov. 2017, pp. 15–26.

[21] W. Brendel, J. Rauber, and M. Bethge, "Decision-based adversarial attacks: Reliable attacks against black-box machine learning models," 2017, *arXiv:1712.04248*.

[22] Y. Zhang et al., "Principal component adversarial example," *IEEE Trans. Image Process.*, vol. 29, pp. 4804–4815, 2020.

[23] S. Das and P. N. Suganthan, "Differential evolution: A survey of the state-of-the-art," *IEEE Trans. Evol. Comput.*, vol. 15, no. 1, pp. 4–31, Feb. 2011.

[24] R. Storn and K. Price, "Differential evolution—A simple and efficient heuristic for global optimization over continuous spaces," *J. Global Optim.*, vol. 11, no. 4, pp. 341–359, 1997.

[25] J. Su, D. Vargas, and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Trans. Evol. Comput.*, vol. 23, no. 5, pp. 828–841, Oct. 2019.

[26] A. Benton, H. Khayrallah, B. Gujral, D. A. Reisinger, S. Zhang, and R. Arora, "Deep generalized canonical correlation analysis," 2017, *arXiv:1702.02519*.

[27] K. Fukui and A. Maki, "Difference subspace and its generalization for subspace-based methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 11, pp. 2164–2177, Nov. 2015.

[28] H. Hotelling, "Relations between two sets of variates," in *Breakthroughs in Statistics*. New York, NY, USA: Springer, 1992, pp. 162–190.

[29] S. N. Afriat, "Orthogonal and oblique projectors and the characteristics of pairs of vector spaces," in *Mathematical Proceedings of the Cambridge Philosophical Society*. Cambridge, U.K.: Cambridge Univ. Press, 1957.

[30] H. Yanai, "Some generalized forms a least squares g-inverse, minimum norm g-inverse, and Moore–Penrose inverse matrices," *Comput. Statist. Data Anal.*, vol. 10, no. 3, pp. 251–260, Dec. 1990.

[31] S. Rifai et al., "The manifold tangent classifier," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 24, 2011, pp. 1–9.

[32] P. Simard et al., "Tangent prop—A formalism for specifying selected invariances in an adaptive network," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 4, 1991, pp. 895–903.

[33] 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, Feb. 2010.

[34] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[35] A. Kurakin, I. Goodfellow, and S. Bengio, "Adversarial machine learning at scale," 2016, *arXiv:1611.01236*.

[36] A. Ilyas et al., "Black-box adversarial attacks with limited queries and information," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 2137–2146.

[37] S. Liu et al., "SignSGD via zeroth-order oracle," in *Proc. Int. Conf. Learn. Represent.*, 2018, pp. 1–24.

[38] R. Shin and D. Song, "JPEG-resistant adversarial images," in *Proc. NIPS Workshop Mach. Learn. Comput. Secur.*, 2017, pp. 1–6.

[39] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer, "Deflecting adversarial attacks with pixel deflection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 8571–8580.

[40] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, no. 11, pp. 2579–2605, 2008.

[41] A. W.-K. Kong and D. Zhang, "Competitive coding scheme for palmprint verification," in *Proc. 17th Int. Conf. Pattern Recognit.*, 2004, pp. 520–523.

[42] A. Shafahi et al., "Adversarial training for free!" in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–12.

[43] F. Tramer and D. Boneh, "Adversarial training and robustness for multiple perturbations," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019, pp. 1–11.

[44] D. Wang, C. Li, S. Wen, S. Nepal, and Y. Xiang, "Defending against adversarial attack towards deep neural networks via collaborative multi-task training," *IEEE Trans. Dependable Secure Comput.*, vol. 19, no. 2, pp. 953–965, Mar. 2022.

[45] A. Shafahi et al., "Universal adversarial training," in *Proc. Conf. Artif. Intell.*, 2020, pp. 5636–5643.