A Novel Image Restoration Technique for Face Adversarial Robustness Improvement

Chiranjeevi Sadu[1], Pradip K. Das[2], Ramanjaneyulu Y[3] and Anand Nayyar[4*]

[1]Department of Computer Science and Engineering, RGUKT Nuzvid, Andhra Pradesh, India.
[2]Department of Computer Science and Engineering, IIT Guwahati, Assam, India.
[3]School of Technology, Woxsen University, Hyderabad, Telangana, India.
[4*]Graduate School, Faculty of Information Technology, Duy Tan University, Da Nang 550000, Viet Nam.

*Corresponding author(s). E-mail(s): anandnayyar@duytan.edu.vn;
Contributing authors: schiranjeevi@rguktn.ac.in; pkdas@iitg.ac.in; ramanjaneyulu.yannam@woxsen.edu.in;

Abstract

Machine Learning (ML) in specific Deep Learning (DL) models have rapid advances and significant accomplishments in numerous applications, including in many safety-critical contexts. However, these models have recently been discovered to be susceptible to adversarial attacks, which are well-crafted input images. Adversarial attacks are invisible to humans but are quite effective at tricking DL models during testing and deployment. We focus on a novel method based on deep image restoration networks that significantly improves facial adversarial robustness of various image-classification models. Adversarial images are created using Private Fast Gradient Sign Method (P-FGSM), StyleGAN and Fast Landmark Manipulation (FLM) methods. The adversarial images are then enhanced using deep image restoration networks to bring back them into the original space. The encoded weighted local magnitude patterns (WLMP) are extracted and provided to different types of classifiers to detect facial adversarial images from the clean images. The

effectiveness of the proposed method has been demonstrated on two real-world datasets and experimental outcomes show that it significantly improves facial adversarial robustness on all evaluating classifiers. It improves the highest classification accuracy from **98.75%** to **99.00%** on P-FGSM attacks, from **77.94%** to **85.25%** on adversarial attacks generated by StyleGAN and from **65.52%** to **69.50%** for FLM attacks.

Keywords: Facial Images, Adversarial Attacks, Image Restoration, Adversarial Robustness, Image Classification.

1 Introduction

The field of DL has seen rapid development, marked by a significant rise in both performance and the range of practical applications over the past two decades. Specifically, the emergence of deep neural networks (DNNs) represented a major leap forward in several applications such as object detection [1], image classification [2], speech recognition [3], natural language processing [4], sentiment analysis [5] and multi-modal [6]. These DNNs, with their diverse architectures, quickly gained immense popularity and showcased exceptional performance that could rival, and in some cases exceed, human capabilities in tasks related to perception and decision-making. As a result, DNNs are increasingly being utilized in safety-critical domains.

However, the majority of existing ML classifiers exhibit significant susceptibility to adversarial examples. An adversarial example is an input data sample that has undergone minute modifications with the specific intent of leading a ML classifier to misclassify it. Often, these alterations are so unnoticeable that they escape the notice of a human observer entirely, yet they are sufficient to cause the classifier to make an erroneous prediction [7, 8].

Adversarial attacks can broadly be classified into two categories: white-box attacks and black-box attacks. In a white-box attack on a ML model, the adversary possesses complete knowledge about the model used for classification. This knowledge includes details about the model architecture like the type of neural network, the number of layers, etc. Additionally, the attacker is aware of the algorithm (e.g., gradient-descent optimization) used during the training process and has access to information about the distribution of the training data. Furthermore, the attacker has knowledge of the model's parameters after it has been fully trained. In contrast to a white-box attack, a black-box attack assumes no prior knowledge about the model being targeted. Instead, it relies on information about the model's settings and previous inputs to exploit the model's vulnerabilities. For instance, the adversary probes the model by submitting a series of meticulously crafted inputs and then observes the outputs produced by the model. This iterative process allows the attacker to gradually learn about the model's behavior and identify its weaknesses without having any direct access to the model's architecture, parameters, or internal workings.

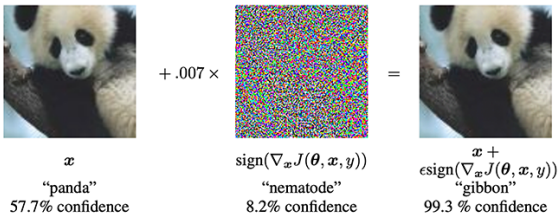

Fig. 1 Shows the adversarial image generation by FGSM.

It's a more challenging and resource-intensive approach compared to white-box attacks, as the attacker needs to experiment and gather information iteratively through interactions with the model.

Multiple techniques have been introduced for generating adversarial attacks [8–14]. Fig. 1 shows procedure for generating FGSM attack [8]. For instance, Fig. 2 depicts examples of original facial images and corresponding adversarial images generated by StyleGAN[15], FLM [16] and P-FGSM [14], respectively. Adversarial images are specifically designed to closely resemble the original images, leading to misclassification by the classifier. Adversarial examples present significant security concerns because they can potentially be employed to launch attacks on ML systems, even when the attacker has no access to the underlying model. Furthermore, it has uncovered that it's possible to execute adversarial attacks on ML systems operating in the physical world, where inputs are received through imperfect sensors rather than precise digital data [11].

In the long-term, as ML and Artificial Intellengence (AI) systems continue to advance in power and capability, the sensitivity of DL models to adversarial images can pose significant challenges and raise concerns, particularly in security and safety-critical applications. It could be leveraged to compromise and gain control over highly potent AI systems. For instance, when adversarial attack is applied to the DNN model, which is part of an autonomous vehicle, the model reads the present scene differently and a tragic accident may result. Despite several defense techniques have been proposed to avoid misclassification by the classifier [17–20], many of these defenses are not effective against various and more powerful adversarial attacks [7, 21]. Thus, the existence of such adversarial attacks shows frailties in ML model and ensuring robustness against adversarial examples becomes a crucial aspect of addressing AI safety concerns.

