Two-Stage Image Restoration: A Shield against Adversarial Attacks on Face Identification

Min Young Lim, Pyo Min Hong, and Youn Kyu Lee*
Department of Computer Engineering
Hongik University
Seoul, Republic of Korea
minyounge17@g.hongik.ac.kr, pyomindl@g.hongik.ac.kr, younkyul@hongik.ac.kr

***Abstract*—Various methods have been proposed to defend face identification systems against adversarial attacks by eliminating adversarial perturbations from target face images. However, it is challenging to eliminate such perturbations while preserving the crucial facial features within the images. In this paper, we propose a novel purification method for an effective defense against adversarial attacks on target face images. Our method incorporates a two-stage image restoration utilizing diffusion, consisting of sequential super-resolution-based image restoration followed by colorization-based image restoration. The experimental results demonstrate the effectiveness of our method in eliminating perturbations while preserving the identity of the facial features.**

***Index Terms*—defense against adversarial attacks, deep learning, face identification, image restoration**

I. Introduction

With the development of deep learning techniques, deep learning-based face identification systems have become widely used as biometric authentication methods. However, these deep learning models are inherently vulnerable to adversarial attacks, which can lead to misclassification through the utilization of adversarial examples containing small amounts of perturbations. In particular, adversarial attacks on face identification systems can result in user misclassification, leading to security threats such as personal information leaks and banking fraud.

To address this problem, several defense methods against adversarial attacks have been proposed, aiming to eliminate perturbations added to adversarial examples. Existing methods achieved this by employing image restoration techniques such as denoising or super-resolution, which convert low-resolution images into high-resolution ones [1]. However, the task of eliminating perturbations while preserving the facial features of images, known as purification, continues to pose a significant challenge [2]. Specifically, when the intensity of image restoration is high, it effectively eliminates most perturbations, but it also poses challenges in preserving facial features, leading to higher false rejection rates [3]. On the contrary, when the intensity of image restoration is low, the defense performance against adversarial attacks can be ineffective due to the persistence of remaining perturbations [2]. Therefore, it is crucial to develop an effective purification method that can eliminate perturbations without distorting facial features.

*Corresponding Author

In this paper, we present a novel purification method that effectively enhances defense against adversarial attacks in face identification systems. Our proposed method comprises two main stages: (Stage#1) super-resolution-based image restoration and (Stage#2) colorization-based image restoration. Specifically, in Stage#1, our method performs super-resolution while ensuring the preservation of facial features in the images. During this process, the intensity of super-resolution is adjusted based on the ratio between the target resolution and the resolution of the input image. In Stage#2, our method transforms the restored image from Stage#1 into a grayscale and subsequently colorizes it. During this process, the remaining perturbations from Stage#1 are eliminated by reconstructing the RGB (i.e., Red, Green, and Blue) values for each pixel. The image restoration in each stage is performed based on the diffusion model, which is known for its high performance in image generation.

The contributions of this paper are as follows: (1) Designing a novel two-stage adversarial attack defense mechanism that supports the effective purification of face images; (2) Proposing and validating a super-resolution&colorization-based perturbation elimination method that minimizes distortions of facial features; (3) Implementing a prototype and conducting a systematic evaluation of the proposed mechanism using the diffusion model.

The paper is organized as follows. Section 2 discusses related work, Section 3 describes our proposed method, Section 4 presents our evaluations, and Section 5 concludes the paper.

II. Related Work

Liu et al. [4] proposed a feature distillation method utilizing JPEG compression to eliminate perturbations. However, it does not support identity preservation for clean images, where perturbations are not added. Deb et al. [3] introduced a self-supervised defense method that automatically detects adversarial faces within input images and eliminates perturbations in the detected regions. While this approach minimizes the distortion of clean images, it still has the limitation of effectively eliminating perturbations in undetected regions. Consequently, achieving precise elimination of perturbations while preserving the identity of clean images remains a challenging task.

