Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2023.045854

ARTICLE

Local Adaptive Gradient Variance Attack for Deep Fake Fingerprint Detection

Chengsheng Yuan[1,2], Baojie Cui[1,2], Zhili Zhou[3], Xinting Li[4,*] and Qingming Jonathan Wu[5]

[1]Engineering Research Center of Digital Forensics, Ministry of Education, Nanjing University of Information Science and Technology, Nanjing, 210044, China

[2]School of Computer Science, Nanjing University of Information Science and Technology, Nanjing, 210044, China

[3]Institute of Artificial Intelligence and Blockchain, Guangzhou University, Guangzhou, 510006, China

[4]School of International Relations, National University of Defense Technology, Nanjing, 210039, China

[5]Department of Electrical and Computer Engineering, University of Windsor, Windsor, N9B 3P4, Canada

*Corresponding Author: Xinting Li. Email: lixt@tju.edu.cn

Received: 09 September 2023 Accepted: 20 November 2023 Published: 30 January 2024

ABSTRACT

In recent years, deep learning has been the mainstream technology for fingerprint liveness detection (FLD) tasks because of its remarkable performance. However, recent studies have shown that these deep fake fingerprint detection (DFFD) models are not resistant to attacks by adversarial examples, which are generated by the introduction of subtle perturbations in the fingerprint image, allowing the model to make fake judgments. Most of the existing adversarial example generation methods are based on gradient optimization, which is easy to fall into local optimal, resulting in poor transferability of adversarial attacks. In addition, the perturbation added to the blank area of the fingerprint image is easily perceived by the human eye, leading to poor visual quality. In response to the above challenges, this paper proposes a novel adversarial attack method based on local adaptive gradient variance for DFFD. The ridge texture area within the fingerprint image has been identified and designated as the region for perturbation generation. Subsequently, the images are fed into the targeted white-box model, and the gradient direction is optimized to compute gradient variance. Additionally, an adaptive parameter search method is proposed using stochastic gradient ascent to explore the parameter values during adversarial example generation, aiming to maximize adversarial attack performance. Experimental results on two publicly available fingerprint datasets show that our method achieves higher attack transferability and robustness than existing methods, and the perturbation is harder to perceive.

KEYWORDS

FLD; adversarial attacks; adversarial examples; gradient optimization; transferability

1 Introduction

Recently, biometric identification technologies such as fingerprint recognition [1], face recognition [2], and iris recognition [3], etc., have seen extensive deployment in a wide range of real-world applications. Fingerprint recognition, in particular, is favored due to its versatility, uniqueness and convenience. However recent studies have shown that these systems are vulnerable to fraudulent

This work is licensed under a Creative Commons Attribution 4.0 International License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

attacks involving fake fingerprints, the proposed fingerprint liveness detection technology can solve the above problems well, and its main task is to identify whether the fingerprint to be authenticated is from a human or a forged imitation.

In recent years, the iterative updating and development of deep learning technology has provided a whole new set of solutions for multiple types of tasks in the field of computer vision, such as image classification [4], object recognition [5], semantic segmentation [6], natural language processing [7]. These solutions have already been introduced to real-world scenarios, such as face recognition [8] and pedestrian detection [9], etc., and have obtained good feedback. Given the excellent performance of deep learning in image classification, it has also been applied to fingerprint liveness detection tasks. Notably, research on fingerprint liveness detection based on deep learning has garnered substantial attention from both academia and industry, showcasing remarkable achievements in deep fake fingerprint detection tasks [10], [11]. However, the latest research [12] has pointed out that in addition to the problem of spoofing attacks by forged fingerprints, DFFD models also face the problem of adversarial attacks. That is, by adding some fine perturbations to the fingerprint image, the constructed adversarial example enables the model to make false classifications with a high degree of confidence. The adversarial fingerprint is more destructive compared with the spoofing attack, and it is easier for the DFFD model to make wrong predictions. This poses a major threat to the integrity and security of the DFFD system. According to the level of knowledge, adversarial examples are generally divided into two categories: white-box attack entails crafting adversarial examples using knowledge of the target model's internal structure and parameters. In contrast, a black-box attack is a method for generating adversarial examples without access to the internal architecture or parameters of the target model. Generally speaking, white-box attacks can design customized perturbations according to the structure and parameters of the model, and achieve a high success rate of attack, but poor transferability in the face of unknown models. In a real-world scenario, it is not realistic to discover and know the parameters and structure of the target model in advance. It is more based on the black box attack scenario, that is, the parameters and structure of the target model are not known in advance, so studying and improving the transferability between different models is more in line with the real scenario. In addition, adversarial attacks can be divided into targeted and untargeted attacks depending on whether the model is incorrectly classified into a particular category. There has been limited research in the realm of adversarial attacks on DFFD systems, despite the significant threats they pose. To deal with adversarial attacks and improve the security of fingerprint recognition systems, a novel adversarial example generation method based on local adaptive gradient variance is proposed in this paper. The main contributions of this paper are as follows:

- To enhance the visual quality of the adversarial fingerprint without affecting the original attack performance, Grad-CAM is used to visualize the attention area of the fingerprint image and as the additional area of the subsequent perturbation, which is difficult to perceive by human eyes.
- To improve the transferability in the gradient-based adversarial example generation method, this paper proposes a local adaptive gradient variance attack method, which realizes the gradient update direction controllable by computing gradient variance at each iteration.
- In this paper, an adaptive parameter search method is proposed to search the optimal hyper-parameters by using the stochastic gradient ascent method, thus reducing manual intervention and balancing the success rate of white-box attacks.
- In this paper, the performance evaluation is tested on two publicly available fingerprint datasets, LiveDet2019 [13] and LiveDet2017 [14]. Experimental results indicate that the proposed method can improve the transferability of adversarial examples between different DFFD models, and obtain good visual quality.