Research on adversarial attacks and defenses poses several challenges, one of which is the complexity of evaluation. Unlike traditional ML, where evaluation is straightforward by measuring the loss on a test set drawn independently and identically distributed from the training set, adversarial ML presents a more complicate problem. In adversarial settings, defenders face an open-ended challenge where attackers can send inputs from an unknown distribution. It's not enough to assess a defense against a single predefined attack or even a set of attacks prepared in advance by the researcher proposing the defense. Even

Fig. 2 Examples of original image, StyleGAN attack, FLM attack, and P-FGSM attack in column-wise.

if a defense performs well in such experiments, it might be vulnerable to a new attack strategy that the defender did not anticipate. Ideally, a defense should be theoretically proven to be sound. However, ML in general, and DNNs in particular, are challenging to analyze theoretically due to their complexity.

Detection-based defense methods have recently gained significant attention as potential alternative solutions to counter adversarial attacks. In [22], ML models are enhanced with an additional class dedicated to classifying adversarial images. Metzen et al. [23] proposed a method to train neural networks (NN) specifically for recognizing and classifying adversarial attacks. Another technique [24] involves training an extra classifier to identify the presence of adversarial attacks. In [25], the input image dimensions are reduced and fed into a classifier, which is then trained using a small dataset and a fully-connected NN. Furthermore, in [26], a cascaded classifier is created, with each individual classifier implemented as a linear SVM on the Principal Component Analysis (PCA) for effective detection of adversarial examples among the original images. However, many defense methods face challenges when applied to real-world datasets, resulting in lower accuracy. As adversarial attacks can be highly sophisticated and adaptive, it becomes difficult to develop robust defense mechanisms that can consistently detect and mitigate these attacks.

In our work, we introduce a novel defense method that utilizes deep image restoration networks to enhance the robustness of ML classifiers against adversarial attacks. This defense mechanism is an extension of our previous work [27]. First, we generate facial adversarial attacks by employing StyleGAN [15],

FLM [16] and P-FGSM [14] techniques. Next, the crafted adversarial images are enhanced using deep image restoration networks to bring back into the original space. We then extract the encoded WLMP features from the input image and utilize them as input for various classifiers, including different types of SVMs, RF, and k-NN. These classifiers are used to evaluate the performance of our proposed defense method. Through comprehensive experiments, we demonstrate the effectiveness of our defense approach in discriminating adversarial images from real ones. The experimental results show that our defense method significantly improves the performance of various classification models.

The objectives of the paper are:

- To study the background and enlighten literature review with regard to the adversarial attacks, defense methods and improve adversarial robustness of ML classifiers.
- To propose Novel Methodology based on deep image restoration to solve the adversarial robustness of ML classifiers. Deep image restoration networks to restore the crafted adversarial images back to their original space. Then, the encoded features extracted to distinguish adversarial images from the original.
- To test and validate the proposed methodology on performance metrics like precision, recall, F1-score and accuracy. Such metrics are validated before and after the applying image restoration techniques.
- To compare the proposed methodology with existing techniques- various types of SVMs, RF and k-NN.

The remaining sections of the paper are organized as follows: Section II discusses related works to adversarial attacks and defense methods. Our proposed method is detailed in Section III. We present experimental results in Section IV and conclude the paper in Section V.

2 Literature Survey

In the field of face manipulation and detection, numerous approaches have been proposed over the years. However, it is observed that not all of these approaches for creating fake images or image attacks, as well as their detection methods, are suitable for effectively detecting facial adversarial attacks. We focus on presenting only state-of-the-art works that specifically address the generation of adversarial images and its detection in order to provide a comprehensive understanding of the current advancements in the field and highlight the most relevant and effective approaches for detecting facial adversarial attacks.

2.1 Adversarial Attacks Generation

Biggio et al. [28] proposed an initial work for generating adversarial examples. It is straightforward yet potent gradient-based method that can be systematically applied to assess the vulnerability of numerous widely-used classification

algorithms to evasion attacks. They targeted conventional machine learning classifiers like SVM and a fully-connected three-layer neural network. The proposed method evaluated by generating adversarial examples on the MNIST dataset [**?**].

The first work to attack Deep Neural Networks (DNNs) was presented by Szegedy et al. [7]. They formulated an optimized approach for generating minimal distorted adversarial examples. They found that DNNs often acquire input-output mappings that exhibit a significant degree of discontinuity. This means that even small, hardly noticeable perturbations to an input can lead the network to misclassify the image. These perturbations are discovered by maximizing the prediction error of the network, and they illustrate the networks' sensitivity to subtle changes in input data. They performed experiments on MNIST [**?**], ImageNet [**?**] and image samples from YouTube.

Goodfellow et al. [8] proposed introduced a one-step method called the Fast Gradient Sign Method (FGSM) for generating adversarial examples efficiently. FGSM computes the gradient of the model's loss with respect to the input and uses it to perturb the input in a way that maximizes the loss. They demonstrated that it is possible to generate these adversarial examples systematically, even when the perturbations are nearly undetectable by humans and showed that DNNs are highly vulnerable to adversarial examples. The authors also explored the concept of transferability, where adversarial examples generated for one model can often be used to deceive other models, even those with different architectures. It primarily focused on generating adversarial examples, it also briefly discussed potential defense mechanisms based on adversarial training, which involve augmenting the training dataset with adversarial examples to improve model robustness. They performed experiments on MNIST [**?**].

