LC-GAN: Improving Adversarial Robustness of Face Recognition Systems on Edge Devices

Peilun Du, Xiaolong Zheng, *Member, IEEE*, Liang Liu, *Member, IEEE*, and Huadong Ma, *Fellow, IEEE*

***Abstract*—Deep-learning-based (DL-based) face recognition has become an important application in the Internet of Things (IoT) environment. However, recent studies demonstrate that elaborate adversarial examples can mislead the results of DL-based face recognition on mobile and edge devices. Such vulnerability threats the robustness of face recognition systems and causes security issues. Generative adversarial defense methods can reform adversarial examples before input into the face recognition model to improve the accuracy under adversarial attacks. Unfortunately, the existing generative adversarial defense methods cannot completely remove the misleading features of adversarial examples due to the lack of robust encoding ability. In this article, we propose a local consistency generative adversarial network (LC-GAN) framework by adding the constraint of local consistency to force the encoder to mine consistent features in each local area, achieving robust encoding ability consequently. The framework includes three main novel designs. First, we present a patch-wise contrastive learning-based refinement stage with local consistency loss to encode robust identity features from nonsalient areas that are undamaged by adversarial attacks. Second, we use a powerful expert network to guide the training of LC-GAN for eliminating adversarial identity features. Third, we design a multilevel identity loss to enhance the identity preservation ability by unifying the local and global identity features. Experimental results on four widely used face data sets show that LC-GAN outperforms other generative adversarial defense methods.**

***Index Terms*—Adversarial attack, adversarial defense, face recognition, generative adversarial network.**

I. Introduction

FACE recognition applications are ubiquitous in the Internet of Things (IoT) environment, including surveillance, mobile business payment, access control, smartphone unlocking, etc. In recent years, deep learning methods [1], [2], [3], [4], [5], [6], [7] have achieved impressive performance in face recognition systems. For example, Yang et al. proposed SCFR [5] and SILR [6] to balance the accuracy and time-consuming training requirements of face recognition under low-quality training data in IoT environments. Yang et al. proposed RRAN [6] that utilizes metric learning to boost the performance of video face recognition. They have become the most popular face recognition methods on mobile and edge devices [5], [6], [7].

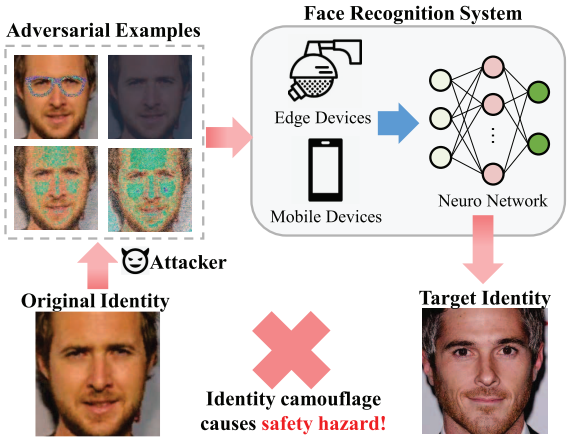


Fig. 1. Attackers transform original face images into adversarial examples, such as perturbations on the eyeglass frames. The adversarial examples will mislead the DL-based face recognition system and camouflage the original identity into the target identity.

However, recent studies find deep-learning-based (DL-based) face recognition vulnerable to adversarial examples [8], [9], [10], [11], [12], [13], [14], [15]. The threatened applications include autonomous cars [9], [11], unmanned aerial vehicles [12], IoT device classification [10], and face recognition [13], [14], [15].

The adversarial examples for face recognition include the physical adversarial eyeglasses frames [13], [14] or imperceptible perturbation scattered on the face image [16], [17]. Fig. 1 demonstrates the worrisome consequence caused by the adversarial examples above, which becomes one of the major gaps for applying DL-based face recognition in the real-world IoT system. Therefore, defending against adversarial attacks is important for robust DL-based face recognition.

Existing defense methods can be grouped into adversarial training, robust architecture, and generative adversarial defense. Adversarial training methods augment training data sets with adversarial examples to force the models to learn robust features [18], [19], [20], [21], [22]. Robust architecture methods achieve defense with augmented model architecture or inference mechanism to improve the robustness of the inference process [23]. Generative adversarial defense methods

Manuscript received 10 June 2022; revised 3 November 2022; accepted 9 December 2022. Date of publication 20 December 2022; date of current version 25 April 2023. This work was supported in part by the Funds for Creative Research Groups of China under Grant 61921003; in part by the National Natural Science Foundation of China under Grant 61932013 and Grant 62225204; and in part by the A3 Foresight Program of NSFC under Grant 62061146002. *(Corresponding author: Huadong Ma.)*

The authors are with the Beijing Key Laboratory of Intelligent Telecommunications Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China (e-mail: dupeilun1995@bupt.edu.cn; zhengxiaolong@bupt.edu.cn; liangliu@bupt.edu.cn; mhd@bupt.edu.cn).

Digital Object Identifier 10.1109/JIOT.2022.3230427

2327-4662 © 2022 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission.
See https://www.ieee.org/publications/rights/index.html for more information.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

modify the adversarial examples with an additional generative model to eliminate adversarial perturbations for a clean inference [24], [25], [26], [27], [28], [29]. Face recognition applications are usually trained with large-scale data sets [1], [30] and deployed in the authentication system on the edge and mobile devices. Augmenting such large-scale data sets with adversarial examples or retraining a new robust model on each heterogeneous IoT environment is unrealistic. Moreover, as the diversity of attacks, injecting specific types of adversarial samples to augment the training data set makes the model overfitting to such specific attacks and loses the generalization of other attacks. Generative adversarial defense is an effective plug-in method to defend against adversarial examples [27] without changing the deployed face recognition model. They encode adversarial examples into identity features and then decode these features into reformed coarse-grained images with learned distribution on clean image data sets in one-step inference [26], [27], or improve the decoding with the iterative strategies [25], [28], [29].

Unfortunately, based on the analysis of the existing generative adversarial defense methods [25], [26], [27], [28], we find that they mainly focus on improving the decoding ability while ignoring the robustness of encoding. MagNet [26] trains the classical encoder–decoder network with a simple image reconstruction loss for improving the image decoding quality. ShieldNet [27] introduces the probabilistic adversarial robustness during the image decoding process of PixelCNN [31]. Defense-GAN [25] iteratively optimizes the initial random input vector for the decoder of basic GAN [32] to achieve better image decoding quality. EGC-FL [28] utilizes feedback loops to generate cleaning images, which is similar with the pipeline of Defense-GAN. However, the powerful decoding ability cannot compensate the vulnerable encoding process. Once the encoder of the generator fails to encode adversarial samples into clean identity features, the corresponding decoder can never decode such adversarial identity features into clean images. Specifically, the limited perturbations will aggregate on salient areas (mouth, nose, and eyes) to achieve successful misleading since salient areas have significant features for face recognition. Such aggregation will destroy the local consistency of original face images. During the encoding process, the adversarial features of perturbations on the salient areas become the highest activated values in the global feature map and then submerge original clean features of nonsalient areas (forehead or cheek). Therefore, training a generator with robust encoding ability for extracting original identity features from nonsalient areas is the cornerstone of generative adversarial defense methods.

However, achieving such robust encoding ability has three major challenges. First, the deep learning models always tend to extract features in the salient areas due to the high activated values of the areas. Such a value-oriented mechanism ensures the high representational ability of the extracted features for successful decoding. Compared with salient areas, nonsalient areas are congenitally deficient in low-dimensional features, such as color and texture, which limits the encoder to extract representational original identity features. Second, the adversarial and original clean features are mixed during the layer-by-layer encoding process. Moreover, the diverse adversarial features of different adversarial examples [13], [14], [33] cannot be directly modeled with one fixed template. Therefore, filtering out the adversarial parts from mixed features is difficult. Third, the ability of identity preservation is important for robust encoding. As identity preservation focuses on features in the global view, introducing this constraint into generative adversarial defense methods may cause conflict in extracting original identity features from local nonsalient areas.

To address the above challenges, we propose a local consistency generative adversarial network (LC-GAN) to reform adversarial face images by our trained generator. First, we find that the extracted features between salient and nonsalient areas of clean images show more local consistency than those of adversarial images (Section III). Therefore, the local consistency is positively correlated with the representational ability of original clean identity features from nonsalient areas. Based on the insight of local consistency, we design a refinement stage with local consistency loss to improve the robustness of encoding. Specifically, after each face image generation, we randomly crop the image into partial face patches and utilize patch-wise contrastive learning to refine the local consistency between extracted features from salient and nonsalient areas in the training process. Second, for filtering out the adversarial parts from mixed features, instead of modeling diverse adversarial features, we extract robust clean identity features with the guidance of a well-trained expert network. By distilling the prior knowledge of the expert network, the encoder will focus on the original clean part of mixed features. Third, to achieve identity preservation of extracted features, we design a multilevel identity loss to enhance identity preservation with shallow feature maps and high-level identity features extracted by the expert network. Such a multilevel combination can unify the local and global identity features and achieve local consistency and identity feature preservation at the same time. Experimental results show that our defense method outperforms other generative adversarial defense methods in both face verification and identification tasks. Our contributions are summarized as follows.