This paper is an extension of our previous conference paper [15]. Compared with the [15], this paper extends and improves it. The main differences are summarized as follows: (1) We have improved the introduction section, restated the research motivation, provided a more comprehensive introduction and solved the actual problems. (2) In Section 3, we constrain the region where perturbation is added, and discuss and analyze the feasibility. (3) In Section 4, we demonstrate that the proposed method can be combined with input transformations to improve the visual quality of adversarial examples via extensive experimentation and interpretation.

The rest of this paper is structured as follows: Section 2 presents a review of related work. In Section 3, we introduce the proposed method for generating adversarial examples. Section 4 provides the experimental results. Finally, the conclusion and future work are given.

2 Related Work

2.1 Adversarial Examples

Szegedy et al. [16] first disclosed the flaws in image classification tasks: although deep learning has achieved impressive performance in image classification, it faces a serious challenge, that is, adding some subtle perturbations to the original image can cause the model to make incorrect predictions, and the human visual system can hardly catch the anomaly. They also give a mathematical formula for the calculation of perturbation, expressed in ρ, which induces the model to give a wrong judgment:

$$minimize \quad |\rho|_p \quad s.t.\ y' = f\left(x+\rho\right), \tag{1}$$

$$x+\rho \in [0,1], y' \neq y$$

In formulation (1), ρ is constrained by norm L_p, where x represents the original input image, y is the ground-truth label associated with it, and y' signifies the incorrect label.

To calculate the global optimal solution, Szegedy tried to transformed the adversarial attack into a convex optimization problem, and presented a L-BFGS method [17]. After that, more and more work has been proposed. Moosavi-Dezfooli et al. [18] designed an iterative method to calculate the minimum perturbation for input images and added perturbations to guide the output image toward the decision boundary of the classifier. Carlini et al. [19] presented a series of three attacks along with a novel loss function designed to deceive target networks via defensive distillation.

Concurrently, the research landscape has seen the emergence of various black-box attack techniques. Sarkar et al. [20] introduced the UPSET network, capable of generating adversarial examples with universal perturbations applied to original images, effectively causing the model to misclassify specific target classes. Bhagoji et al. [21] proposed a finite difference-based method (FD attack) rooted in finite difference principles, wherein pixel data is adjusted to estimate the gradient direction approximately, subsequently conducting iterative attacks based on this estimated gradient. Dong et al. [22] introduced momentum into the iterative adversarial example generation process. Su et al. [23] designed a single pixel attack using a differential evolution algorithm to explore extreme conditions by modifying a single pixel in an image to trick the classifier. Furthermore, Fei et al. [12] conducted pioneering research into the feasibility of adversarial examples within the context of DFFD networks. A series of improved and optimized adversarial example generation methods have been proposed, showing great potential in this field. In this paper, focusing on DFFD model, we focus on how to improve the mobility and visual quality of the detection model.

2.2 Gradient-Based Methods

In this section, our primary focus is on presenting gradient-based attacks aimed at enhancing the transferability of adversarial attacks.

2.2.1 Fast Gradient Sign Method (FGSM)

To solve the nonlinear and vulnerability problems of the model, Goodfellow et al. [24] first proposed a FGSM, which realizes the generation of the adversarial example x^{adv} by optimizing the loss function and one-step update method, which is defined as follows:

$$x^{adv} = x + \epsilon \cdot sign\left(\nabla_x J\left(x, y;\ \theta\right)\right) \tag{2}$$

here, x^{adv} represents the adversarial example, x is the original image, ϵ is the perturbation size, $sign(.)$ denotes the sign function, and $\nabla_x J$ represents the gradient derived from the loss function $J(.)$. While FGSM performs attack rapidly, it exhibits a moderate success rate in adversarial attacks.

2.2.2 Iterative Fast Gradient Sign Method (I-FGSM)

In contrast to FGSM, which relies on a single iteration, Kurakin et al. proposed I-FGSM [25] conducts multiple iterations during adversarial example generation, employing a smaller step size α for each iteration, as expressed by:

$$x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign\left(\nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right)\right) \tag{3}$$

where x_t^{adv} signifies the adversarial example at the t-th iteration, with $t = 0$ implying $x_t^{adv} = x$, I-FGSM perform a better attack performance compared to FGSM but incurs higher computational costs.

2.2.3 Momentum Iterative Fast Gradient Sign Method (MI-FGSM)

To enhance the update stability and avoid local maximum, MI-FGSM [22] is proposed to extend I-FGSM, that is, the momentum of previous iterations was included in gradient calculation to boost the transferability of adversarial examples, which is expressed as:

$$\begin{aligned} g_{t+1} &= \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right)}{||\nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right)||_1}, \\ x_{t+1}^{adv} &= x_t^{adv} + \alpha \cdot sign\left(g_{t+1}\right), \end{aligned} \tag{4}$$

where g_t is the gradient at the t-th iteration, and the attenuation factor μ accumulates gradients from previous iterations.

2.2.4 Nesterov Iterative Fast Gradient Sign Method (NI-FGSM)

NI-FGSM [26] introduces a Nesterov momentum, during the gradient update, predicts the gradient direction of the next iteration. NI-FGSM substitutes x_t^{adv} in Eq. (4) with $x_t^{adv} + \alpha \cdot \mu \cdot g_t$.

2.2.5 Variance Tuning Momentum-Based Iterative Method (VMI-FGSM)

VMI-FGSM [27] calculates the gradient of the neighborhood data points during the update to optimize the gradient update direction in the next iteration. It can be expressed as:

$$g_{t+1} = \mu \cdot g_t + \frac{\nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right) + v_t}{||\nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right) + v_t||_1} \tag{5}$$

where $v_{t+1} = \frac{1}{n}\sum_{i=1}^{N} \nabla_{x^i} J\left(x^i, y;\ \theta\right) - \nabla_x J\left(x, y;\ \theta\right)$, $x^i = x_t^{adv} + r^i$, and r^i is randomly selected in the neighborhood. Essentially, v_t captures the difference between the current gradient and the average gradient of its neighborhood during the t-th iteration. This method further enhances the transferability of the gradient-based adversarial attacks.