Kurakin et al. [11] extended the FGSM to a more advanced technique known as the Basic Iterative Method (BIM). Unlike the original FGSM, which applies a single perturbation to an input, the BIM applies multiple perturbations iteratively. These perturbations are computed based on the gradients of the loss with respect to the input at each iteration. Each iteration adds a small perturbation to the input, gradually increasing its deviation from the original, legitimate input. The BIM allows for fine-tuning of parameters, such as the step size and the number of iterations. These parameters can be adjusted to control the trade-off between the strength of the adversarial perturbations and the perceptibility of the resulting adversarial examples. They performed experiments on samples of ImageNet [**?**].

Moosavi et al. [29] examined the decision boundary of a classifier around a specific data point, aiming to find a path for that data point which leads to a different prediction by the classifier. This method, known as DeepFool, identifies the minimum perturbation required to cross the decision boundary. DeepFool typically employs an iterative approach. It starts with the original data point and iteratively adjusts it while considering the classifier's output at each step. The perturbation is gradually refined to reach the minimum necessary to cause a different prediction. The primary aim of the DeepFool is to

determine the smallest possible perturbation that can be applied to the input data point to shift it across the decision boundary, resulting in a change in the classifier's prediction. They tested the proposed method on various DNNs applied to CIFAR-10 [**?**] and MNIST [**?**].

In [30], the auhtors proposed a technique called the Jacobian-based Saliency Map Attack (JSMA), which crafts adversarial examples by exploiting the saliency information derived from the Jacobian matrix of the model. JSMA relies on the Jacobian matrix, which is a matrix of partial derivatives that describes how small changes in input features affect the model's output. Specifically, it focuses on the derivatives of the model's predicted class probabilities with respect to the input features. It is often used for targeted adversarial attacks. It aims to generate adversarial examples that are misclassified into specific target classes chosen by the attacker. JSMA typically employs an iterative approach to craft adversarial examples. It starts with the original input and iteratively perturbs the most salient features to maximize the likelihood of the target misclassification. The proposed algorithm is validated on MNIST [**?**].

When conducting black-box attacks, the attacker is unaware of the classifier's settings and training data. Only the model's input data and the accompanying outputs may be seen by the attacker. Adversaries can find flaws in the model and use them to launch attacks based on the input-output connection. The first successful method for taking down DNN classifiers in a black-box environment was developed by Papernot et al. [13]. The classifier's parameters or training dataset are unknown to the attacker. An optimization-based approach is proposed for black-box attacks in [31]. The attacker assumes no access to the prediction confidence of the classifier. Ilyas et al. [32] proposed a method to estimate gradient information and leveraged to generate adversarial examples in a black-box setting. A genetic algorithm to generate adversarial examples utilized by Alzantot et al. [33]. It evolves the input data to find perturbations that can deceive the target model.

In semi white-box attacks, a generative model is first trained in a white-box manner to create adversarial samples. Once the generative model is trained, the attacker can use it to generate adversarial samples in a black-box manner. This approach allows the attacker to leverage the power of generative models to craft adversarial examples. The work by Xiao et al. [34] introduced a semi white-box adversarial attack model. The authors trained a Generative Adversarial Network (GAN) [35] to target a specific model. Adversarial samples are then generated directly from the trained generative model. Deng et al. [36] proposed the Additive Angular Margin Loss (ArcFace), a Convolutional Neural Network (CNN)-based method that maximizes face image classification accuracy. Advfaces [37] utilized GAN to craft adversarial face images with minimal perturbations in salient facial regions. SemanticAdv [38] proposed a method based on attribute-conditioned image editing to generate adversarial samples that appear semantically realistic. For a detailed review on adversarial attacks

and defenses, Xu et al. [39] provide an extensive overview of various techniques and approaches in the field.

2.2 Adversarial Attacks Detection

One of the commonly used strategies to protect classifiers from misclassification is to detect adversarial attacks from the original inputs. Rather than directly predicting the class label for the input, these methods first determine whether the input is an original or an adversarial example. If the classifier detects that the input is adversarial, it refrains from making a class prediction. This approach aims to effectively distinguish between adversarial images and original ones, thereby reducing the chances of misclassification by the classifier. Adversarial attack detection algorithms play a crucial role in this process.

Su et al. [**?**] introduced a novel technique for creating one-pixel adversarial perturbations, leveraging differential evolution (DE). This approach requires minimal adversarial information, making it a black-box attack, and possesses the ability to deceive a broader range of neural network architectures. Experimental findings revealed that a significant portion of natural images in datasets like Kaggle CIFAR-10 and ImageNet (ILSVRC 2012) can be manipulated to be misclassified into at least one target class by modifying just a single pixel. Furthermore, these manipulated images achieved relatively high confidence scores in their misclassification, with an average confidence of 74.03% for CIFAR-10 and 22.91% for ImageNet.

Many approaches have been proposed to distinguish between adversarial examples and original inputs. Here are some notable methods. Grosse et al. [22] proposed an auxiliary model to differentiate adversarial examples from original inputs. Gong et al. [24] trained a binary classifier to separate adversarial examples and then trained the classifier on the detected original examples. Metzen et al. [23] introduced an auxiliary neural network-based method for adversarial example detection. Hendrycks et al. [40] proposed a statistical method based on PCA, where original images tend to have higher weights on early principal components, while adversarial images have larger weights on latter principal components. Massoli et al. [41] presented a detection method that combines a deep learning model with k-NN classifier. Agarwal et al. [42] developed a method based on intensity values of pixels and PCA as features for detecting universal perturbations, using SVM classifier. Xie et al. [43] improved adversarial robustness by incorporating feature denoising blocks in neural networks. Mustafa et al. [44] proposed a defense method based on image super-resolution to enhance adversarial example detection. Recently, Sadu et al. [27] proposed a defense method that generates adversarial attacks using P-FGSM and provides the extracted features to different types of SVM classifiers for detection.