1) We demonstrate the threat of adversarial attacks to DL-based face recognition systems and find that the existing generative adversarial defense methods lack robust encoding ability for achieving robust face recognition systems. Then, we present analysis with the expert face recognition model to show that local consistency is a significant factor for improving the robustness of encoding.
2) We propose the LC-GAN framework to extract robust original identity features from nonsalient areas. To achieve the robust encoding, LC-AGN has three novel designs, including a refinement stage with local consistency loss, the knowledge transfer of the expert network, and a multilevel identity loss.
3) We conduct extensive experiments with four state-of-the-art attack methods on four widely used face data sets to test the performance of LC-GAN. The results demonstrate that LC-GAN can train a generator with robust encoding ability for reforming adversarial examples to

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

achieve higher face recognition accuracy. Besides, LC-GAN has the minimal inference computational overhead than widely used baselines.

The remainder of this article is organized as follows. Section II reviews the related works on basic adversarial attacks, adversarial attacks on face recognition, and generative adversarial defenses. Section III demonstrates the characteristics of adversarial face images and our problem definition. Section IV describes the details of LC-GAN. Section V shows the experimental details and defense performance on four face data sets. Section VI illustrates the visualization results of our method and baselines. Finally, Section VII concludes this article.

II. Related Work

A. Basic Adversarial Attacks

Recent adversarial attacks are divided into two types: black-box and white-box attacks. White-box attackers can access all parameters, architecture, and training process of threat models, while black-box attackers cannot have any access to parameters. Bruna et al. [34] bridged the gap between black box and white box by proving that an adversarial example for one model can be generalized to others. We mainly focus on white-box attacks in this article. The fast gradient sign method (FGSM) is a classical one-step attack [18]. Given an input image x with true label y, FGSM finds the adversarial example $\tilde{x}$ by gradient direction of cross-entropy loss $J(x, y)$ and disturbance size ϵ with sign function. Project gradient descent (PGD) [33] is an iterative version of FGSM. PGD finds the perturbation for x by splitting disturbance size ϵ into several small step size α and accumulating perturbations with projection operation. The Carlini–Wagner (C&W) attack is an optimization-based attack [35] against model distillation defense by iteratively finding the adversarial examples with the smallest perturbation which leads to a high probability of misclassification. Deepfool is an iterative adversarial attack with variable step size [36]. It keeps finding optimal adversarial perturbation with minimal classification distance to other categories in each iteration. Based on Deepfool, Moosavi-Dezfooli et al. [37] proposed a classical universal adversarial attack to generate a fixed image-agnostic perturbation for attacking most images in the data set.

B. Attacks on Face Recognition

Recent evidence [13], [14], [16], [17] suggests that adversarial attacks can be employed in face recognition models, such as VGG-Face [1] and ArcFace [4]. Sharif et al. [13] utilized I-FGSM to generate small perturbations on faces, which is physically realized through printing a pair of eyeglass frames with well-designed perturbation. Furthermore, Sharif et al. [14] designed a GAN-based method to train a generator to provide adversarial examples satisfying desired objectives that can realize large-scale automated adversarial attacks. This series of physical attacks seriously endanger the security of face recognition. Dong et al. [15] proposed an evolutionary attack algorithm to improve the efficiency of decision-based black-box attack. Though this evolutionary attack improves the effectiveness of a decision-based black-box attack, it is still impractical due to a large number of query times. Zhong and Deng [17] proposed a transferable adversarial attack against face recognition. This transferable adversarial attack applies random dropout to convolutional filters to achieve ensemble performance with only one model.

C. Generative Adversarial Defenses

Benefiting from the powerful generative ability of DNNs, recent generative adversarial defense methods have resisted many adversarial attacks. We divide them into manifold transfer methods [24], [26], [27], [28] and optimization-based methods [25], [29] according to the iterative requirements.

Manifold transfer methods are one-step methods that utilize learned knowledge to transfer the adversarial distribution into clean distribution. MagNet [26] is a two-stage manifold transfer method that includes one or more separate detector networks and a reformer network based on an autoencoder–decoder mechanism. The detector learns to differentiate normal and adversarial examples by approximating the manifold distribution, and then, the reformer moves adversarial examples toward the manifold of normal examples. However, MagNet cannot produce high-resolution images limited by the architecture of autoencoder–decoder. ShieldNets [27] introduces the theoretic framework of probabilistic adversarial robustness and transfers sample probability to adversarial-free zones with PixelCNN. This method is a generative model based on conditional probability which needs to calculate the pixel of the generated image point by point to reform the input image. The resolution of images taken in the real world is much higher than test data sets. The point-by-point calculation of probabilistic adversarial robustness will affect the inference speed of the authentication system which is impractical in real-world applications.

Optimization-based methods iteratively optimize the distribution of initial random variables to approach the distribution of clean images while keeping the reformed image similar to adversarial inputs at the pixel level. Defense-GAN [25] achieves optimization-based defense with a basic GAN framework and iteratively uses gradient descent minimization to find suitable hidden variables z as the input of the learned generator to reform the input image. Zhou et al. [29] performed manifold projection for adversarial face pairs with VAE [38] and Style-GAN [39] structure. The generated faces of [29] may have misleading details due to the inherent randomness of VAE. EGC-FL [28] utilizes feedback loops to generate cleaning images for the effective defense of deep neural networks. However, the iterative strategy of the above methods is not fit for face recognition in edge computing conditions due to high computational cost. They cannot satisfy the high-resolution and real-time requirements. Moreover, all existing generative adversarial defense methods only focus on overall pixel loss of decoding while ignoring the identity feature preservation and local feature consistency of encoding.

III. Measurements of Adversarial Samples

In this section, we quantitatively analyze the characteristics of adversarial face images and find the local consistency

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

TABLE I
STATISTICS OF COSINE SIMILARITY BETWEEN PATCHES AND IMAGES ON VARIOUS SAMPLES. *P & P* IS THE COSINE SIMILARITY BETWEEN PATCH PAIRS AND *P & I* IS THE COSINE SIMILARITY BETWEEN PATCHES AND THEIR HOLISTIC CLEAN FACE IMAGES

	Clean	$FGSM_{20}$	PGD_8	PGD_{20}	C&W	Physical
P & P	**0.966**	0.923	0.893	0.906	0.884	0.901
P & I	**0.771**	0.602	0.624	0.604	0.576	0.612

problem of adversarial defense. After that, we give an interpretable insight to training a generator for defending adversarial face images and the definition of the optimization problem of LC-GAN.

A. Local Consistency of Adversarial Face Images

Face recognition needs to perform identification within the same category (human) in general tasks. Therefore, face recognition needs more "fine-grained" features of local facial details while the existing generative adversarial defense methods only focus on the global loss of pixel values and ignore the local loss of detailed facial features. Therefore, we need to pay more attention to the changes of local features after the adversarial attacks. The stability of these features is required for correct face recognition.

The idea of local consistency first comes from our observations on the physical adversarial attack with eyeglasses frames. They [13], [14] perform physical adversarial attacks with highly aggregated adversarial perturbations that are on the salient areas of the face image. It is obvious that such perturbations on eyeglasses frames will cause local inconsistency of identity features between salient and nonsalient areas. However, for the imperceptible perturbations [33], [35] scattered on the face images, the adversarial pixels are not obviously aggregated like adversarial eyeglasses frames. Therefore, a question arises: *Is the problem of local consistency still hold on to imperceptible perturbations?*

We use a well-trained expert network to assist our analysis of local consistency on imperceptible perturbations [33]. We simulate different parts of a face by randomly cropping the face images into smaller patches. During this experiments, the edge length of the cropped image patches are 2/3 of the original image. Then, we extract the identity features through the expert network and analyze the semantic distances between these patches. The semantic distances are measured by the cosine similarity between identity features. The statistics are conducted with the average cosine similarity of 1000 images on VGGFace data set [1]. The results are shown in Table I, the *P&P* denotes the cosine similarity between patches within the same face image. *P&P* of clean images is larger than adversarial examples, which means that the adversarial attacks will widen the semantic distances between patches and reduces the cosine similarity. Moreover, we analyze the semantic distance between the patch and its face image, denoted with *P&I*. The *P&I* of clean images is also larger than adversarial examples, which demonstrates that the feature identity of clean patches is closer to the original identity. Both *P&I* and *P&P* show larger cosine similarity on clean images, which means that the clean images have higher local consistency between patches and their corresponding images.

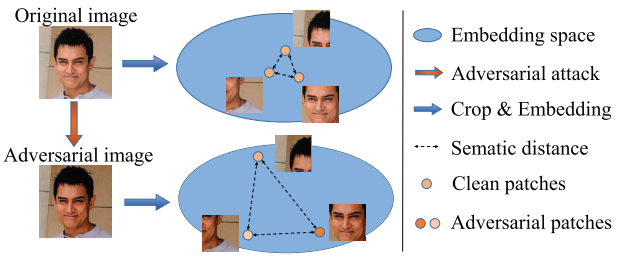


Fig. 2. Illustration of the semantic distance between image patches. The local features of the patches after adversarial attacks are farther than the original face image. The patches with more adversarial perturbation will be embedded into other identities in embedding space.