2.2.6 Transformation Robust Attack (TRA)

TRA [12] stands as the pioneering adversarial attack method in the realm of DFFD, and it confirms the feasibility of adversarial attacks on DFFD.

2.3 Input Transformations

This section introduces various input transformations to enhance the attack transferability.

2.3.1 Diverse Input Method (DIM)

DIM [28] implements stochastic alterations involving resizing and padding on input data using a fixed probability. Subsequently, the modified images are directed through the classifier to calculate gradients, thereby enhancing the potential for transferability.

2.3.2 Translation-Invariant Method (TIM)

TIM [29] employs a set of images to compute gradients, proving particularly effective, especially when confronting black-box models equipped with defensive mechanisms. To mitigate gradient calculations, Dong et al. introduce slight positional shifts to the images, followed by an approximation of gradient computation through convolving gradients from unaltered images with a kernel matrix.

2.3.3 Scale-Invariant Method (SIM)

SIM [26] introduces the concept of scale-invariant property and computes gradients across an array of images scaled by a factor of $1/2^i$ relative to the input image, enhancing the adaptability of generated adversarial examples, where i is treated as a hyper-parameter.

It’s important to emphasize that different input transformation methods, namely DIM, TIM, and SIM, can be seamlessly incorporated into gradient-based attack methodologies, that is, the Composite Transformation Method (CTM) can improve mobility more effectively. In this study, the proposed method seeks to enhance the transferability of gradient-based attacks (e.g., MI-FGSM, VMI-FGSM). It can be synergistically employed alongside diverse input transformations to further bolster the transferability of the attack.

3 Proposed Method

VMI-FGSM establishes gradient variance as the distinction between the average gradient in the vicinity neighborhood and the gradient from the previous iteration. We believe that simply combining the difference between the previous iteration and the current iteration gradient is not enough to solve the transferability problem of adversarial examples. Consequently, this paper designs a novel attack approach based on local adaptive gradient variance under lower perturbation levels, adversarial examples generated using our method demonstrate enhanced attack performance and increased transferability against unknown DFFD models. Fig. 1 provides visual insights into the effects of various attacks on both live and counterfeit fingerprints.

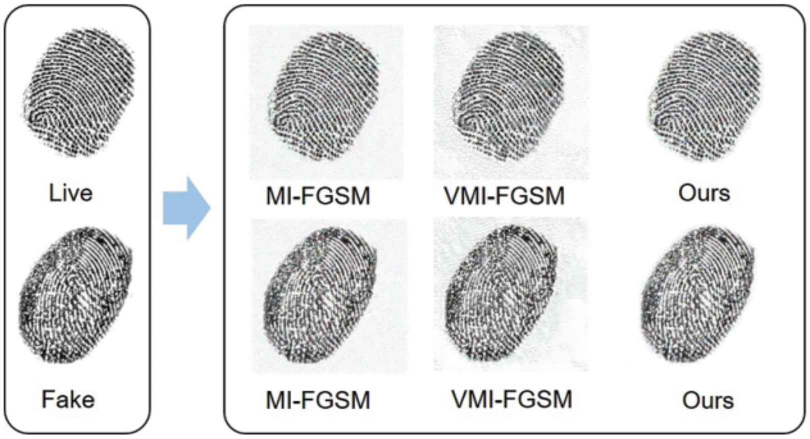

Figure 1: When the perturbation size of ϵ is 0.06 and the attack target model is Inception-v3, compared with MI-FGSM and VMI-FGSM, the adversarial examples of the proposed method is better

3.1 Adversarial Fingerprint Area Location

The generation of adversarial examples involves adding subtle perturbations to the original image, typically constrained by specific norms like L_0 or L_∞. The requirement of these interference generation is that the human visual system cannot be observed without successfully attacking the task model. To delve into the underlying principles of adversarial fingerprints, we employ the Grad-CAM to highlight sensitive regions that influence DFFD classification.

As illustrated in Fig. 2, the layers of the VGG-16 model concentrate on discerning the texture of the original fingerprint within the image. When dealing with an adversarial fingerprint generated using the I-FGSM method, we observe that the introduced perturbations not only avoid diverting neural network attention to irrelevant areas but also focus on the fingerprint texture region. Consequently, our research uses YOLO-v5 [30] to restrict the addition of perturbations for adversarial fingerprints exclusively to the interior of the fingerprint texture region. This strategy not only improves the image quality without affecting the performance of the original task, but also makes the perturbations imperceptible to humans.

Algorithm 1: Local Adaptive Gradient Variance Attack

```
Input: A classifier $f$ with parameter $\theta$, loss function $J$, a raw image $x$, and the corresponding label $y$. The magnitude of perturbation $\epsilon$, decay factor $\mu$, total iteration number $T$, number of sampled examples $N$, the variance factor $\lambda$ and the neighborhood bound $b$.
Output: An adversarial example $x^{adv}$
$\alpha = \epsilon / T$
for $t = 0$ to $T - 1$ do
    Get the gradient $\tilde{g}_{t+1}$ by
        $\tilde{g}_{t+1} = \nabla_{x_t^{adv}} J\left(x_t^{adv}, y;\ \theta\right)$
    Update $\tilde{g}_{t+1}$ by
        $g_{t+1} = \mu \cdot g_t + \frac{\tilde{g}_{t+1} + \lambda \cdot v_t}{||\tilde{g}_{t+1} + \lambda \cdot v_t||_1}$
    Update $v_{t+1}$ by
```