We propose a new defense method against facial adversarial attacks for improving adversarial robustness, which is based on deep restoration networks and feature denoising. Image restoration enhances adversarial images and improve their similarity to original images. Our method is different from the state-of-the-art methods. First, adversarial images are generated based on

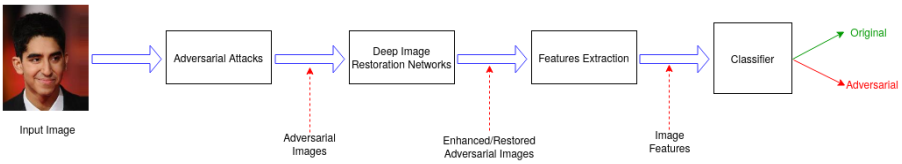

Fig. 3 The overall procedure of the proposed defense method.

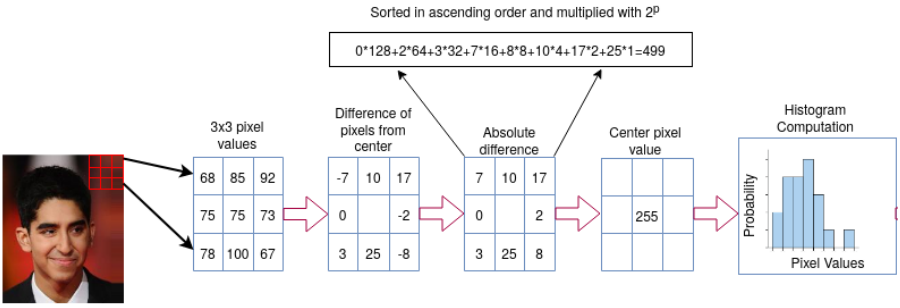

Fig. 4 The procedure for extraction of WLMP features

P-FGSM, StyleGAN and FLM. These adversarial images are designed to protect the image scenes and effectively mislead classifiers into high-confidence misclassifications. To restore the adversarial images to their original form, we employ techniques such as image super resolution and feature denoising. This process aims to bring the adversarial images closer to the original images, thereby improving their robustness against adversarial attacks. The performance is evaluated on various types of SVM classifiers, RF and k-NN. The results show that our defense method significantly improves the adversarial robustness.

3 Proposed Method

We detail the procedure of the proposed defense method in this section and summarized as follows: 1) Adversarial images are generated based on P-FGSM, StyleGAN and FLM, 2) Image restoration is used to enhance adversarial images, 3) WLMP encoded features are extracted from the input facial images, 4) Trained and tested on different types of classifiers for demonstrating the detection performance. Workflow of the proposed defense method is as shown in Fig. 3.

3.1 Adversarial Attacks

The adversarial face images are generated based on attacks P-FGSM[14], StyleGAN [15] and FLM [16]. The attacks are briefly discussed as follows:

3.1.1 P-FGSM[14]

Let an image I and $\hat{y}_i$ is its true class label of one of the scene types shown in I. Let a set of N scene classes of an image be $y_1, ..., y_i, ..., y_N$. Then, a multiclass classifier M is applied to image I to generate one-hot vector y of size N-dimensional, is given by:

$$y = M(I) \tag{1}$$

where $y = \{y_1, ..., y_i, ..., y_N\}$ is obtained from a selection on the probability vector $p = \{p_1, ..., p_i, ..., p_N\}$. Here, p_i is the probability of the scene class y_i of the image I.

$$p_i = p(y_i/I) \tag{2}$$

A transformation T is defined such that $\hat{I} = T(I)$ to induce M to classify the image I with a different scene label:

$$y \neq M(\hat{I}) \tag{3}$$

The transformation T aims to apply a minimal distortion to the image I in order to make it unnoticeable. Additionally, T should be designed to ensure that the true class label $\hat{y}_i$ cannot be inferred from the predicted class $M(\hat{I})$ or from the probability distribution of the predicted classes. Thus, T is defined as follows:

$$\hat{I} = T(I) = I + \delta_I^* \tag{4}$$

where δ_I^* is an adversarial perturbation. It is generated as follows:

$$\delta_I^* = arg_{\delta_I} max J_M(\theta, I + \delta_I, y) \tag{5}$$

In the P-FGSM, adversarial images are generated by adaptively targeting a class label $\hat{y}$ based on the classification probability vector p obtained from the classifier. To achieve a high misclassification rate, P-FGSM takes advantage of the fact that the true class labels are often among the class labels with the highest collective probabilities. It selects the target class label $\hat{y}$ from a subset of classes based on a specified threshold $\sigma \in [0, 1]$. To determine the target class label $\hat{y}$, the elements of the probability vector p are sorted in non-increasing order and denoted as $p' = \{p'_1, ..., p'_N\}$. The cumulative probability of each class label is calculated by summing the probabilities up to that class label.

$$\hat{y} = R(\{y_j : \sum_{i=1}^{j-1} p'_j > \sigma\}), \tag{6}$$

where R is a function that selects a class label arbitrarily from the input set and σ is a threshold to control the number of classes to select $\hat{y}$: a higher σ

denotes a smaller subset of target classes. P-FGSM generates the adversarial image $\hat{I} = \hat{I}_N$ iteratively, starting from $\hat{I}_0 = I$, as

$$\hat{I} = \hat{I}_{N-1} - \epsilon \times sign(\Delta_I J_M(\theta, \hat{I}_{N-1}, \hat{y})), \tag{7}$$

by increasing the prediction probability of class label $\hat{y}$ until a desired classification probability or a threshold on maximum number of iterations is reached.

3.1.2 FLM [16]

FLM utilizes the gradient of the model's prediction with respect to the facial landmarks to determine the displacement field. By computing the gradient, FLM obtains the direction in which each landmark should be moved to generate the adversarial landmark locations. This iterative process aims to find the optimal displacement field f that manipulates the facial landmarks to create the desired adversarial effect. The adversarial landmark location x_i^{adv} can be obtained by adding the displacement vector f_i to the original landmark x_i.