Such results are caused by the uneven perturbations scattered on face images. The imperceptible adversarial attack aims to mislead the predictions of target model with imperceptible perturbations under limited budget. The process can be denoted as

$$\text{argmax}\mathcal{L}(f(\tilde{x}), y) \quad \text{s.t.} \quad \|\tilde{x} - x\|_\rho \leq \epsilon \tag{1}$$

where $\mathcal{L}$ is the loss function of target model f, $\tilde{x}$ is adversarial examples, and ρ is the distance metric, usually is the Euclidean or infinite norm. The imperceptible adversarial attack will limit the budget of perturbations. As the salient areas have significant features for face recognition, the limited perturbations are more likely aggregated there to achieve successful misleading. Therefore, the experiments prove that the problem of local consistency still holds in imperceptible perturbations. Fig. 2 illustrates the phenomenon of local consistency. The adversarial perturbations will cause adversarial identities on salient areas, which will further change the detailed features of local patches.

Based on the above insight, we can pull the semantic distance between patches of salient and nonsalient areas closer to make the patches achieve local consistency like a clean image. After achieving local consistency, the features extracted from the patches will be similar to those from the nonsalient areas since the nonsalient areas have a higher proportion of face images. As face recognition maps the image into an identity feature vector of embedding space. We need to ensure the identity feature of the reformed image are consistent before and after the adversarial attacks. After that, we propose a novel optimization objective for generative adversarial defense methods with local feature consistency and identity preservation.

B. Problem Definition

The classical optimization-based generative adversarial methods are optimized by

$$\mathcal{L}_o = \mathbb{E}_{x \sim p_{\text{data}}}\left[\min_z ||G_o(z) - x||_\rho\right] \tag{2}$$

where z is the hidden variable, G_o is the generator of optimization-based generative adversarial defense methods, and ρ is the metric for measuring the distance between the

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

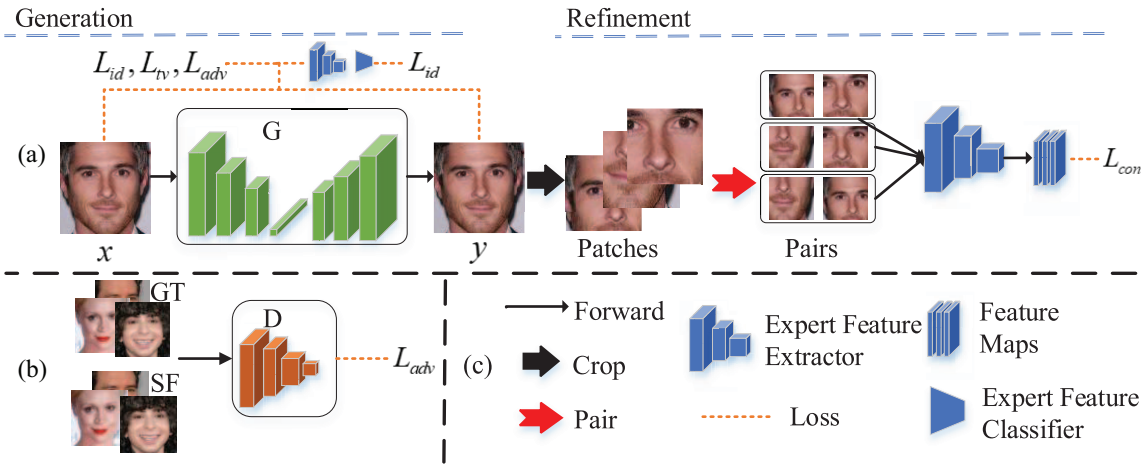


Fig. 3. Illustration of LC-GAN training framework. It illustrates (a) training process of generator G with the generation and refinement stage, (b) training process of discriminator D to distinguish between synthesized face images (SF) and ground-truth (GT) face images, and (c) legends of LC-GAN. Note that (a) and (b) are performed alternately during the whole training process. After the training, the inference only needs to use G for reforming adversarial images.

adversarial and reformed image. The goal of (2) is to completely reconstruct the input images, which is unreasonable for defending against adversarial attacks. Completely reconstructing the input images means that the adversarial perturbation is recovered together. Manifold projection [29] uses VAE to introduce variability and optimizes hidden variable z with KL-divergence to approach clean distribution of face images. However, it is inefficient to iteratively optimize the hidden variable. Moreover, the inherent randomness of VAE may cause misleading details on generated face images. Therefore, we utilize the structure of autoencoder to avoid the uncontrollable randomness and improve the efficiency of the generative methods

$$\mathcal{L}_{\text{target}} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\min ||G_t(x) - x||_\rho\right] \tag{3}$$

where the input of generator is replaced with face image x, and G_t is the generator of manifold transfer methods.

We improve the objective of (3) with local consistency and identity preservation. The identity feature preservation minimizes the distance between identity features of input and reformed face images

$$\mathcal{L}_{\text{identity}} = \mathbb{E}_{x\sim p_{\text{data}}}\left[\min_{x^*} ||\text{id}_{x^*} - \text{id}_x||_\rho\right] \tag{4}$$

where id is the identity feature, and x^* is the reformed face. The local feature consistency of x^* can be denoted as

$$\mathcal{L}_{\text{consisitency}} = \mathbb{E}\left[\min_{x^*} ||\text{con}_{p_i} - \text{con}_{p_j}||_\rho\right] \quad p_{i,j} \in x^* \tag{5}$$

where p_i and p_j denote the patches randomly cropped from x^*, and con is the consistency feature from the image patches. Combining $\mathcal{L}_{\text{identity}}$ and $\mathcal{L}_{\text{consisitency}}$ with $\mathcal{L}_{\text{target}}$, the main objective in this article can be denoted as

$$\arg\min \mathcal{L}_{\text{target}} + \mathcal{L}_{\text{identity}} + \mathcal{L}_{\text{consistency}}. \tag{6}$$

The details of conducting these loss functions are described in Section IV.

IV. Local Consistency Generative Adversarial Network

A. Overview of Training Framework

The training process of LC-GAN is shown in Fig. 3. LC-GAN includes a generator G to reform the input face image into the original identity, and a discriminator D as an adversary of G to distinguish the realness generated by G. We design a refinement stage for G after image generation. Fig. 3(a) and (b) are performed alternately during the whole training process. First, x from the face set is fed into G to extract latent features and generate the reformed face y. Then, y is randomly cropped, and the patches are paired into face patch pairs for local consistency refinement. Our novel generator loss can be divided into two parts: 1) multilevel identity loss $\mathcal{L}_{\text{id}}$ and 2) local consistency loss $\mathcal{L}_{\text{con}}$. Moreover, we utilize the expert face recognition model to provide powerful extracted features for our generator losses. We use the conventional discriminator loss $\mathcal{L}_{\text{adv}}$ for training D and G. Besides, we apply total variation loss $\mathcal{L}_{\text{tv}}$ to smooth the reformed faces and pixel-wise loss $\mathcal{L}_p$ to reconstruct input face images.

B. Generator With Refinement Stage

To realize the consistency of local features, we design a refinement stage with patch-wise contrastive learning to compute local consistency loss. As described in Section III, most of the face area is not covered or affected (adversarial identity) by the perturbations. Therefore, we take advantage of this incomplete coverage defect to mine the clean local identity feature of undisturbed nonsalient areas to defend against adversarial attacks on face recognition.

First, we use the generator to reconstruct input faces and use multilevel identity loss $\mathcal{L}_{\text{id}}$ and pixel-wise loss $\mathcal{L}_p$ to reduce the identity error and overall pixel error between the input and reformed face images. In the refinement stage, we design a patch-wise contrastive learning method to mine the local consistency of the face image, which enables us to eliminate the influence of the adversarial texture when the attacks are agnostic. The reformed face image is cropped into patches

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

with different sizes through our pipeline, and then the patches are randomly combined into patch pairs. We use the expert network to extract the features of patch pairs and compare the similarity of the patches within each patch pair to measure the consistency as local consistency loss $\mathcal{L}_{\text{con}}$. This forces our model to pay more attention to extracting original identity from nonsalient areas and local consistency of face image.

C. Multilevel Identity Loss With the Expert Model

We utilize the classical face recognition network as our expert network to provide powerful feature extraction. As shown in Fig. 3(c), the expert network consists of the feature extractor and feature classifier. The identity feature is the output of the feature classifier and we penalize the feature distance between the reformed face and input face to preserve face identity. Based on ensuring the identity feature similarity between the two images, we further use the distance of feature maps extracted by the feature extractor to improve identity preservation. The feature maps have more detailed spatial information on local features. The combination of the two-level features can unify the local and global identity features for achieving better identity preservation. In this article, we use VGG-Face [1] and ArcFace [4] to efficiently map face images to identity features.

D. Loss Functions of LC-GAN

For training LC-GAN with the main objective in (6), the loss functions are composed of $\mathcal{L}_D$ for training D and $\mathcal{L}_G$ for training G. In this section, we describe the mathematical definition and actual calculation process of each component for $\mathcal{L}_D$ and $\mathcal{L}_G$.

1) Adversarial Loss: The adversarial loss $\mathcal{L}_{\text{adv}}$ is originally introduced by Goodfellow et al. [32]. Generator G and discriminator D in the GAN architecture are learned in an adversarial fashion with $\mathcal{L}_{\text{adv}}$

$$\mathcal{L}_{\text{adv}} = \log D(x) + \log(1 - D(G(x))) \tag{7}$$

where the first item $\log D(x)$ is the classification loss of real face images, and the second item $\log(1 - D(G(x)))$ is the loss of reformed face images.