(Continued)

Algorithm 1 (continued)

```
        $v_{t+1} = \left(\frac{1}{N}\sum_{i=1}^{N}\nabla_{x^i}J\left(x^i, y;\ \theta\right) - \nabla_x J\left(x, y;\ \theta\right)\right)^2$
    Update $x_{t+1}^{adv}$ by gradient
        $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign\left(g_{t+1}\right)$
end for
$x^{adv} = x_T^{adv}$
return $x^{adv}$
```

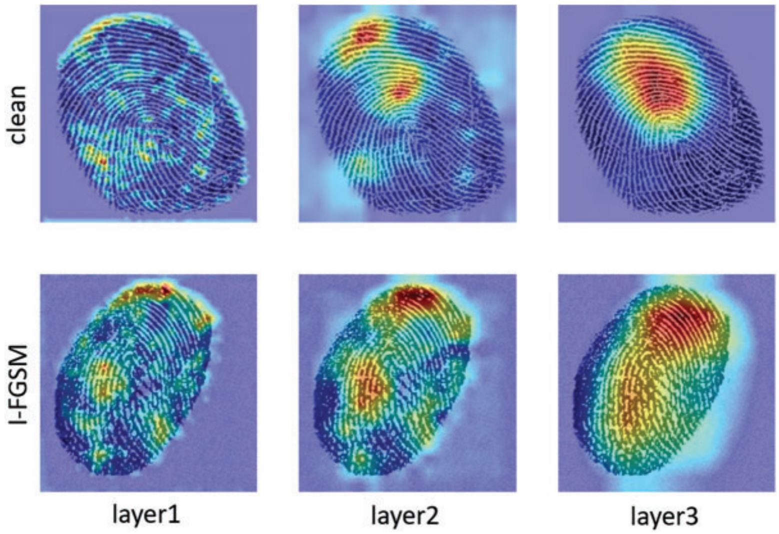

Figure 2: The Grad-CAM visualization of VGG-16 classifies a clean fingerprint image and an FGSM generated adversarial fingerprint, layer1 to layer3 represents the layer from shallow to deep

3.2 Gradient Variance

Given a clean image x, a corresponding label y, a classifier f parameterized by θ, and a loss function J. Additionally, we consider an example x' sampled from the neighborhood, along with a neighborhood bound ϵ' for that region, and define the gradient variance as follows:

$$v\left(x\right) = \left(E_{\left\|x'-x\right\|_{p<\epsilon'}}\left[\nabla_{x'}J\left(x', y;\ \theta\right) - \nabla_x J\left(x, y;\ \theta\right)\right]\right)^2. \tag{6}$$

We establish $\epsilon' = \epsilon \cdot b$, where ϵ denotes the perturbation size, and b functions as the bound of the neighborhood. Due to the discontinuous nature of the input, calculating $E_{\|x-x\|_{p<\epsilon'}}\left[\nabla_{x'}J\left(x', y;\ \theta\right)\right]$ directly is not feasible. Consequently, we approximate this value by sampling N data points from the neighborhood of x to compute $v(x)$:

$$v_{t+1} = \left(\tfrac{1}{N}\sum_{i=1}^{N}\nabla_{x^i}J\left(x^i, y;\ \theta\right) - \nabla_x J\left(x, y;\ \theta\right)\right)^2 \tag{7}$$

Here, x^i represents an example sampled from distribution $U[-(b \cdot \epsilon)^d, (b \cdot \epsilon)^d]$. During the t-th iteration, we employ the gradient variance obtained from the preceding $(t-1)$-th iteration to modify the gradient's update direction. In addition, we add a factor λ to control impact of gradient variance on transferability.

```
Algorithm 2: Parameter Adaptive Searching
Input: A classifier $f$ with parameter $\theta$, dataset of fingerprints $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, perturbation size $\epsilon$, decay factor $\mu$, parameters $\theta^{adv}$, number of iteration $T$, the loss function $J'(\theta^{adv})$ and learning rate $lr$.
Output: Parameters $\tilde{\theta}^{adv}$
$lr = 0.1, g_0 = 0$
for $i \leq n$ do
    Get adversarial example $x_i^{adv}$ by Algo. 1
    Get the gradient $g_i$ by
        $g_i = \nabla_{x_i^{adv}} J'\left(x_i^{adv}, y_i; \theta^{adv}\right)$
    Update $\theta_i$ by
        $\theta_{i+1}^{adv} = \theta_i^{adv} + lr \cdot g_i$
    Update $x_{t+1}^{adv}$ by gradient
        $x_{t+1}^{adv} = x_t^{adv} + \alpha \cdot sign\,(g_{t+1})$
end for
$\tilde{\theta}^{adv} = \theta_n^{adv}$
return $\tilde{\theta}^{adv}$
```

3.3 Adaptive Parameter Selection

Optimizing the parameters in the above methods manually can be resource-intensive and susceptible to subjective biases. Hence, we have devised an adaptive parameter optimization method rooted in gradient descent. When provided with an adversarial example denoted as x^{adv}, the model's prediction $y = f(x^{adv})$, and loss function $L = J(x^{adv}, y)$. In this approach, we replace N, b, and λ from Algorithm 1 with θ^{adv}. With an original image, a learning rate lr, and the corresponding adversarial example $x^{adv} = g(x; \theta^{adv})$, the loss function can be expressed as $L = J'(x, y;\ \theta^{adv})$. The gradient $\nabla_{\theta^{adv}} J'$ can be derived from $J'(.)$, facilitating the parameter update process:

$$\theta_{t+1}^{adv} = \theta_t^{adv} + lr \cdot \nabla_{\theta^{adv}} J_i\left(\theta^{adv}, x_i^{adv}, y_i^{adv}\right) \tag{8}$$

Notably, our primary aim here is to increase the loss, effectively constituting a gradient ascent procedure. Given the need to generate corresponding adversarial examples in each iteration of the gradient update, we opt for a stochastic gradient ascent approach to manage computational costs. The parameter search process is concisely outlined in Algorithm 2.