Let ϕ be a function for landmarks detection that maps the input face image into a set of n 2D landmark locations $X = \{x_1, x_2, ...x_n\}$, where $x_i = (u_i, v_i)$. Let $x_i^{adv} = (u_i^{adv}, v_i^{adv})$ is the obtained after transformation of x_i, and defines the i-th landmark location of the corresponding adversarial face image x^{adv}. FLM defines flow or displacement field f per landmark to produce the location of the corresponding adversarial landmarks for manipulating the input face image based on X. It optimizes the spatial displacement vector $f_i = (\Delta u_i, \Delta v_i)$ for the i-th landmark $x_i^{adv} = (u_i^{adv}, v_i^{adv})$. It uses the direction of the gradient of the prediction same as FGSM [8] to find the landmark displacement field f in an iterative manner. The adversarial landmark x_i^{adv} can be obtained from the original landmark x_i and the displacement vector f_i as:

$$x_i^{adv} = x_i + f_i, (u_i^{adv}, v_i^{adv}) = (u_i + \Delta u_i, v_i + \Delta v_i) \tag{8}$$

3.2 Feature Denoising and Deep Image Restoration Networks

An effective denoising technique can help to mitigate the effect of added perturbations if not eliminated because all adversarial attacks add noise to an input image in the form of well-crafted small perturbations. Image denoising either in the spatial or frequency domain causes a loss of textural information, which is counterproductive to our goal of producing clean image-like performance on denoised images. We denoise adversarial images using bilateral (BL) filter [45], which is the mostly used edge-preserving denoising technique. It combines both range and domain filtering to smooth images and preserves edges in a way similar to human performance. It averages only perceptually similar colors and preserves only perceptually visible edges.

Algorithm 1 A Robust Defense against Adversarial Facial Images with Image Restoration (BL + SR)

/* Image de-noising Input */
Input: Adversarial image x^{adv}
Output: Denoised Image $x_D = D(x^{adv})$
1. Convert the RGB image into gray color image using the transformation $0.299 * R + 0.587 * G + 0.114 * B$.
2. Denoise noisy patterns in the image using BL Filter.
3. Revert the denoised image back to RGB.

/* Image Super-Resolution (SR) */
Input: Denoised image $x_D = D(x^{adv})$
Output: Super Resolved Image $x_{SR} = N(x_D)$
4. Transform adversarial images back to normal image space using deep image restoration networks: N(.).

/* Adversarial Images Detection */
5. Extract encoded features for the recovered or super resolved images.
6. Forward the extracted features to the classifier model for correct prediction.

Fig. 5 Examples of original images, adversarial images generated using FLM, Restored adversarial images by BL Filter, Restored adversarial images by BL+SR.

Fig. 6 Examples of original images, adversarial images generated using P-FGSM, Restored adversarial images by Bilateral Filter, Restored adversarial images by BL+SR.

Image super-resolution (SR) reconstructs a high-resolution image I^{SR} from a low-resolution image I^{LR}. Depending on the situation, the relationship between I^{LR} and the original high-resolution image I^{HR} can change. Recently, DNNs [46, 47] have shown to significantly enhance peak signal-to-noise ratio (PSNR) in the SR problem. We reconstruct high-resolution images for denoised images based on an enhanced deep super-resolution network (EDSR) [48]. It consists of residual blocks and ResNet architecture and produced significantly improved performance in the single image SR problem.

3.3 Facial Adversarial Attacks Detection

Once the adversarial images restored into the original space, we extract WLMP features [49] from each facial image. It is observed that the use of smoothing and blending techniques in digital image editing is common for removing abnormalities in fake or altered face images. These techniques aim to create a more visually consistent appearance by reducing or eliminating noticeable artifacts or inconsistencies. As a result, the texture surfaces in these edited images can often appear mostly unaltered, making it challenging to detect the alterations visually. To handle this issue, the proposed defense method leverages WLMP features to highlight the most altered regions of the face images when they are being attacked or altered. The WLMP features encode the differences between a center pixel and its adjacent pixels, giving more weight to the nearest pixels compared to the ones farther away. By doing so, these features can effectively capture and emphasize the local changes in the image,

even if the overall texture surfaces appear largely unaltered. The procedure for extracting the WLMP features for each facial image is described as shown in Fig. 4. These extracted features are then provided as input to various types of classifiers to discriminate between the face adversarial images and the original ones.

The process of extracting WLMP features for each facial image involves the following steps:

1. Divide the input face image into multiple blocks of size 3×3. Each block represents a local region of the image.
2. Compute the differences between the center pixel of each block and its adjacent pixels. These differences are calculated as the absolute values of the pixel intensity differences.
3. To give higher weightage to the pixels that are closer to the center pixel, sort the obtained differences in increasing order. Then, multiply each absolute difference value by a weight factor of 2^p, where $p = 0, 1, 2, ..., 7$. This weight factor increases with the proximity of the adjacent pixel to the center pixel. Adjust the resulting value of the center pixel to ensure it falls within the range of 0 to 255. For example, if the computed value exceeds 255, set it to 255.
4. Compute the histogram feature vector based on the modified center pixel values across all the blocks in the image. The histogram represents the distribution of the modified pixel values.
5. Finally, the extracted feature vectors obtained from the training dataset are provided as input to different types of classifiers. These classifiers learn the presence of adversarial attacks based on the extracted WLMP features and can be used to classify facial images as either original or adversarial.

3.4 Algorithm Description

An algorithm of the proposed defense method is provided in Algorithm 1. First, we denoise the adversarial face image using BL filter. It smooths the effect of adversarial noise. After that, SR is performed as a mapping function to enhance the visual quality of images, which brings the images in the adversarial space into the original space in high-resolution. Then, encoded WLMP features are extracted for each facial image and trained with different types of SVM classifiers, Random Forest and k-NN in adversarial training fashion. Our defense method minimizes the effect of adversarial perturbations in the image domain and significantly improves the overall performance of the classifier.