2) Multilevel Identity Loss: To maintain the identity of the face image, we penalize the identity feature distance between the reformed and input face image. We introduce an expert network to provide the ability of identity feature extraction. Different from the previous work that only uses outputs of the feature classifier as identity features [40], we use the middle feature maps of the feature extractor to enhance our identity preservation ability like perceptual loss [41]. The feature maps extracted by CNN will retain spatial features of the face image, which describe the spatial distribution of facial semantic information on the image. Our multilevel identity loss is the sum of feature map loss and identity feature loss. Let F denote the expert network, we express the multilevel identity loss function as follows:

$$\mathcal{L}_{\text{id}} = \|F_e(x) - F_e(y)\|_2^2 + \|F_c(F_e(x)) - F_c(F_e(y))\|_2^2 \tag{8}$$

where $y = G(x)$ is the reformed faces from generator G, and F_e and F_c are the feature extractor and feature classifier of the expert model, respectively. Specifically, the output of F_e is 2-D feature maps through the last CNN layer, the output of F_c is 1-D feature vector through the last FC layer.

3) Local Consistency Loss: In the refinement stage, we crop the reformed face images into patches with random positions to simulate different areas of the face image. As shown in Table I, the *P&I* cosine similarity of adversarial samples is larger than 0.6 on average, which means some adversarial remain the original identity. For example, when the image contains the whole face, the salient areas may be the facial landmarks, so the perturbation pixels mainly gather there. When our image patches are randomly sampled to the cheek, the salient areas of the original face will disappear, so this place will retain more original identity information rather than perturbations. We randomly pair the generated patches into patch pairs in pair set P. Some of these patches contain more identity information and texture features of adversarial perturbations, and others contain more identity and texture features of the original clean images.

Then, we use the expert network to extract the features of patch pairs and penalize the similarity of the patches in each pair to measure the local consistency. Since the nonsalient areas have a higher proportion of face images, the features of all patches will be pulled toward clean features from the nonsalient areas during the optimization. For measuring the semantic similarity, we use an encoder of the expert model to extract the feature maps. However, such 2-D feature maps cannot directly be used for measurements. We transform the feature maps into 1-D vectors and apply the softmax function to make them a kind of computable probability distribution. Let e_n^0 and e_n^1 denote the feature maps of patches extracted by F_e, $\mathcal{L}_{\text{con}}$ can be denoted as

$$\mathcal{L}_{\text{con}} = \sum_{n=1}^{N} KL\left(\log\left(\mathcal{T}\left(e_n^0\right)\right), \mathcal{T}\left(e_n^1\right)\right) \tag{9}$$

where N is the size of patch pair set P, and $\mathcal{T}$ is the transform operation with softmax that transforms the feature maps into 1-D vectors. KL is the Kullback–Leibler divergence loss function which measures the similarity between two vectors. We set $N = 3$ for appropriate training costs.

4) Pixel-Wise Loss: We use the pixel-wise loss to penalize the content error of reformed images. Specifically, the pixel-wise loss measures the pixel-level global differences like landmark location or texture features between input images and reformed images. We adopt the L^1 loss function to calculate pixel-wise loss, where W and H denote the width and height of the image, respectively

$$\mathcal{L}_p = \frac{1}{W \times H} \sum_{w=1}^{W} \sum_{h=1}^{H} |x_{w,h} - G(x)_{w,h}|. \tag{10}$$

5) Total Variation Loss: To encourage spatial smoothness in the reformed face images, we apply the total variation regularization with L^1 loss function to maintain the smoothness

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

Algorithm 1: LC-GAN Training Process

Input: Generator G, discriminator D, training set $\mathcal{D}_{face}$, expert network F, including extractor F_e and classifier F_c.
Output: Trained generator G.

1 Initialize the parameters of G and D;
2 **for** *epoch* < *Maximum iterations* **do**
3 Set *inner_iter* ← 0;
4 **for** x *in* $\mathcal{D}_{face}$ **do**
5 Reformed face image $x^* \leftarrow G(x)$;
6 $\mathcal{L}_{adv} \leftarrow$ Eq. (7);
7 **if** *inner_iter* < 5 **then**
8 Update D with backpropagation of $\mathcal{L}_{adv}$;
9 Clean the gradient;
10 *inner_iter* = *inner_iter* + 1;
11 **else**
12 Set *inner_iter* ← 0;
13 Random crop x^* and generate patch pair set P;
14 Get $\mathcal{L}_{id}$ and $\mathcal{L}_{con}$ according to Eq. (8)(9) with the guidance of F;
15 Get $\mathcal{L}_G$ according to Eq. (10)(11)(13);
16 Update G with backpropagation of $\mathcal{L}_G$;
17 Clean the gradient;
18 Return trained generator G;

by limiting the variation between adjacent pixels

$$\mathcal{L}_{\text{tv}} = \sum_{w,h} |y_{w,h} - y_{w,h+1}| + |y_{w,h} - y_{w+1,h}| \tag{11}$$

where $y = G(x)$ is the reformed face images from generator G, and w and h are the 2-D coordinate of each pixel in y.

6) Overall Loss Function: Combining the loss functions above, we conduct $\mathcal{L}_D$ and $\mathcal{L}_G$ to train LC-GAN. As shown in Fig. 3, we use the conventional discriminator loss $\mathcal{L}_{\text{adv}}$ for D. We use other components plus $\mathcal{L}_{\text{adv}}$, including $\mathcal{L}_p$, $\mathcal{L}_{\text{id}}$, $\mathcal{L}_{\text{con}}$, and $\mathcal{L}_{\text{tv}}$, for training G with multiple optimization objectives. The overall loss function with the weight coefficient λ_i is

$$\mathcal{L}_D = \mathcal{L}_{\text{adv}}, \tag{12}$$

$$\mathcal{L}_G = -\lambda_1 \mathcal{L}_{\text{adv}} + \lambda_2 \mathcal{L}_p + \lambda_3 \mathcal{L}_{\text{tv}} + \lambda_4 \mathcal{L}_{\text{id}} + \lambda_5 \mathcal{L}_{\text{con}}. \tag{13}$$

E. Overall Training Process

During the training, we use the above loss functions to iteratively optimize D and G. As shown in Algorithm 1, we first forward x into G to generate reformed face image x^* and calculate conventional adversarial loss (lines 5 and 6). The training of D and G is optimized alternatively. When we train D, we will fix the parameters of G and vice versa. According to [40], we update G after five updates of D due to the unstable update of D (lines 7–10). For training G, we get $\mathcal{L}_{\text{id}}$ and $\mathcal{L}_{\text{con}}$ with the guidance of the expert network (lines 13 and 14). Then, we calculate the overall loss of G and update the model (lines 15–17). Note that the refinement stage of LC-GAN is silent after training, only the generator G is employed for adversarial defense.

Algorithm 2: Robust Face Recognition

Input: Testing face images $\tilde{X}$, Trained generator G, protected face recognition model f, detection threshold T, face database $\mathcal{D}_{face}$.
Output: Face recognition result.

1 **for** $\tilde{x}$ *in* $\tilde{X}$ **do**
2 Transform $\tilde{x}$ into appropriate size;
3 Generate reformed face image $x^* \leftarrow G(\tilde{x})$;
4 $Dis \leftarrow \|\tilde{x} - x^*\|_2^2$;
5 **if** $Dis > T$ **then**
6 Resist $\tilde{x}$ for face recognition;
7 **else**
8 $Logit \leftarrow f(x^*)$;
9 Return identity in $\mathcal{D}_{face}$ according to *Logit*;

F. Robust Face Recognition

After the training, we use the well-trained generator to reform the adversarial face images into their original identity during testing. However, the generator cannot reform the adversarial samples that are greatly distorted by the large attacking budget. The perturbation of such adversarial samples can destroy the identity features of all areas, including non-salient areas. The defense process is shown in Algorithm 2, under this situation, LC-GAN can be an adversarial detector by measuring the L^2 distance between the reformed and the input face image (lines 3–6). Therefore, we design a filter threshold of T to further improve reforming performance for face recognition tasks. When the distance between input and reformed face images exceeds the threshold, we determine the restorations are failed and resist these adversarial face images.

The data structure of the database is a matrix that stores the identity feature vector of all legal identities. The identity feature vector of each person is the average of the 1-D identity feature vector extracted from multiple sampled images of this person. We return the identity that has the minimal cosine distance between the average identity feature vector and the identity feature vector of the input face image.

V. Experiments

This section compares LC-GAN with recent state-of-the-art defensive methods, including manifold transfer methods [26], [27], [28], optimization-based methods [25], [29], and some image processing methods [42], [43] on widely used face data sets [1], [44], [45], [46]. We evaluate LC-GAN in both adversarial attack [13], [18], [35], [47] and clean situations.

A. Data Sets and Baselines

1) Data Sets: As for the training stage, we train our network on two large-scale data sets: 1) VGGFace [1] and 2) CASIA-WebFace [30]. VGGFace contains 2622 celebrities

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

with roughly 1000 images per celebrity. CASIA-WebFace contains 494 414 face images with 10 575 identities. All the face images are preprocessed with face alignment and resized to 224×224.