4 Experiments

In this section, a set of experiments is carried out using the LiveDet2019 and LiveDet2017 datasets to assess the effectiveness of the method. Initially, we delineate the specific experimental configurations and subsequently compare our success rate in performing attacks with other methods across various settings. It is important to emphasize that FLD constitutes a binary classification task. In the case of non-targeted attack on N categorized datasets, the model is tasked with classifying images into the remaining N-1 classes. In contrast, within the context of the binary classification task, images are categorized exclusively as either genuine or counterfeit. This inherent difference results in an appreciably lower success rate for binary classification attacks compared to their multi-class counterparts. Empirical findings also demonstrate that our method can enhance the transferability of adversarial attacks while improving visual effects compared to alternative methods.

4.1 Experimental Setup

4.1.1 Dataset

In terms of data distribution and image quality, the dataset originates from the Liveness Detection Competitions of 2019 and 2017 (LiveDet2019 and LiveDet2017). These datasets consist of fingerprint images gathered by different sensors such as Digital Persona, Orcathus Sensors, and Green Bit. Each sensor's image collection comprises both genuine fingerprint images and counterfeit fingerprints crafted from a variety of materials such as Ecoflex, Latex and Gelatine.

Since the different quality of fingerprint images collected by different sensors, we have made the deliberate choice to exclusively utilize images gathered by the Digital Persona sensor for the dataset. Additionally, to maintain a balance in the quantities of counterfeit and live fingerprints, our dataset includes only counterfeit fingerprints derived from the initial three materials: Ecoflex, Gelatine, and Latex. This selection aims to ensure a more even distribution of features extracted during the training of our network model, thereby enhancing the model's classification accuracy. Every image in our dataset has undergone resizing to align with the input size required by our model, transforming the original 252 × 324 dimensions to 224 × 224. This meticulously curated dataset serves as the training data for our network model. In the context of testing adversarial attacks, 1000 images that can be reliably classified are selected to test the attack performance.

4.1.2 Metrics

In this paper, the performance of adversarial examples is evaluated from two aspects: (1) Transferability, as the ability to deceive the black-box model is essential for effective adversarial attacks. (2) Image quality, that is, the size of the introduced perturbation, we use the Peak Signal-to-Noise Ratio (PSNR) value as the evaluation metric.

4.1.3 Baselines

We have chosen gradient-based iterative adversarial attacks commonly employed in the field as our baseline, specifically, I-FGSM, MI-FGSM, and VMI-FGSM. It is worth noting that VMI-FGSM has been empirically demonstrated to exhibit superior transferability compared to other attack methods.

4.1.4 Models Choice

In the performance test, five classic networks, specifically, Inception-v3, Inception-v1, Inception-resnet-v2, VGG-16 and Mobilenet-v1, are selected to analyze the performance of different schemes. In addition, we combined two adversarial training models, namely Inception-v3$_{\text{adv}}$, Inception-resnet-v2$_{\text{adv}}$, to evaluate the robustness of adversarial attacks.

4.1.5 Hyper-Parameters Setting

In the experimental setup, we still use the relevant parameters in our previous work [15], where the magnitude of the perturbation is set to $\epsilon = 0.16$. The pixel values range within [0, 1], iteration number $T = 10$, and factor $\mu = 1.0$. For our method, we establish neighborhood bound $b = 1.5$ and factor $\lambda = 1.5$. Notably, we have observed that the attack success rate tends to stabilize when the number of samples reaches 10, prompting us to set $N = 10$ for our experiments. Furthermore, we introduce variations in ϵ, within different attack methods to conduct a comparative analysis. Specifically, we adjust ϵ to take on values of 0.03, 0.06, 0.09, 0.12, and 0.16, respectively, enabling us to assess the performance of our method alongside other techniques under varying perturbation levels.

4.2 Experimental Results

4.2.1 Transferability

First, the attack success rates of different network models I-FGSM, MI-FGSM, VMI-FGSM and ours are evaluated using LiveDet2019 and LiveDet2017 datasets at a fixed perturbation size, and the outcomes are summarized in Tables 1 and 2. The rows represent the attacked model, and adversarial examples are generated based on models in columns. Each neural network is trained using the designated dataset. Notably, the results in the table indicate that our proposed method consistently achieves a superior attack success rate when pitted against unknown models. Furthermore, it sustains its attack performance when confronted with the white-box model.

Table 1: Attack success rate on different models on LiveDet2019

Model	Attack	Inc-v3	Inc-v1	IncRes-v2	VGG-16	Mobilenet
Inc-v3	I-FGSM	92.9%	22.3%	7.4%	3.4%	43.9%
	MI-FGSM	97.2%	39.4%	19.1%	6.9%	45.3%
	VMI-FGSM	97.4%	40.8%	27.7%	8.7%	52.5%
	Ours	96.1%	47.3%	29.2%	11.7%	46.8%
Inc-v1	I-FGSM	15.2%	99.3%	13.7%	8.4%	42.1%
	MI-FGSM	37.5%	99.1%	35.1%	19.6%	50.4%
	VMI-FGSM	53.7%	99.2%	59.7%	30.7%	70.4%
	Ours	58.4%	99.2%	60.9%	36.5%	71.7%
IncRes-v2	I-FGSM	20.8%	26.9%	97.8%	9.2%	41.9%
	MI-FGSM	53.2%	60.3%	97.6%	23.8%	43.8%
	VMI-FGSM	80.3%	87.2%	97.4%	39.7%	45.2%
	Ours	82.1%	91.5%	97.6%	48.1%	49.9%
VGG-16	I-FGSM	11.1%	14.7%	8.3%	95.4%	35.4%
	MI-FGSM	25.5%	33.7%	29.6%	95.1%	45.3%
	VMI-FGSM	39.8%	48.1%	46.8%	95.6%	48.2%
	Ours	45.4%	52.0%	45.9%	96.1%	48.8%
Mobilenet	I-FGSM	9.9%	4.5%	2.0%	3.1%	97.4%
	MI-FGSM	22.4%	12.4%	5.7%	8.2%	97.6%
	VMI-FGSM	30.1%	10.7%	3.3%	10.3%	97.2%
	Ours	45.1%	25.3%	19.3%	17.5%	97.8%