4 Experimental Results

The performance of the proposed defense method is trained and tested on two real-world image datasets. Its performance is demonstrated with different types of classifiers.

Table 1 The overall results of the proposed defense method on CeleA Dataset

S.No	Dataset	Adversarial Attack	Image Restoration Networks	Classifier	Precision	Recall	F1-score	Accuracy(%)
1	CelebA	P-FGSM	No	Linear SVM	1.0	0.98	0.99	98.75
				Polynomial SVM	1.0	0.97	0.98	98.5
				Sigmoid SVM	0.72	0.73	0.72	72
				Gaussian SVM	1.0	0.96	0.98	98
				Random Forest	1.0	0.97	0.97	98.5
				k-NN	0.98	0.97	0.98	97.5
			BL+SR	Linear SVM	0.99	1.0	1.0	99
				Polynomial SVM	0.98	1.0	0.99	98
				Sigmoid SVM	0.93	1.0	0.96	93
				Gaussian SVM	0.99	1.0	0.99	99
				Random Forest	0.99	1.0	0.99	99
				k-NN	0.98	1.0	0.99	98

Table 2 The overall results of the proposed defense method on FFHQ Dataset

Dataset	Adversarial Attack	Classifier	No Image Restoration Networks				BL				BL+SR			
			Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)	Precision	Recall	F1-score	Accuracy (%)
FFHQ	StyleGAN	Linear SVM	0.71	0.7	0.7	**70.1**	0.85	0.85	0.85	**80.07**	0.85	0.85	0.85	**80.07**
		Polynomial SVM	0.78	0.78	0.78	**77.94**	0.89	0.89	0.89	**85.29**	0.89	0.89	0.89	**85.29**
		Sigmoid SVM	0.66	0.65	0.65	**65.69**	0.83	0.82	0.82	**77.12**	0.83	0.81	0.82	**76.47**
		Gaussian SVM	0.62	0.56	0.59	**60.29**	0.8	0.71	0.75	**68.63**	0.8	0.71	0.75	**68.95**
		Random Forest	0.66	0.67	0.66	**66.18**	0.83	0.82	0.82	**76.47**	0.83	0.81	0.82	**75.82**
		k-NN	0.71	0.48	0.57	**63.73**	0.88	0.73	0.8	**75.16**	0.88	0.71	0.79	**74.18**
	FLM	Linear SVM	0.67	0.58	0.62	**64.66**	0.78	0.66	0.72	**68.09**	0.78	0.67	0.72	**68.45**
		Polynomial SVM	0.65	0.43	0.52	**60.21**	0.77	0.53	0.63	**62.29**	0.77	0.53	0.63	**62.49**
		Sigmoid SVM	0.56	0.58	0.57	**55.86**	0.65	0.56	0.6	**55.24**	0.65	0.57	0.61	**55.68**
		Gaussian SVM	0.57	0.6	0.58	**57.66**	0.71	0.71	0.71	**64.81**	0.71	0.71	0.71	**64.65**
		Random Forest	0.59	0.62	0.6	**59.26**	0.71	0.72	0.71	**66.01**	0.71	0.72	0.71	**65.77**
		k-NN	0.59	1	0.74	**65.52**	0.67	0.95	0.79	**69.5**	0.67	0.94	0.78	**68.65**

4.1 Dataset

Our defense method is evaluated on two real-world datasets **CelebFaces Attributes Dataset (CelebA)**[50] and **Flickr-Faces-HQ, (FFHQ)**[15].

4.1.1 CelebA

It is a large-scale face attributes dataset consists of a total of 202,599 celebrity face images. Each face image in the dataset is annotated with 40 different attribute labels, providing additional information about various facial characteristics. It also includes face images captured in diverse conditions, such as different poses, background clutter and other variations. Moreover, it also exhibits a high level of diversity, containing a total of 10,177 unique identities.

4.1.2 FFHQ

It consists of a total of 70,000 real faces sourced from Flickr and an equal number of 70,000 fake faces generated using the StyleGAN technique [15]. It

encompasses a significant level of variation in terms of age, ethnicity and image backgrounds. It also includes a wide range of accessories such as eyeglasses, sunglasses, hats and more.

Fig. 5 shows examples of restored adversarial images for FLM attacks. The first row shows the original facial images and their corresponding FLM attacks are shown in the second row. In the third and forth rows, the restored adversarial images after BL and BL+SR are shown respectively. Similarly, Fig. 6 shows examples of original face images, their adversarial images generated by P-FGSM, the restored adversarial images after BL and BL+SR are shown in row-wise respectively.

4.2 Performance of the Proposed Defense Method

The proposed defense method is trained and tested on two real-world datasets. Its performance is demonstrated with various types of classifiers. The overall statistics of the proposed defense method on CeleA dataset is presented in Table 1. The results show that before employing image restoration, the classifiers Linear SVM, Polynomial SVM, Sigmoid SVM, Gaussian SVM, Random Forest and k-NN classifier detect with an accuracy of 98.75%, 98.5%, 72%, 98%, 98.5% and 97.5% respectively. Among the classifiers, Linear SVM shows its effectiveness in detecting facial adversarial images from the original with the highest accuracy of 98.75%. After employing image restoration such as BL followed by SR (BL+SR) to the adversarial images, the classification accuracy improves from 98.75% to 99% for Linear SVM, from 72% to 93% for Sigmoid SVM, from 98% to 99% for Gaussian SVM, from 98.5% to 99% for Random Forest and from 97.5% to 98% for k-NN on CelebA dataset with P-FGSM adversarial attack.