We evaluate LC-GAN on four widely used face data sets, including LFW [44], AgeDB-30 [45], CFP-FP [46], and VGGFace [1]. We make verification tests on LFW, AgeDB-30, and CFP-FP. LFW [44] has 13 233 face images from 5749 identities. The images make up 6000 face pairs that 3000 pairs are from the same identity and another 3000 pairs. AgeDB-30 [45] contains 12 240 images with 568 different identities of celebrities, such as actors, writers, scientists, and politicians, each with identity, age, and gender attributes. The images make up 6000 face pairs. CFP-FP [46] contains 7000 images with 500 different identities and 7000 face pairs. We measure the identification accuracy on VGGFace [1]. We randomly select 6000 images with 300 celebrities from VGGFace to perform the evaluation. All the face images are preprocessed same as the training stage.

2) Baselines: We study the performance of LC-GAN and compare it with widely used baseline methods that are denoted as *Random*, *Transform*, *Mag*, *Shield*, *Ensemble*, *Manifold*, and *De-GAN*. *Random* [43] and *Transform* [42] are preprocessing methods that defend against adversarial attacks by general image processing, including image compression, pixel quantization, random resizing, and padding. *Mag* [26] is a classical manifold transfer method that uses an autoencoder to extract latent features with learned distribution and reforms the adversarial input with an autodecoder. *Shield* [27] is the model with improved probabilistic adversarial robustness. *Ensemble* [28] is the recent state-of-the-art generative adversarial defense method with the iterative mechanism. *De-GAN* [25] is a classical optimization-based method that utilizes GAN [32]. *Manifold* [29] is the recent SOTA generative adversarial defense method for face verification which combines with VAE [38], image processing, and style-GAN [39].

B. Training Details of LC-GAN

We follow the network architecture in [40] to build our generator. The batch size is 32, the initial learning rate is 0.001 adjusted with training iteration. For training on VGGFace, we empirically set $\lambda_1 = 10^{-3}$, $\lambda_2 = 1$, $\lambda_3 = 10^1$, $\lambda_4 = 1$, and $\lambda_5 = 10^3$, and the gradient penalty coefficient is 10. The training iteration is 3000. For training on CASIA-WebFace, we set $\lambda_1 = 1$, $\lambda_2 = 20$, $\lambda_3 = 10$, $\lambda_4 = 10$, and $\lambda_5 = 5$. The training iteration is 1000. The principle of setting hyperparameters is keeping the magnitude of each loss at a balanced level to achieve a better training effect. For example, when training with VGGFace data set, the magnitude of $\mathcal{L}_{\text{adv}}$ is very large, up to hundreds or even thousands. Therefore, we set λ_1 to 10^{-3} to make the overall training process in a balanced state. After the operation, we keep each loss term within the interval of 0.5–1. We utilize VGGFace as the expert model for training our generator on VGGFace, and ArcFace for CASIA-WebFace. Our LC-GAN and baselines are trained on VGGFace for identification and trained on CASIA-WebFace for verification.

C. Adversarial Attacks for Testing

We evaluate the defense performance under the widely used adversarial attacks for face recognition, including FGSM [18], PGD [33], C&W [35], and physical attack [13]. Note that the physical attack can only attack the front face images due to the fixed mask of the eyeglasses frames. As the face images in verification tasks have diverse perspectives, we only apply physical attack on the identification task with adversarial eyeglasses frames. For verification tasks, we perform the C&W attack for 500 maximum iterations in L^2 norm to attack most of the faces. The learning rate is 0.1, ϵ is 8, and initial c is 100. For FGSM and PGD, we conduct attacks with $\epsilon = 8$ under the L^∞ constraint. The number of iterations of PGD is 50. For the identification task, we apply FGSM and PGD with diverse attack budgets to achieve different intensity attacks. The ϵ of PGD includes 8, 20, and 50. The number of iterations of PGD is 50. The ϵ of FGSM includes 20 and 50. We perform the C&W with L^1 and L^2 norm. In the physical attack, we set the step size α is 20 according to [13], and the maximum number of iterations per face attack is 300. When the identity with the highest confidence is different from the ground truth, the attack is successful.

The adversarial data set is composed of face images from LFW [44], AgeDB-30 [45], CFP-FP [46], and VGGFace [1]. The adversarial LFW, AgeDB-30, and CFP-FP are verification data sets with 6k, 6k, and 7k face pairs, respectively. The adversarial attacks for verification data sets are target attacks, we attack the second face image of each pair with the identity of the first face image. The adversarial VGGFace contains 300 celebrities, each with 20 adversarial examples. The adversarial attacks for VGGFace are non-target attacks to dodge the original identity of face images.

D. Testing Metric

We evaluate the defense performance by measuring the robust verification and identification accuracy of reformed faces. We utilize the two SOTA pretrained face recognition models, VGG-Face [1] and ArcFace [4], as the protected model. We evaluate the robust verification accuracy on LFW [44], AgeDB-30 [45], and CFP-FP [46] and identification accuracy on VGGFace [1]. Note that the face images in verification tasks have diverse perspectives while the generator trends to generate face images with the front perspective. The L^2 distance between the reformed and the input face image cannot be the basis for detecting adversarial samples. Therefore, we only test the performance of the adversarial detector on the identification task. We set $T = 10$ in our identification experiments for *Mag* and *Ours*. Both resistance and reformation are successful defense to improve the robust accuracy. We report the resist and robust accuracy, respectively, in the following detailed experiments.

E. Experimental Results

We conduct experiments of verification task on LFW [44], AgeDB-30 [45], and CFP-FP [46], and experiments of identification task on VGGFace [1]. The verification tasks aim to find whether the input pair of face images is the same identity. The

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

TABLE II
VERIFICATION ACCURACY (%) OF DIFFERENT BASELINES DEFENSE AGAINST STATE-OF-THE-ART ATTACKS ON LFW, AGE-DB, AND CFP-FP

LFW (Same identity pairs/ Different identities pairs / Average)				
Defense method	Clean	FGSM	PGD	C&W
No Defense	98.7/99.8/99.2	44.4/81.7/63.1	26.8/74.8/50.8	24.0/52.7/38.3
Random [43]	97.7/98.4/98.1	49.8/75.4/62.6	42.2/70.5/56.3	54.2/68.8/61.5
Transform [42]	**98.8/99.4/99.1**	54.1/68.0/61.0	40.5/73.3/56.9	34.3/73.3/53.8
Mag [26]	91.2/96.5/93.9	59.1/68.5/63.8	56.2/70.4/63.3	56.9/71.0/63.9
Ensemble [28]	92.7/85.1/88.9	56.0/62.5/59.3	52.3/67.2/59.7	58.3/60.7/59.5
Manifold [29]	90.7/92.3/91.5	33.6/81.6/57.6	37.0/78.4/57.7	34.1/82.5/58.3
Ours	90.4/93.0/91.7	**57.4/77.7/67.6**	**51.3/83.9/67.6**	**48.5/80.3/64.4**
Age-DB (Same identity pairs/ Different identities pairs / Average)				
Defense method	Clean	FGSM	PGD	C&W
No Defense	92.8/95.2/94.0	5.8/35.7/20.7	0.3/35.3/17.7	1.0/1.7/1.3
Random [43]	89.9/91.0/89.5	55.6/49.4/52.5	37.3/37.5/37.4	45.9/44.9/45.4
Transform [42]	**91.1/94.0/92.6**	48.8/50.7/49.7	31.3/23.5/27.4	38.1/50.1/44.1
Mag [26]	89.6/88.2/88.8	49.6/58.3/54.0	56.7/51.3/54.0	40.5/65.6/53.1
Ensemble [28]	88.9/85.7/87.3	59.7/45.2/52.4	58.7/45.8/52.3	61.3/44.5/52.9
Manifold [29]	90.0/89.0/89.5	47.7/57.6/52.7	48.4/55.1/51.8	47.3/57.6/52.5
Ours	91.5/90.1/90.8	**57.7/64.9/61.3**	**49.3/59.5/54.4**	**51.3/63.2/57.3**
CFP-FP (Same identity pairs/ Different identities pairs / Average)				
Defense method	Clean	FGSM	PGD	C&W
No Defense	93.3/97.3/95.3	12.1/50.2/31.2	0.0/47.4/23.7	1.7/4.1/2.9
Random [43]	90.2/93.9/92.1	48.5/70.5/59.5	42.1/77.0/49.6	41.6/60.4/51.1
Transform [42]	**94.2/94.7/94.5**	45.5/66.4/55.9	27.2/47.2/37.2	48.3/57.7/49.5
Mag [26]	91.8/93.7/92.8	44.2/65.7/54.9	48.2/61.1/54.6	48.2/64.9/56.5
Ensemble [28]	87.8/90.0/88.9	52.2/56.5/54.4	**48.3/60.9/54.6**	50.3/59.5/54.9
Manifold [29]	90.2/93.2/91.7	42.0/69.9/56.0	47.4/59.7/53.6	40.5/70.5/55.5
Ours	90.0/90.5/90.2	**60.0/62.3/61.2**	48.9/57.2/53.1	**55.8/58.7/57.3**

identification tasks use the face recognition model to query the identity of input face images.

1) Results of Verification Tasks: We examine the effectiveness of LC-GAN on three widely used verification face data sets: 1) LFW [44]; 2) AgeDB-30 [45]; and 3) CFP-FP [46]. The protected model is well-trained ArcFace [4] with CASIA-WebFace [30]. From the results on verification tasks in Table II, we obtain the following observations.