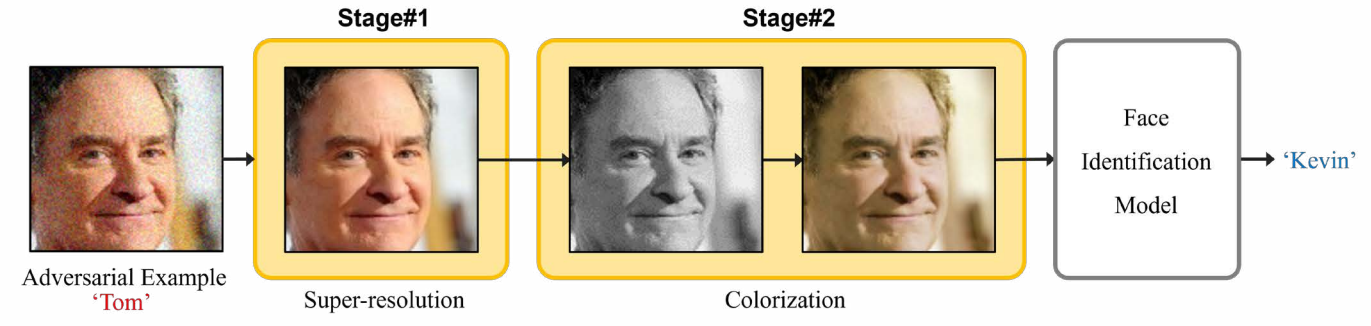

Fig. 1. An Overview of Our Proposed Method

III. Our Approach

In this paper, we propose a novel defense method against adversarial attacks for purifying face images. As shown in Fig. 1, our proposed method performs a two-stage image restoration. In Stage#1, our method focuses on conducting super-resolution while preserving facial features of the images to the greatest extent possible. In Stage#2, our method involves a process of transforming the image to grayscale and subsequently colorizing it to eliminate any remaining perturbations that may have persisted after Stage#1. With the implementation of our two-stage image restoration, our method enables the effective elimination of perturbations. For example, both the adversarial example that disturbs the target model, causing it to misclassify 'Kevin' as 'Tom' and the clean image of 'Kevin' can be correctly identified as 'Kevin' through our two-stage image restoration. The details of our proposed method are as follows.

A. Stage#1: Super-resolution-based Image Restoration

Our method employs super-resolution-based image restoration, which transforms low-resolution images into high-resolution images. During this process, each pixel value is reconstructed, resulting in the elimination of any perturbations. To ensure the maximum preservation of facial features, our method employs a resolution scale that represents the ratio of resolution between the input and restored images. The selection of the optimal resolution scale, which preserves the facial features, involves two steps: (1) Evaluating the extent of perturbation elimination and identity preservation by monitoring face identification indicators at different resolution scale levels (e.g., 2, 4, and 8) for face image samples, which include both adversarial examples and their corresponding clean images, obtained from the target environment; (2) Performing a visual inspection of the restored images to determine whether the facial features exhibit any distortions (e.g., changes in the shape of the eyes, nose, mouth and presence of artifacts) at each resolution scale.

B. Stage#2: Colorization-based Image Restoration

Our method incorporates colorization-based image restoration to eliminate any remaining perturbations, which involves transforming the image to grayscale and subsequently colorizing it. Fig. 2 illustrates the accuracy of identity prediction at each step of the colorization-based image restoration, using

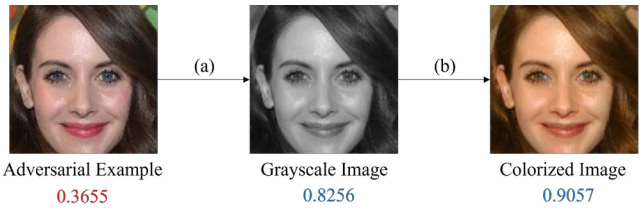

Fig. 2. A Process of Image Restoration based on Colorization

a sample image from the CASIA-WebFace dataset [5]. The elimination of the remaining perturbation through colorization is achieved as follows: (a) During the conversion of the image to grayscale, the perturbation represented in the RGB channel is consolidated into a single channel, resulting in the elimination of perturbations. Consequently, the accuracy of identity prediction improves from 0.3655 to 0.8256; (b) During the conversion of the grayscale image to a colorized image, the RGB values for each pixel are reconstructed, leading to the elimination of perturbations. This further improves the accuracy of identity prediction improved from 0.8256 to 0.9057. Colorization-based image restoration enables the elimination of perturbations while preserving the facial features of images to the greatest extent possible by reconstructing only the RGB values for each pixel.

IV. Evaluation

To assess the effectiveness of our proposed defense method against adversarial attacks, we conducted evaluations addressing the following research questions.

- **RQ#1**: How well does our method eliminate the perturbation of adversarial examples?
- **RQ#2**: How well does our method preserve the identity of the clean images?
- **RQ#3**: How well does our two-stage-based method defend against adversarial attacks compared to the individual method?

A. Experimental Setting

- **Dataset**: We selected the CASIA-WebFace as our target face identification dataset. Due to the significant variation in the number of images per identity in the CASIA-WebFace (ranging from 2 to 802 images), which may have a negative impact on the performance of face recognition models [6], we conducted the following procedure

TABLE I
THE PERFORMANCE OF OUR DEFENSE METHOD

Clean Image	ADV Example	Clean Image + Our Method	ADV Example + Our Method
93.03%	36.40%	90.78%	69.27%

to select the target dataset from CASIA-WebFace: (a) selecting identities containing a minimum of 44 images; (b) randomly selecting 1,000 identities; (c) randomly selecting 40 training images and 4 testing images per identity. As a result, our selected dataset consists of 44,000 face images (train: 40,000, test: 4,000) from 1,000 identities.