Taking Table 1 as an example, in the context of a white-box attack, the attack success rate of each method based on Inception-resnet-v2 exceeds 97%. For black-box attacks targeting Mobilenet and Inception-v1, our proposed method attains success rates of 91.5% and 49.9%, respectively. In comparison, VMI yields success rates of 87.2% and 45.2% for these models. It is also observed that when the architecture of the black-box model closely resembles that of the target model, the attack exhibits higher success rates. This underscores the robust transferability of our approach across a variety of models. When using the LiveDet2017 dataset, Table 2 presents similar results.

Table 2: Attack success rate on different models on LiveDet2017

Model	Attack	Inc-v3	Inc-v1	IncRes-v2	VGG-16	Mobilenet
Inc-v3	I-FGSM	95.8%	22.6%	12.8%	6.1%	43.8%
	MI-FGSM	97.5%	38.4%	20.4%	9.6%	45.2%
	VMI-FGSM	97.1%	45.0%	25.6%	13.5%	54.9%
	Ours	96.2%	48.5%	27.5%	14.2%	56.0%
Inc-v1	I-FGSM	17.2%	97.3%	15.6%	10.2%	45.1%
	MI-FGSM	35.4%	98.1%	33.1%	22.7%	53.9%
	VMI-FGSM	47.7%	98.2%	55.8%	33.9%	68.4%
	Ours	52.9%	98.2%	62.3%	35.2%	70.3%
IncRes-v2	I-FGSM	23.5%	27.5%	97.4%	8.8%	42.3%
	MI-FGSM	55.4%	61.5%	97.6%	20.1%	44.2%
	VMI-FGSM	75.6%	86.3%	97.4%	43.4%	46.8%
	Ours	80.3%	91.0%	96.9%	46.1%	48.8%
VGG-16	I-FGSM	9.2%	12.0%	7.2%	94.4%	36.5%
	MI-FGSM	20.6%	36.2%	23.1%	96.2%	45.2%
	VMI-FGSM	33.8%	44.7%	42.7%	96.9%	45.9%
	Ours	38.7%	48.9%	42.3%	96.7%	47.3%
Mobilenet	I-FGSM	8.1%	5.1%	3.5%	2.9%	97.2%
	MI-FGSM	22.5%	13.6%	6.1%	7.9%	97.6%
	VMI-FGSM	36.7%	15.8%	10.4%	11.2%	97.6%
	Ours	41.8%	28.0%	15.4%	15.1%	97.7%

Additionally, we test the robustness of different methods by challenging three network models that have undergone adversarial training, as depicted in Table 3. The experimental findings demonstrate that our method exhibits greater robustness when faced with adversarially trained models.

To further investigate the transferability of adversarial examples under varying levels of perturbation, we conducted experiments as presented in Table 4. Rows represent the attack methods employed, while columns denote the perturbation sizes set at 0.03, 0.06, 0.09, 0.12, and 0.16, respectively, with pixel values confined to [0, 1]. All attacks are conducted on Inception-v3, and the reported results represent the average black-box attack success rates, consistent with those detailed in Tables 1 and 2. In Fig. 3, we provide visualizations of a counterfeit fingerprint image and corresponding adversarial examples with different perturbation magnitudes. Due to the limitation of the perturbation region, the noise in the image is not obvious. As the perturbation increases, the black-box attack success rate will be higher. For instance, at $\epsilon = 0.16$, the proposed method achieves a remarkable 35.1% success rate in attacks across various models, surpassing the performance of any other method.

Table 3: Attack success rate on adversarially trained models

Model	Attack	Inc-v3$_{adv}$	Inc-Res $_{adv}$
Inc-v3	I-FGSM	11.2%	2.4%
	MI-FGSM	30.1%	9.5%
	VMI-FGSM	43.8%	17.6%
	Ours	46.9%	18.7%
Inc-v1	I-FGSM	1.9%	2.0%
	MI-FGSM	7.3%	8.9%
	VMI-FGSM	12.5%	13.1%
	Ours	15.6%	17.1%
IncRes-v2	I-FGSM	3.6%	15.4%
	MI-FGSM	10.2%	33.8%
	VMI-FGSM	18.3%	45.1%
	Ours	22.3%	47.5%

Table 4: Attack success rate on Inception-v3 with different perturbation sizes ϵ

Attack	0.03	0.06	0.09	0.12	0.16
I-FGSM	5.4%	7.1%	9.7%	13.2%	19.3%
MI-FGSM	6.2%	10.1%	13.5%	16.9%	27.7%
VMI-FGSM	8.9%	13.8%	17.6%	22.7%	32.4%
Ours	9.5%	14.5%	19.1%	24.5%	35.1%

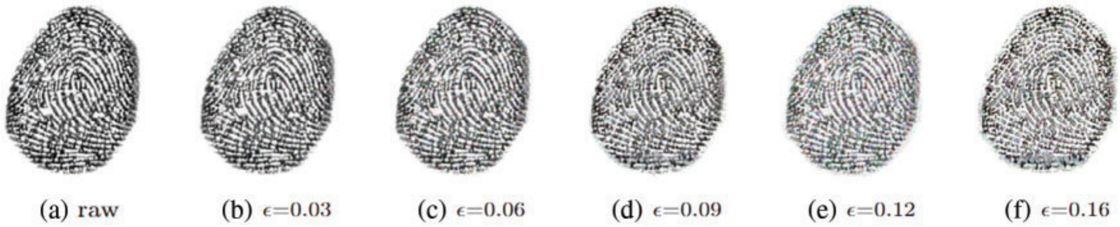

Figure 3: Adversarial examples generated by Ours with different ϵ on Inception-v3

It is worth noting that although it may be difficult for human observers to detect alterations in the image with the increase of perturbation, the model is more vulnerable to deception because its feature perception is different from that of human beings.