1) We find that LC-GAN improves our verification accuracy from the *No Defense* setting and gains over the best baseline under FGSM, PGD, and C&W attacks.
2) From the perspective of attacking intensity, C&W is the most threatening adversarial attack that causes the lowest robust accuracy under *No Defense* condition. The attack intensity of PGD is higher than that of FGSM, ranking second.
3) The PGD robust accuracy of *Ours* on CFP-FP is slightly behind *Ensemble*. We believe that the reason for such a result is that the data in CFP-FP is composed of front and side face images. The side faces lack symmetry, which reduces the effect of local consistency.
4) The robust accuracy of different identity pairs is higher than the same identity pairs. The reason is that the adversarial attack performs target attack on different identity pairs, which is more difficult than untarget attack.
5) The preprocess methods *Random* and *Transform* achieve better verification accuracy with clean image data but the worst defensive performance.

The results show the tradeoff of accuracy between clean and adversarial images. Our method achieves the highest defensive performance while keeping reasonable accuracy on clean images (above 90%).

TABLE III
IDENTIFICATION RESULTS (%) WITH DIFFERENT BASELINES AGAINST STATE-OF-THE-ART ATTACKS

Attack	No Defense	De-GAN	Mag	Shield	Ours
$FGSM_{20}$	70.85	13.53	52.36	69.64	**81.85**
$FGSM_{50}$	23.68	20.41	56.50	32.15	**87.56**
PGD_{8}	33.64	28.56	55.43	61.41	**78.68**
PGD_{20}	18.04	15.88	58.87	31.30	**77.84**
PGD_{50}	6.83	11.52	42.45	24.07	**89.75**
$C\&W_{L^1}$	8.22	13.75	50.71	38.48	**67.94**
$C\&W_{L^2}$	5.56	11.76	55.84	36.41	**71.56**
Physical	67.82	14.32	86.17	59.17	**87.09**

2) Results of Identification Task: We test LC-GAN on the VGG-Face data set against four different attack methods: 1) FGSM; 2) PGD; 3) C&W attack; and 4) *Physical* attack. The clean identify accuracy of the threat model is 92%. As shown in Table III:

1) LC-GAN significantly outperforms the three other baseline defense methods. Compared with the baseline method with the second-highest performance, the accuracy can be improved by up to 65% under various adversarial attacks.
2) The defense performance of FGSM, PGD, and *Physical* is better than the C&W attack, which means the C&W attack is more destructive than other attacks.
3) LC-GAN can resist the attack of FGSM and PGD with the accuracy of more than 80% on average.
4) The accuracy of Defense-GAN is lower than No Defense accuracy under each adversarial attack which means its

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

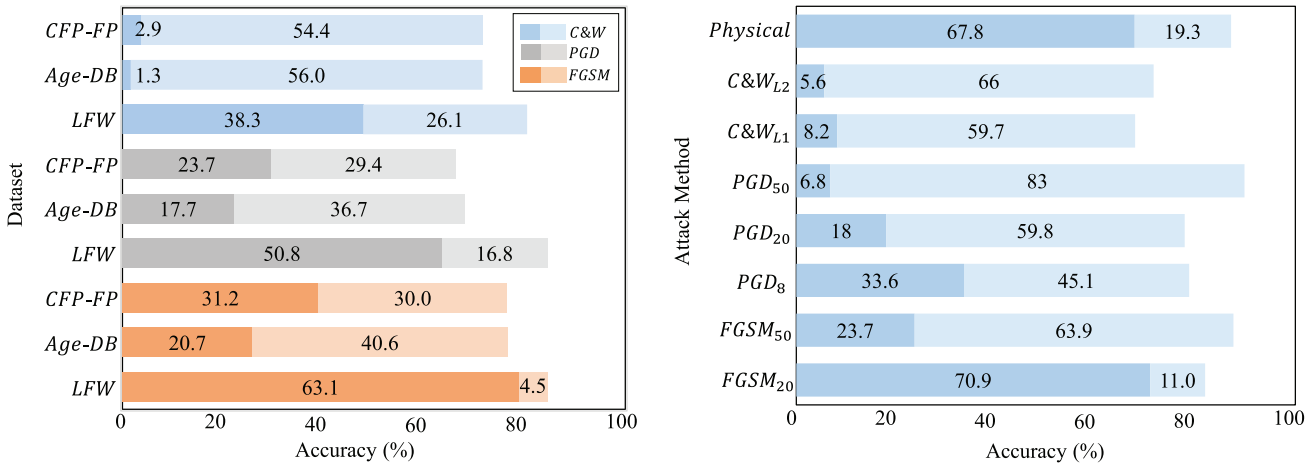


Fig. 4. Security analysis results. The dark color part of each bar denotes the vulnerable accuracy of the face recognition model under various adversarial attacks, and the light color part denotes the improved accuracy by our proposed method. (a) Security analysis on face verification task. (b) Security analysis on face recognition task.

optimization method for input hidden variables lacks a customized design for face recognition.

F. Security Analysis

Based on Tables II and III, we conduct a security analysis to further illustrate the vulnerability of existing face recognition systems and the robustness gained from our proposed method in Fig. 4. The dark color part of each bar denotes the vulnerable accuracy of the face recognition model under various adversarial attacks, and the light color part denotes the improved accuracy by our proposed method. As shown in Fig. 4(a), adversarial attacks will reduce the accuracy of face verification tasks on LFW, Age-DB, and CFP-FP data sets. Our proposed method can improve the verification accuracy significantly. As shown in Fig. 4(b), our proposed method achieves similar improvements as Fig. 4(a). Moreover, our proposed method can defend against various adversarial attacks with different attacking budgets. We consider such generality is due to the clean training process of LC-GAN which does not inject any adversarial samples during the training of our generator. Therefore, applying our proposed method can effectively avoid the security problems caused by the damage of adversarial attacks.

G. Ablation Studies

1) Influence of Local Consistency Loss: As shown in Table IV, the model training with our local consistency loss achieves better defense performance than the GAN model without local consistency loss. It shows that our consistency loss does learn local information from adversarial samples and promotes the performance in eliminating adversarial identities during local consistency refinement. We can obtain the following observations from Table IV.

1) C&W is still the most threatening adversarial attack that causes the lowest robust accuracy under the *No Defense* condition, from the perspective of attacking intensity.

TABLE IV
EXPERIMENTAL RESULTS (%) AGAINST DIFFERENT ATTACKS WITH LOCAL CONSISTENCY LOSS ABLATION STUDY ON IDENTIFICATION TASK

	No Defense	Ours w/o L_{con}	Ours w/ L_{con}
Clean	92.00	73.75	**79.27**
FGSM_{20}	70.85	57.31	**81.85**
PGD_{8}	33.64	62.95	**78.68**
PGD_{20}	18.04	68.12	**77.84**
$\text{C\&W}_{L^2}$	5.56	71.14	**71.56**
Physical	67.82	39.57	**87.09**

2) Our method has the best overall performance under a variety of attack conditions, especially under the physical attack [13], [14]. Training with local consistency loss improves the robust accuracy by about 47.5% compared with training without local consistency loss.
3) The robust accuracy on clean images is lower than the robust accuracy under FGSM_{20} and *Physical*. We consider this result is the necessary tradeoff for improving the adversarial robustness.

The overall average accuracy remains above 80% under all conditions, which is acceptable.

Compared with digital adversarial attacks, including FGSM, PGD, and C&W, the physical adversarial attack is more likely to be implemented in real-world face recognition applications. *Physical* can perform the adversarial attack by generating meaningful patterns by limiting the attacking area, such as tattoos on the face, glasses frame, and patterns on the hat. Therefore, we believe that training with local consistency loss is essential due to it has a significant defense effect against such harmful attacks.

2) Influence of Patch Size: We investigate the impact of different patch sizes in the refinement stage. There are three kinds of sizes of patches: 1/2, 2/3, and 3/4. In essence, the patch size is a tradeoff between local and global features, and a smaller size will theoretically mine more subtle local textures. As shown in Table V, the best average result occurs in the middle patch size. Small size (1/2) is not suitable for patch-wise contrastive learning and large size (3/4) will ignore

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

TABLE V
EXPERIMENTAL RESULTS (%) OF LC-GAN TRAINED WITH DIFFERENT PATCH SIZES

Attack	No Defense	1/2	2/3	3/4
$FGSM_{20}$	70.85	74.75	**81.85**	78.95
PGD_{20}	18.04	58.20	**77.84**	73.56
C&W	5.56	62.65	**71.56**	61.43
Physical	67.82	**89.34**	87.09	86.17

TABLE VI
EXPERIMENTAL RESULTS W/ AND W/O DEFENDING FILTER. THE FILTER CAN HELP LC-GAN FURTHER IMPROVE THE DEFENSE PERFORMANCE

Method	$FGSM_{20}$	PGD_{20}	C&W	Physical
Ours w/o threshold	69.4	59.2	65.8	54.3
Ours w/ threshold	81.9	77.8	71.6	87.1

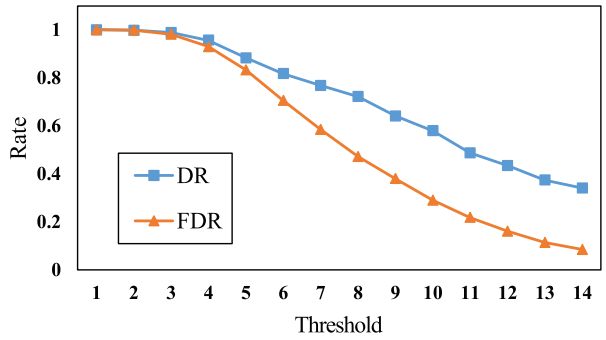


Fig. 5. Detection performance under various thresholds on PGD. *DR* is the detection rate, and *FDR* is the false detection rate.

local features of face images. Therefore, we use the middle patch size in this article.