On FFHQ dataset with both adversarial attacks StyleGAN and FLM, our method achieves $5-10\%$ improvement in the classification accuracy in almost all classification models even if it achieves low classification accuracy before applying image restoration. The results on FFHQ dataset before and after applying image restoration (BL+SR) with adversarial attacks StyleGAN and FLM are shown in Table 2. Our experimental results show that BL alone is sufficient sometimes to bring back the adversarial images into the original space, leading the classifier towards correct prediction. Thus, the results show that significant improvement in the detection accuracy after employing the image restoration BL+SR on the adversarial images.

5 Evaluating Robustness of Intensity-based and Geometric-based Adversarial Attacks

Almost all intensity-based attacks augment the input samples with high-frequency components and employ a $l_p - norm$ constraint to regulate the distortion. The adversarial samples may not necessarily sit on the same manifold as the natural samples since the $l_p - norm$ is not a perfect similarity metric. On the other hand, geometric-based adversarial attacks are extremely

robust against adversarial training compared to intensity-based adversarial attacks because they are targeting the most important locations in the images using geometric perturbations. We use P-FGSM [14] and FLM [16] for intensity-based and geometric-based adversarial attacks, respectively. We evaluate the robustness of intensity-based and geometric-based adversarial attacks by extracting the encoded WLMP features with various classifiers on CelebA dataset. Geometric-based adversarial attacks are much more resistant against all evaluating classifiers except Sigmoid SVM. The overall statistics for evaluating the robustness of both adversarial attacks are presented in Table 3.

Table 3 Robustness comparison of intensity-based and geometric-based adversarial attacks on CelebA dataset

S.No	Classifier	Intensity-based Adversarial Attack (P-FGSM)				Geometric-based Adversarial Attack (FLM)			
		Precision	Recall	F1-score	Accuracy(%)	Precision	Recall	F1-score	Accuracy(%)
1	Linear SVM	0.96	0.97	0.96	**96.4**	0.77	0.75	0.76	**76.16**
2	Polynomial SVM	0.89	0.96	0.92	**91.89**	0.75	0.79	0.77	**76.16**
3	Random Forest	0.91	0.98	0.94	**94.14**	0.74	0.8	0.77	**75.99**
4	Sigmoid SVM	0.52	0.54	0.53	**51.35**	0.58	0.63	0.6	**58.77**
5	Gaussian SVM	0.91	0.94	0.92	**92.34**	0.7	0.79	0.74	**72.35**
6	k-NN	0.87	0.99	0.93	**91.89**	0.74	0.49	0.59	**65.73**

6 Conclusions

A new defense method for improving robustness against facial adversarial attacks is proposed based on deep image restoration networks. We generate a well-protected version of adversarial face images based on P-FGSM, FLM and StyleGAN and have proved that these images can mislead the classifier to misclassification with high confidence. Image restorations such BL followed by SR are performed on adversarial images to enhance the visual quality of images, which brings back the low resolution adversarial images into the high resolution original space. The encoded features are extracted for the recovered images and trained on various types of classifiers. The results are demonstrated on two real-world datasets for different adversarial attacks. The experimental results show that there is a significant improvement in the classification accuracy after employing the image restoration in the classification models and also geometric-based adversarial attacks are more robust to defend than intensity-based adversarial facial adversarial attacks.

We will focus on various denoising filters and deep image super resolution networks to improve adversarial robustness further for different adversarial attacks in future work. We will also investigate whether SR or denoising alone is sufficient for all types of DL-based adversarial attacks.

Conflict of Interest:

The authors have no conflicts of interest to declare.

References

[1] R. Girshick, "Fast r-cnn," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1440–1448.

[2] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in neural information processing systems*, vol. 25, pp. 1097–1105, 2012.

[3] D. Amodei, S. Ananthanarayanan, R. Anubhai, J. Bai, E. Battenberg, C. Case, J. Casper, B. Catanzaro, Q. Cheng, G. Chen *et al.*, "Deep speech 2: End-to-end speech recognition in english and mandarin," in *International conference on machine learning*. PMLR, 2016, pp. 173–182.

[4] L. Deng and Y. Liu, *Deep learning in natural language processing*. Springer, 2018.

[5] A. Ortis, G. M. Farinella, and S. Battiato, "An overview on image sentiment analysis: Methods, datasets and current challenges." in *ICETE (1)*, 2019, pp. 296–306.

[6] F. Carrara, A. Esuli, T. Fagni, F. Falchi, and A. M. Fernández, "Picture it in your mind: Generating high level visual representations from textual descriptions," *Information Retrieval Journal*, vol. 21, no. 2, pp. 208–229, 2018.

[7] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," *arXiv preprint arXiv:1312.6199*, 2013.

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

[9] N. Carlini and D. Wagner, "Towards evaluating the robustness of neural networks," in *2017 ieee symposium on security and privacy (sp)*. IEEE, 2017, pp. 39–57.

[10] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 9185–9193.

[11] A. Kurakin, I. Goodfellow, S. Bengio *et al.*, "Adversarial examples in the physical world," 2016.

[12] Y. Liu, X. Chen, C. Liu, and D. Song, "Delving into transferable adversarial examples and black-box attacks," *arXiv preprint arXiv:1611.02770*, 2016.

[13] N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, and A. Swami, "Practical black-box attacks against machine learning," in *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, 2017, pp. 506–519.

[14] C. Y. Li, A. S. Shamsabadi, R. Sanchez-Matilla, R. Mazzon, and A. Cavallaro, "Scene privacy protection," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 2502–2506.

[15] T. Karras, S. Laine, and T. Aila, "A style-based generator architecture for generative adversarial networks," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4401–4410.

[16] A. Dabouei, S. Soleymani, J. Dawson, and N. Nasrabadi, "Fast geometrically-perturbed adversarial faces," in *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2019, pp. 1979–1988.

[17] S. Gu and L. Rigazio, "Towards deep neural network architectures robust to adversarial examples," *arXiv preprint arXiv:1412.5068*, 2014.