4.2.2 Attack with Input Transformations

Input transformations, such as DIM, TIM, and SIM, can be seamlessly integrated with gradient-based adversarial attacks to significantly bolster transferability. In this paper, these input transformations have been incorporated into our method, resulting in a demonstrable enhancement of transferability. As detailed in Table 5, success rates exhibit further improvement across various models, with our proposed method consistently outperforming the baseline. These findings provide additional compelling evidence for the efficacy of our approach.

Table 5: Attack success rate of adversarial examples generated on different models enhanced by CTM

Model	Attack	Inc-v3	Inc-v1	IncRes-v2	VGG-16	Mobilenet
Inc-v3	MI-CT-FGSM	96.8%	67.8%	53.2%	32.0%	61.4%
	VMI-CT-FGSM	97.4%	71.4%	55.5%	39.3%	72.5%
	Ours-CT	95.8%	72.9%	58.4%	39.8%	67.8%
Inc-v1	MI-CT-FGSM	67.2%	98.7%	73.0%	51.8%	57.4%
	VMI-CT-FGSM	53.7%	98.5%	68.7%	71.4%	75.8%
	Ours-CT	57.1%	98.5%	69.5%	72.6%	76.2%
IncRes-v2	MI-CT-FGSM	83.7%	85.3%	97.7%	76.6%	58.0%
	VMI-CT-FGSM	87.5%	87.2%	98.1%	75.7%	61.8%
	Ours-CT	88.1%	92.3%	98.3%	77.0%	63.9%
VGG-16	MI-CT-FGSM	48.4%	53.1%	49.9%	96.2%	56.8%
	VMI-CT-FGSM	52.7%	59.2%	52.3%	96.1%	61.9%
	Ours-CT	55.4%	60.0%	53.2%	96.2%	59.8%
Mobilenet	MI-CT-FGSM	39.1%	23.5%	19.6%	23.3%	98.4%
	VMI-CT-FGSM	43.2%	28.8%	13.6%	24.3%	98.6%
	Ours-CT	58.4%	44.5%	37.6%	31.0%	98.8%

4.2.3 Quantitative Analysis of Visual Quality

Moreover, the PSNR metric has also been introduced to evaluate the quality of adversarial fingerprint images. As presented in Table 6, when compared to alternative attack methods, the adversarial fingerprints generated by our method exhibit a notably higher PSNR value, signifying superior visual quality. Again, the scheme proposed in this paper is effective.

Table 6: PSNR values of adversarial examples generated by different methods

Datasets	I-FGSM	MI-FGSM	VMI-FGSM	Ours
LiveDet2019	18.7	18.9	18.6	22.3
LiveDet2017	18.8	18.9	18.8	22.2

5 Conclusion and Future Work

FLD based on deep learning not only suffers from spoofing attacks of forged fingerprints, but also faces the deceptive attacks problem of adversarial fingerprints. The existing FLD research tasks lack the study of adversarial examples, and the transferability of adversarial attacks in the face of unknown network models is generally poor. To solve the above problems, we propose an adversarial attack method based on local adaptive gradient variance, which is designed to enhance the transferability of adversarial attacks and improve the visual quality, to further enhance the security of the fingerprint recognition system. Initially, we constrain the perturbation generation range and formulate gradient variance as the squared difference between the current gradient and the average gradient of the neighborhood during each iteration. Subsequently, during the generation of adversarial examples at each iteration, we optimize the current gradient direction based on the gradient variance from the previous iteration. To address the challenge of selecting appropriate parameters, this paper proposes an adaptive parameter search method that employs gradient ascent to identify the optimal solution.

Experimental results reveal that our proposed method can effectively enhance the transferability of adversarial attacks and further improve the visual quality while maintaining a high success rate for white-box attacks. These findings underscore the current vulnerabilities of DFFD systems, which struggle to withstand adversarial attacks. While black-box attacks have demonstrated feasibility, there remains room for improving their success rates, albeit at the cost of elevated computational complexity in generating adversarial examples. These challenges merit further exploration in future research, with an emphasis on developing more robust defenses against such attacks.

Acknowledgement: We are grateful to Nanjing University of Information Science and Technology for providing a research environment and computing equipment.

Funding Statement: This work is supported by the National Natural Science Foundation of China under Grant (62102189; 62122032; 61972205), the National Social Sciences Foundation of China under Grant 2022-SKJJ-C-082, the Natural Science Foundation of Jiangsu Province under Grant BK20200807, NUDT Scientific Research Program under Grant (JS21-4; ZK21-43), Guangdong Natural Science Funds for Distinguished Young Scholar under Grant 2023B1515020041.

Author Contributions: Study conception and design: C. Yuan, B. Cui; data collection: B. Cui; analysis and interpretation of results: C. Yuan, B. Cui; draft manuscript preparation: C. Yuan, B. Cui, Z. Zhou, X. Li and Q. M. J. Wu. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: All datasets and materials are publicly available.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] K. Karampidis, M. Rousouliotis, E. Linardos and E. Kavallieratou, “A comprehensive survey of fingerprint presentation attack detection,” *Journal of Surveillance, Security and Safety*, vol. 2, no. 4, pp. 117–161, 2021.