3) Influence of Filter for Identification: As shown in Table VI, the defending filter for identification tasks improves the reforming accuracy of LC-GAN which achieves more than 10% improvements. When the image is greatly damaged by adversarial attacks, the reformed performance is limited. It is necessary to filter the damaged images to further defend against adversarial attacks. The setting of the threshold will affect the recognition results. Small thresholds will filter out the originally clean samples, and large settings will filter out less adversarial samples.

We measure the detection rate (DR) and false DR (FDR) to evaluate the detection performance under various thresholds on PGD with a budget of 8. FDR is the proportion of filtered correctly recognized faces to correctly recognized faces. The results are shown in Fig. 5, the DR and FDR are decreasing with the increase of the threshold. We select $T = 10$ for the appropriate DR to avoid a large FDR.

4) Influence of Expert Network: We conduct experiments of LC-GAN in three different settings of the expert network, with VGGFace [1], with MobileFace [48], and without any expert network. The results are shown in Table VII, training with VGGFace achieves better performance than MobileFace and without any expert network. The VGGFace has more parameters than MobileFace, which means more prior knowledge. It demonstrates that the distillation of the powerful expert network is an important module. The reason is that the expert network can indirectly pass the learned knowledge to our generator with the feature distance measurement operations of LC-GAN.

TABLE VII
IDENTIFICATION RESULTS (%) COMPARISON OF DIFFERENT EXPERT NETWORKS ON LC-GAN. *w/o Exp.* IS THE LC-GAN TRAINED WITHOUT ANY EXPERT NETWORK

Attack	w/o Expert	MobileFace [48]	VGGFace [1]
PGD_8	62.17	73.29	**78.68**
PGD_{20}	58.81	68.77	**77.84**
PGD_{50}	53.25	76.23	**89.75**

TABLE VIII
EFFICIENCY COMPARISON OF LC-GAN AND BASELINES. THE METRIC INCLUDES MODEL SIZE, FLOPS, AND INFERENCE TIME

Method	Size (MB)	FLOPS (M)	Time (s)
Mag [26]	3.83	469.60	0.43
De-GAN [25]	**1.02**	**114.09**	12.87
Shield [27]	13.28	75831.64	11.43
Ensemble [28]	3.85	1053.60	1.14
Manifold [29]	1.67	177.01	3.78
Ours	3.83	469.60	**0.42**

H. Efficiency of LC-GAN

We conduct experiments for comparing the efficiency of LC-GAN with SOTA methods from the perspectives of IoT environments. The metric includes model size, FLOPS, and inference time. The *Size* is the total parameter size of a model, *FLOPS* is floating-point operations per second of a model, *Time* is the time required to generate one reformed face. We conduct the comparison with the average value of reforming 500 test images. The results are shown in Table VIII, we can obtain the following observations.

1) *De-GAN* has the minimum model size and FLOPS in one iteration. However, it takes 30× than our method to reform an adversarial face due to its iterative optimization-based defense strategy. It is obviously not suitable for face recognition systems on the edge and mobile device. *Ensemble* has the same shortcoming of time cost.
2) The efficiency of *Mag* is comparable to our methods due to the same model architecture, but our training method has better defense performance.
3) All metrics of *Manifold* and *Shield* are much higher than our method. Such high consumption is not fit for the requirements of face recognition systems in IoT environments.

VI. VISUALIZATION OF REFORMED IMAGES

As shown in Fig. 6, our reformed face images are clearer for face recognition than baselines. The reformed face images of *Mag* and *Ensemble* are blurred. The reformed face images of *De-GAN* and *Manifold* are severely deformed. The faces of *Shield* still have some perturbation pixels. Compared with the normal images, our reformed faces have patch blocks that may reduce the smoothness of the images. We consider such results are caused by the random cropping operations of patch-wise contrastive learning. During patch-wise contrastive learning,

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

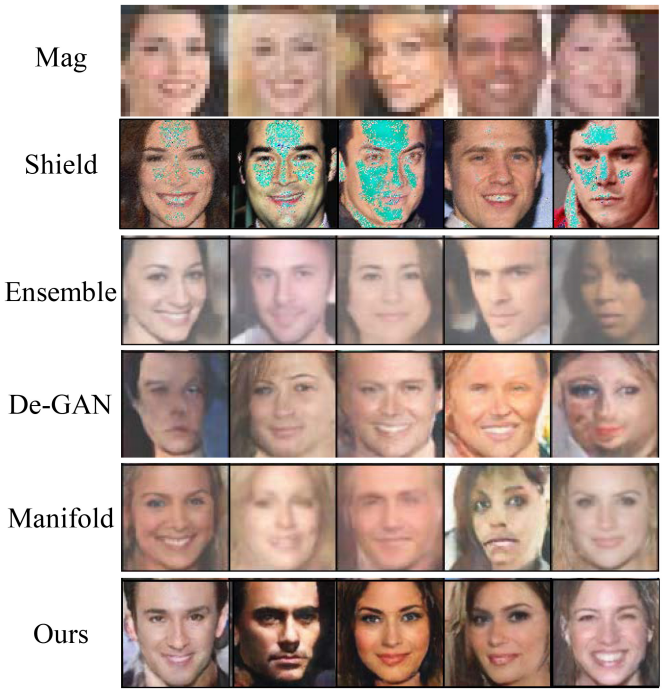


Fig. 6. Visualization of our reformed face images. Compared with the normal images, our reformed face images have patch blocks that may reduce the smoothness of images.

the local consistency loss will force the generator to generate the image with patches, which will reduce the smoothness of the image. This may be the sacrifice for applying local consistency. In future work, we can alleviate the defect of local consistency loss by introducing the constraints for patch-wise edge smoothing or utilizing well-trained super-resolution tools [49] to alleviate the defect of local consistency loss.

VII. Conclusion

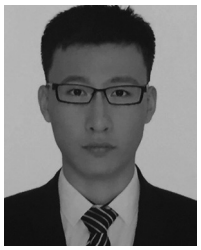
In this article, we improve the robustness of DL-based face recognition system in IoT environments with insights of local consistency. By introducing the constraint of local consistency and other novel designs, we propose LC-GAN to achieve a generative adversarial defense method with the robust encoding ability for defending against adversarial attacks. LC-GAN has a refinement stage with the expert network for extracting the remaining clean identity features on nonsalient areas and a multilevel identity loss to enhance the identity preservation ability. The experimental results demonstrate that LC-GAN can restore higher accuracy than baseline methods under adversarial attacks. The robustness improvement of LC-GAN demonstrates that designing adversarial defense methods with the obtained differences by comparing the characteristics between adversarial examples and clean images is an effective and interpretable solution. In future work, we can consider more characteristics of adversarial examples, such as the unnatural adjacent pixel changes, the mixed patterns of other categories, and the strange activation route of neurons in the deep learning model, to further improve the robustness of LC-GAN.