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

[19] N. Papernot, P. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a defense to adversarial perturbations against deep neural networks," in *2016 IEEE symposium on security and privacy (SP)*. IEEE, 2016, pp. 582–597.

[20] A. Rozsa, E. M. Rudd, and T. E. Boult, "Adversarial diversity and hard positive generation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 25–32.

[21] N. Carlini and D. Wagner, "Adversarial examples are not easily detected: Bypassing ten detection methods," in *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, 2017, pp. 3–14.

[22] K. Grosse, P. Manoharan, N. Papernot, M. Backes, and P. McDaniel, "On the (statistical) detection of adversarial examples," *arXiv preprint arXiv:1702.06280*, 2017.

[23] J. H. Metzen, T. Genewein, V. Fischer, and B. Bischoff, "On detecting adversarial perturbations," *arXiv preprint arXiv:1702.04267*, 2017.

[24] Z. Gong, W. Wang, and W.-S. Ku, "Adversarial and clean data are not twins," *arXiv preprint arXiv:1704.04960*, 2017.

[25] A. N. Bhagoji, D. Cullina, and P. Mittal, "Dimensionality reduction as a defense against evasion attacks on machine learning classifiers," *arXiv preprint arXiv:1704.02654*, vol. 2, 2017.

[26] X. Li and F. Li, "Adversarial examples detection in deep networks with convolutional filter statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 5764–5772.

[27] C. Sadu and P. K. Das, "A defense method against facial adversarial attacks," in *TENCON 2021-2021 IEEE Region 10 Conference (TENCON)*. IEEE, 2021, pp. 459–463.

[28] B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Šrndić, P. Laskov, G. Giacinto, and F. Roli, "Evasion attacks against machine learning at test time," in *Joint European conference on machine learning and knowledge discovery in databases*. Springer, 2013, pp. 387–402.

[29] S.-M. Moosavi-Dezfooli, A. Fawzi, and P. Frossard, "Deepfool: a simple and accurate method to fool deep neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2574–2582.

[30] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, "The limitations of deep learning in adversarial settings," in *2016 IEEE European symposium on security and privacy (EuroS&P)*. IEEE, 2016, pp. 372–387.

[31] P.-Y. Chen, H. Zhang, Y. Sharma, J. Yi, and C.-J. Hsieh, "Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models," in *Proceedings of the 10th ACM workshop on artificial intelligence and security*, 2017, pp. 15–26.

[32] A. Ilyas, L. Engstrom, A. Athalye, and J. Lin, "Black-box adversarial attacks with limited queries and information," in *International Conference on Machine Learning*. PMLR, 2018, pp. 2137–2146.

[33] M. Alzantot, Y. Sharma, S. Chakraborty, H. Zhang, C.-J. Hsieh, and M. B. Srivastava, "Genattack: Practical black-box attacks with gradient-free optimization," in *Proceedings of the Genetic and Evolutionary Computation Conference*, 2019, pp. 1111–1119.

[34] C. Xiao, B. Li, J.-Y. Zhu, W. He, M. Liu, and D. Song, "Generating adversarial examples with adversarial networks," *arXiv preprint arXiv:1801.02610*, 2018.

[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in

Advances in neural information processing systems, 2014, pp. 2672–2680.

[36] J. Deng, J. Guo, N. Xue, and S. Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[37] D. Deb, J. Zhang, and A. K. Jain, "Advfaces: Adversarial face synthesis," in *2020 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2020, pp. 1–10.

[38] H. Qiu, C. Xiao, L. Yang, X. Yan, H. Lee, and B. Li, "Semanticadv: Generating adversarial examples via attribute-conditioned image editing," in *European Conference on Computer Vision*. Springer, 2020, pp. 19–37.

[39] H. Xu, Y. Ma, H.-C. Liu, D. Deb, H. Liu, J.-L. Tang, and A. K. Jain, "Adversarial attacks and defenses in images, graphs and text: A review," *International Journal of Automation and Computing*, vol. 17, no. 2, pp. 151–178, 2020.

[40] D. Hendrycks and K. Gimpel, "Early methods for detecting adversarial images," *arXiv preprint arXiv:1608.00530*, 2016.

[41] F. V. Massoli, F. Carrara, G. Amato, and F. Falchi, "Detection of face recognition adversarial attacks," *Computer Vision and Image Understanding*, vol. 202, p. 103103, 2021.

[42] A. Agarwal, R. Singh, M. Vatsa, and N. Ratha, "Are image-agnostic universal adversarial perturbations for face recognition difficult to detect?" in *2018 IEEE 9th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. IEEE, 2018, pp. 1–7.

[43] C. Xie, Y. Wu, L. v. d. Maaten, A. L. Yuille, and K. He, "Feature denoising for improving adversarial robustness," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 501–509.

[44] A. Mustafa, S. H. Khan, M. Hayat, J. Shen, and L. Shao, "Image super-resolution as a defense against adversarial attacks," *IEEE Transactions on Image Processing*, vol. 29, pp. 1711–1724, 2019.

[45] C. Tomasi and R. Manduchi, "Bilateral filtering for gray and color images," in *Sixth international conference on computer vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 839–846.

[46] J. Kim, J. K. Lee, and K. M. Lee, "Accurate image super-resolution using very deep convolutional networks," in *Proceedings of the IEEE conference*

on computer vision and pattern recognition, 2016, pp. 1646–1654.

[47] ——, "Deeply-recursive convolutional network for image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 1637–1645.

[48] B. Lim, S. Son, H. Kim, S. Nah, and K. Mu Lee, "Enhanced deep residual networks for single image super-resolution," in *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 136–144.

[49] A. Agarwal, R. Singh, M. Vatsa, and A. Noore, "Swapped! digital face presentation attack detection via weighted local magnitude pattern," in *2017 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 2017, pp. 659–665.

[50] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.