[2] M. Wang and W. Deng, “Deep face recognition: A survey,” *Neurocomputing*, vol. 429, pp. 215–244, 2021.

[3] J. Daugman, “How iris recognition works,” in *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 14, no. 1, pp. 21–30, 2004.

[4] L. Niu, A. Veeraraghavan and A. Sabharwal, “Webly supervised learning meets zero-shot learning: A hybrid approach for fine-grained classification,” in *Proc. of CVPR,* Salt Lake City, USA, pp. 7171–7180, 2018.

[5] G. Hinton, L. Deng, D. Yu, G. E. Dahl, A. R. Mohamed *et al.,* “Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups,” *IEEE Signal Processing Magazine*, vol. 29, no. 6, pp. 82–97, 2012.

[6] J. Long, E. Shelhamer and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2015.

[7] I. Sutskever, O. Vinyals and Q. V. Le, "Sequence to sequence learning with neural networks," in *Advances in Neural Information Processing Systems*, vol. 27. pp. 3104–3112, 2014.

[8] H. Wang, Y. Wang, Z. Zhou, X. Ji, D. Gong *et al.,* "CosFace: Large margin cosine loss for deep face recognition," in *Proc. of CVPR*, Salt Lake City, USA, pp. 5265–5274, 2018.

[9] Z. Zheng, X. Yang, Z. Yu, L. Zheng, Y. Yang *et al.,* "Joint discriminative and generative learning for person re-identification," in *Proc. of CVPR*, Long Beach, USA, pp. 2138–2147, 2019.

[10] C. Yuan, X. Chen, P. Yu, R. Meng, W. Cheng *et al.,* "Semi-supervised stacked autoencoder-based deep hierarchical semantic feature for real-time fingerprint liveness detection," *Journal of Real-Time Image Processing*, vol. 17, no. 1, pp. 55–71, 2020.

[11] W. E. N. Jian, Y. Zhou and H. Liu, "Densely connected convolutional network optimized by genetic algorithm for fingerprint liveness detection," *IEEE Access*, vol. 9, pp. 2229–2243, 2020.

[12] J. Fei, Z. Xia, P. Yu and F. Xiao, "Adversarial attacks on fingerprint liveness detection," *EURASIP Journal on Image and Video Processing*, vol. 2020, no. 1, pp. 3104–3112, 2020.

[13] G. Orrù, R. Casula, P. Tuveri, C. Bazzoni, G. Dessalvi *et al.,* "Livdet in action-fingerprint liveness detection competition 2019," in *Proc. of ICB*, Crete, Greece, pp. 1–6, 2019.

[14] V. Mura, G. Orrù, R. Casula, A. Sibiriu, G. Loi *et al.,* "LivDet 2017 fingerprint liveness detection competition 2017," in *Proc. of ICB*, Gold Coast, Australia, pp. 297–302, 2018.

[15] C. Yuan and B. Cui, "Adversarial attack with adaptive gradient variance for deep fake fingerprint detection," in *Proc. of MMSP*, Shanghai, China, pp. 1–6, 2022.

[16] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan *et al.,* "Intriguing properties of neural networks," arXiv preprint arXiv: 1312.6199, 2014.

[17] H. Tang and X. Qin, *Practical Methods of Optimization*, 1st ed., Dalian, China: Academic Press, Dalian University of Technology Press, pp. 138–149, 2004.

[18] S. M. Moosavi-Dezfooli, A. Fawzi and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. of CVPR*, Las Vegas, USA, pp. 2574–2582, 2016.

[19] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. of Symp. on Security and Privacy*, San Jose, USA, pp. 39–57, 2017.

[20] S. Sarkar, A. Bansal, U. Mahbub and R. Chellappa, "UPSET and ANGRI: Breaking high performance image classifiers," arXiv preprint arXiv:1707.01159, 2017.

[21] A. N. Bhagoji, W. He, B. Li and D. Song, "Practical black-box attacks on deep neural networks using efficient query mechanisms," in *Proc. of ECCV*, Munich, Germany, pp. 3104–3112, 2018.

[22] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu *et al.,* "Boosting adversarial attacks with momentum," in *Proc. of CVPR*, Salt Lake City, USA, pp. 9185–9193, 2018.

[23] J. Su, D. V. Vargas and K. Sakurai, "One pixel attack for fooling deep neural networks," *IEEE Transactions on Evolutionary Computation*, vol. 23, no. 5, pp. 828–841, 2019.

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

[25] A. Kurakin, I. J. Goodfellow and S. Bengio, "Adversarial examples in the physical world," in *Proc. of Int. Conf. on Learning Representations Workshop*, Toulon, France, pp. 1–15, 2017.

[26] J. Lin, C. Song, K. He, L. Wang and J. E. Hopcroft, "Nesterov accelerated gradient and scale invariance for adversarial attacks," in *Proc. of ICLR*, Addis Ababa, Ethiopia, 2020.

[27] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *Proc. of CVPR*, Nashville, USA, pp. 1924–1933, 2021.

[28] C. Xie, Z. Zhang, Y. Zhou, S. Bai, J. Wang *et al.,* "Improving transferability of adversarial examples with input diversity," in *Proc. of CVPR*, Long Beach, USA, pp. 2730–2739, 2019.

[29] Y. Dong, T. Pang, H. Su and J. Zhu, “Evading defenses to transferable adversarial examples by translation-invariant attacks,” in *Proc. of CVPR*, Long Beach, USA, pp. 4312–4321, 2019.

[30] W. Zhan, C. Sun, M. Wang, J. She, Y. Zhang *et al.,* “An improved Yolov5 real-time detection method for small objects captured by UAV,” *Soft Computing*, vol. 26, pp. 361–373, 2022.