References

[1] O. M. Parkhi et al., "Deep face recognition," in *Proc. BMVC*, vol. 1, 2015, p. 6.
[2] Q. Cao, L. Shen, W. Xie, O. M. Parkhi, and A. Zisserman, "VGGFace2: A dataset for recognising faces across pose and age," in *Proc. 13th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, 2018, pp. 67–74.
[3] K. Zhao, J. Xu, and M.-M. Cheng, "RegularFace: Deep face recognition via exclusive regularization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 1136–1144.
[4] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "ArcFace: Additive angular margin loss for deep face recognition," in *Proc. CVPR*, 2019, pp. 4690–4699.
[5] S. Yang, Y. Wen, L. He, and M. Zhou, "Sparse common feature representation for undersampled face recognition," *IEEE Internet Things J.*, vol. 8, no. 7, pp. 5607–5618, Apr. 2021.
[6] S. Yang, Y. Wen, L. He, M. Zhou, and A. Abusorrah, "Sparse individual low-rank component representation for face recognition in the IoT-based system," *IEEE Internet Things J.*, vol. 8, no. 24, pp. 17320–17332, Dec. 2021.
[7] Z. Ou, Y. Hu, M. Song, Z. Yan, and P. Hui, "Redundancy removing aggregation network with distance calibration for video face recognition," *IEEE Internet Things J.*, vol. 8, no. 9, pp. 7279–7287, May 2021.
[8] A. Singh and B. Sikdar, "Adversarial attack and defence strategies for deep-learning-based IoT device classification techniques," *IEEE Internet Things J.*, vol. 9, no. 4, pp. 2602–2613, Feb. 2022.
[9] T. Bai, J. Luo, and J. Zhao, "Inconspicuous adversarial patches for fooling image-recognition systems on mobile devices," *IEEE Internet Things J.*, vol. 9, no. 12, pp. 9515–9524, Feb. 2022.
[10] Z. Bao, Y. Lin, S. Zhang, Z. Li, and S. Mao, "Threat of adversarial attacks on DL-based IoT device identification," *IEEE Internet Things J.*, vol. 9, no. 11, pp. 9012–9024, Jun. 2022.
[11] Y. Li, X. Xu, J. Xiao, S. Li, and H. T. Shen, "Adaptive square attack: Fooling autonomous cars with adversarial traffic signs," *IEEE Internet Things J.*, vol. 8, no. 8, pp. 6337–6347, Apr. 2021.
[12] J. Tian, B. Wang, R. Guo, Z. Wang, K. Cao, and X. Wang, "Adversarial attacks and defenses for deep learning-based unmanned aerial vehicles," *IEEE Internet Things J.*, vol. 9, no. 22, pp. 22399–22409, Nov. 2022.
[13] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2016, pp. 1528–1540.
[14] M. Sharif, S. Bhagavatula, L. Bauer, and M. K. Reiter, "A general framework for adversarial examples with objectives," *ACM Trans. Privacy Security*, vol. 22, no. 3, p. 16, 2019.
[15] Y. Dong et al., "Efficient decision-based black-box adversarial attacks on face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 7714–7722.
[16] A. Dabouei, S. Soleymani, J. M. Dawson, and N. M. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *Proc. IEEE Winter Conf. Appl. Comput. Vis. (WACV)*, 2019, pp. 1979–1988.
[17] Y. Zhong and W. Deng, "Towards transferable adversarial attack against deep face recognition," *IEEE Trans. Inf. Forensics Security*, vol. 16, pp. 1452–1466, 2021.
[18] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2014, *arXiv:1412.6572*.
[19] A. Kurakin, D. Boneh, F. Tramèr, I. J. Goodfellow, N. Papernot, and P. McDaniel, "Ensemble adversarial training: Attacks and defenses," in *Proc. ICLR*, 2018, pp. 1–6.
[20] D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at ODDs with accuracy," 2018, *arXiv:1805.12152*.
[21] P. Du, X. Zheng, L. Liu, and H. Ma, "Defending against universal attack via curvature-aware category adversarial training," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2022, pp. 2470–2474.
[22] P. Du, X. Zheng, M. Qi, L. Liu, and H. Ma, "Towards adversarial robust representation through adversarial contrastive decoupling," in *Proc. IEEE Int. Conf. Multimedia Expo*, 2022, pp. 1–6.
[23] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2016, pp. 582–597.
[24] Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman, "PixelDefend: Leveraging generative models to understand and defend against adversarial examples," 2017, *arXiv:1710.10766*.
[25] P. Samangouei, M. Kabkab, and R. Chellappa, "Defense-GAN: Protecting classifiers against adversarial attacks using generative models," 2018, *arXiv:1805.06605*.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.

[26] D. Meng and H. Chen, "MAGNET: A two-pronged defense against adversarial examples," in *Proc. ACM SIGSAC Conf. Comput. Commun. Security*, 2017, pp. 135–147.
[27] R. Theagarajan, M. Chen, B. Bhanu, and J. Zhang, "ShieldNets: Defending against adversarial attacks using probabilistic adversarial robustness," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6988–6996.
[28] J. Yuan and Z. He, "Ensemble generative cleaning with feedback loops for defending adversarial attacks," in *Proc. CVPR*, 2020, pp. 578–587.
[29] J. Zhou, C. Liang, and J. Chen, "Manifold projection for adversarial defense on face recognition," in *Proc. ECCV*, vol. 12375, 2020, pp. 288–305.
[30] D. Yi, Z. Lei, S. Liao, and S. Z. Li. "Learning face representation from scratch." 2014. [Online]. Available: http://arxiv.org/abs/1411.7923
[31] A. van den Oord, N. Kalchbrenner, and K. Kavukcuoglu, "Pixel recurrent neural networks," in *Proc. Int. Conf. Mach. Learn.*, vol. 48, 2016, pp. 1747–1756.
[32] I. J. Goodfellow et al., "Generative adversarial nets," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.
[33] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *Proc. ICLR*, 2018, pp. 1–9.
[34] C. Szegedy et al., "Intriguing properties of neural networks," in *Proc. 2nd Int. Conf. Learn. Represent. (ICLR)*, Banff, AB, Canada, Apr. 2014. [Online]. Available: http://arxiv.org/abs/1312.6199
[35] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *Proc. IEEE Symp. Security Privacy (SP)*, 2017, pp. 39–57.
[36] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "DeepFool: A simple and accurate method to fool deep neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 2574–2582.
[37] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, "Universal adversarial perturbations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1765–1773.
[38] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," in *Proc. ICLR*, 2014, pp. 1–8.
[39] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proc. CVPR*, 2019, pp. 4401–4410.
[40] Y. Qian, W. Deng, and J. Hu, "Unsupervised face normalization with extreme pose and expression in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 9851–9858.
[41] J. Johnson, A. Alahi, and L. Fei-Fei, "Perceptual losses for real-time style transfer and super-resolution," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 694–711.
[42] C. Guo, M. Rana, M. Cissé, and L. van der Maaten, "Countering adversarial images using input transformations," in *Proc. ICLR*, 2018, pp. 1–9.
[43] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. L. Yuille, "Mitigating adversarial effects through randomization," in *Proc. ICLR*, 2018, pp. 1–9.
[44] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," in *Proc. Workshop Faces Real-Life Images Detection Alignment Recognit.*, 2008.
[45] S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou, "AGEDB: The first manually collected, in-the-wild age database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshop*, vol. 2, 2017, p. 5.
[46] S. Sengupta, J.-C. Chen, C. Castillo, V. M. Patel, R. Chellappa, and D. W. Jacobs, "Frontal to profile face verification in the wild," in *Proc. IEEE Conf. Appl. Comput. Vis.*, Feb. 2016, pp. 1–9.
[47] A. Kurakin, I. J. Goodfellow, and S. Bengio, "Adversarial examples in the physical world," 2016, *arXiv:1607.02533*.
[48] S. Chen, Y. Liu, X. Gao, and Z. Han, "MobileFaceNets: Efficient CNNs for accurate real-time face verification on mobile devices," in *Proc. CCBR*, vol. 10996, 2018, pp. 428–438.
[49] K. Prajapati et al., "Unsupervised single image super-resolution network (USISResNet) for real-world data using generative adversarial network," in *Proc. Conf. Comput. Vis. Pattern Recognit. Workshops*, 2020, pp. 1904–1913.



Peilun Du received the B.E. degree from Beijing University of Posts and Telecommunications, Beijing, China, in 2017, where he is currently pursuing the Ph.D. degree.

His research interests include adversarial attack and defense.

Xiaolong Zheng (Member, IEEE) received the B.E. degree from Dalian University of Technology, Dalian, China, in 2011, and the Ph.D. degree from Hong Kong University of Science and Technology, Hong Kong, China, in 2015.

He is currently an Associate Professor with the State Key Laboratory of Networking and Switching Technology and the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China. His research interests include Internet of Things, wireless networks, and ubiquitous computing.

Liang Liu (Member, IEEE) received the B.S. degree from the Department of Computer Science and Technology, South China University of Technology, Guangzhou, China, in 2004, and the Ph.D. degree from the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing, China, in 2009.

He is a Professor with the State Key Laboratory of Networking and Switching Technology and the School of Artificial Intelligence, and the Dean of the School of Artificial Intelligence, Beijing University of Posts and Telecommunications. He was a visiting Ph.D. student with the Networking and Information Systems Laboratory, Texas A&M University at College Station, College Station, TX, USA, from 2007 to 2008. He has published over 170 papers. His current research interests include Internet of Things and intelligent sensing technologies.

Huadong Ma (Fellow, IEEE) received the B.S. degree in mathematics from Henan Normal University, Xinxiang, China, in 1984, the M.S. degree in computer science from Shenyang Institute of Computing Technology, Chinese Academy of Science, Shenyang, China, in 1990, and the Ph.D. degree in computer science from the Institute of Computing Technology, Chinese Academy of Science, Beijing, China, in 1995.

From 1999 to 2000, he held a visiting position with the University of Michigan at Ann Arbor, Ann Arbor, MI, USA. He is currently a Professor with the School of Computer Science, Beijing University of Posts and Telecommunications, Beijing. His current research interests include Internet of Things and sensor networks, and multimedia computing, and he has published more than 400 papers in prestigious journals, such as ACM/IEEE transactions or conferences, such as ACM SIGCOMM, ACM MobiCom/MM, and IEEE INFOCOM and five books.

Dr. Ma received the First Class Prize of the Natural Science Award of the Ministry of Education, China, in 2017. He received the 2019 Prize Paper Award of IEEE TRANSACTIONS ON MULTIMEDIA, the 2018 Best Paper Award from IEEE MULTIMEDIA, the Best Paper Award in IEEE ICPADS 2010, and the Best Student Paper Award in IEEE ICME 2016 for his coauthored papers. He received the national funds for Distinguished Young Scientists in 2009. He was/is an Editorial Board Member of the IEEE TRANSACTIONS ON MULTIMEDIA, IEEE INTERNET OF THINGS JOURNAL, *ACM Transactions on Internet of Things*, and *Multimedia Tools and Applications*. He serves as the Chair of ACM SIGMOBILE China.

Authorized licensed use limited to: Auckland University of Technology. Downloaded on May 06,2024 at 07:27:43 UTC from IEEE Xplore. Restrictions apply.