- **Face Identification**: We selected the FaceNet model [7], one of the popular face identification models, to evaluate the performance of face identification. We trained FaceNet model using our train dataset, which consisted of 40,000 images.
- **Adversarial Attack**: We selected PGD [8], one of the popular adversarial attack methods, to generate adversarial examples for the test images. We evaluated whether the perturbations added to adversarial examples were effectively eliminated using our proposed method.
- **Super-resolution & Colorization**: We selected DDNM (Denoising Diffusion Null-Space Model) [9], which is a state-of-the-art model capable of performing both super-resolution and colorization. The resolution scale value was set to 2 in our experiments.

Our experiments utilized one GPU (NVIDIA GeForce RTX 3090) and the following hyperparameters: (1) FaceNet: Python 3.9.13, PyTorch 1.7.1+cu110, 64 batch size, and 30 epochs; (2) PGD: eps 0.5/255, alpha 2/255, iters 40; (3) DDNM: Python 3.8.10, PyTorch 1.12.1+cu116, eta 0.85, sigma_y 0.

B. Experimental Results

(RQ#1) Effectiveness of our defense method for adversarial examples: We evaluated the performance of face identification on the adversarial examples using our method. Table I shows the performance of face identification for the clean images (93.03%), adversarial examples (36.40%), and adversarial examples restored by our method (69.27%). Our method achieved a significant improvement of 32.87% in face identification accuracy on the adversarial examples. The result highlights the effectiveness of our method in defending against adversarial attacks on face identification.

(RQ#2) Effectiveness of our defense method for clean images: We evaluated the performance of face identification on the clean images using our method. Table I presents the performance of face identification on clean images (93.03%) and clean images restored by our method (90.78%). Our method exhibited a minor decrease of 2.25% in face identification accuracy on the clean images. However, considering the improved performance on adversarial examples, our method effectively defends against adversarial attacks while preserving the identity in target images to the greatest extent possible.

(RQ#3) Effectiveness of our defense method compared to the individual method: We evaluated the effectiveness of our two-stage-based method by conducting an ablation study, comparing it to the individual method. Table II presents the performance of face identification on adversarial examples using only Stage#1 (62.12%), only Stage#2 (68.37%), and the combination of Stage#1 and Stage#2, which represents our method (69.27%). Our method exhibited higher face identification accuracy compared to the individual method. The result demonstrates the effectiveness of our method in defending against adversarial attacks on face identification by combining the individual methods of super-resolution and colorization.

TABLE II
THE ABLATION STUDY OF IMAGE RESTORATION

w/o Stage#2	w/o Stage#1	Our Method (=Two-Stage)
62.12%	68.37%	69.27%

V. CONCLUSION

In this paper, we propose a novel purification method for effective defense against adversarial attacks on face images. Our proposed method protects face identification systems from adversarial attacks through a two-stage image restoration, which sequentially performs super-resolution and colorization-based on the diffusion model. Our experimental results demonstrated that our method effectively eliminates perturbations added to adversarial examples while minimizing distortions of facial features. Our future work involves conducting an extensive evaluation that encompasses a wide range of adversarial attacks and face datasets. Moreover, we plan to conduct a comprehensive study on the defense performance by exploring different hyperparameters of the diffusion model.

ACKNOWLEDGEMENT

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2022-00165648).

REFERENCES

[1] Y. Bakhti, S. A. Fezza, W. Hamidouche, and O. Déforges, "Ddsa: A defense against adversarial attacks using deep denoising sparse autoencoder," *IEEE Access*, vol. 7, pp. 160397–160407, 2019.

[2] Z. Niu, Z. Chen, L. Li, Y. Yang, B. Li, and J. Yi, "On the limitations of denoising strategies as adversarial defenses," *arXiv:2012.09384*, 2020.

[3] D. Deb, X. Liu, and A. K. Jain, "Faceguard: A self-supervised defense against adversarial face images," in *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition*, pp. 1–8, IEEE, 2023.

[4] Z. Liu, Q. Liu, T. Liu, N. Xu, X. Lin, Y. Wang, and W. Wen, "Feature distillation: Dnn-oriented jpeg compression against adversarial examples," in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 860–868, IEEE, 2019.

[5] D. Yi, Z. Lei, S. Liao, and S. Z. Li, "Learning face representation from scratch," *arXiv:1411.7923*, 2014.

[6] Y. Zhang and W. Deng, "Class-balanced training for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 824–825, 2020.

[7] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 815–823, 2015.

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

[9] Y. Wang, J. Yu, and J. Zhang, "Zero-shot image restoration using denoising diffusion null-space model," *arXiv:2212.00490*, 2022.