Improving Transferability of Adversarial Examples on Face Recognition with Beneficial Perturbation Feature Augmentation

Fengfan Zhou, Hefei Ling, Yuxuan Shi, Jiazhong Chen, Zongyi Li, Qian Wang
Huazhong University of Science and Technology
{ffzhou, lhefei, shiyx, jzchen, zongyili, yqwq1996}@hust.edu.cn

Abstract

*Face recognition (FR) models can be easily fooled by adversarial examples, which are crafted by adding imperceptible perturbations on benign face images. To improve the transferability of adversarial examples on FR models, we propose a novel attack method called **B**eneficial **P**erturbation **F**eature Augmentation **A**ttack (BPFA), which reduces the overfitting of the adversarial examples to surrogate FR models by the adversarial strategy. Specifically, in the backpropagation step, BPFA records the gradients on pre-selected features and uses the gradient on the input image to craft adversarial perturbation to be added on the input image. In the next forward propagation step, BPFA leverages the recorded gradients to add perturbations(i.e., beneficial perturbations) that can be pitted against the adversarial perturbation added on the input image on their corresponding features. The above two steps are repeated until the last backpropagation step before the maximum number of iterations is reached. The optimization process of the adversarial perturbation added on the input image and the optimization process of the beneficial perturbations added on the features correspond to a minimax two-player game. Extensive experiments demonstrate that BPFA outperforms the state-of-the-art gradient-based adversarial attacks on FR.*

1. Introduction

Deep learning has achieved great success in various computer vision fields in the last decade [9] [34] [33] [32] [37]. As a typical application of deep learning, the performance of FR has also been greatly improved [25] [35] [6]. FR's performance is nearing saturation on some datasets(*e.g.*, LFW [16] and MegaFace [19]). Due to FR's superior performance, FR has been widely used in scenarios with high-security requirements, such as airport security checks, financial payment, and mobile phone unlocking [39].

However, previous work has shown that FR is vulnerable to adversarial examples [42]. The attacker can successfully fool an FR model by adding imperceptible adversarial perturbations on the input image of the FR model [45]. In addition, the adversarial face examples crafted by the attacker are transferable across different FR models. The transferability of adversarial face examples enables the attacker to attack the victim FR model successfully without knowing any information about it [39] [15] [45]. The feasibility of black-box attacks on FR models poses a significant threat to the existing FR applications.

In recent years, adversarial attacks on FR have made great progress. Among the recently proposed adversarial attacks on FR, patch-based adversarial attacks have been widely studied. These adversarial attacks mainly use generator-based methods or gradient-based methods to add adversarial perturbations on parts of the face images [20] [26] [39] [27]. However, there exist significant visual differences between the adversarial examples crafted by these attack methods and the original attacker images, which makes the crafted adversarial examples very noticeable and easy to cause people's vigilance. In addition, some works have studied the adversarial attacks on FR based on makeup transfer. These attacks craft adversarial face examples in the form of makeup using the technology of makeup transfer [42] [15]. If the attacker image is an image of a female, the adversarial examples crafted by these attacks are often satisfactory. However, if the attacker image is an image of a male, the adversarial examples crafted by these attacks are not very satisfactory, and the crafted adversarial examples are easy to arouse the suspicion of others. There are also some recent works that use adversarial face examples for face encryption [41] [5]. However, these works ignore improving the transferability of adversarial face examples crafted by a single FR model, which leads to their low black-box attack performance.

Few adversarial attacks on FR can solve the above problems and DFANet [45] is one of them. DFANet limits the maximum deviation of the added adversarial perturbations to a predefined value under the L_p norm. To improve the transferability of the crafted adversarial examples, DFANet

performs random dropout on the features of the convolutional layers to obtain ensemble-like effects, which can be regarded as a feature augmentation method. However, we argue that *random* may not be optimal. Adding elaborate noise on the features may bring better black-box attack performance to DFANet.

Then what kind of noise should be added on the features to improve the performance of DFANet is a problem to be solved. To find out this noise, we conducted a survey. We find that adversarial strategies [43] [31] bring significant improvements to the generalization of trained models on computer vision tasks other than adversarial attacks. Therefore, we attempt to add noises that can be pitted against the optimization process of the adversarial perturbation added on the input image on features to improve the transferability of adversarial examples. For the method of crafting such noises, we plan to adopt a similar method as the method of crafting the adversarial perturbations. Specifically, we use the adversarial perturbations that can be pitted against the adversarial perturbation added on the input image as these noises, and these noises are called beneficial perturbations.

Beneficial perturbation was proposed by Wen and Itti [36], and the original purpose of the beneficial perturbation was to improve the adversarial robustness [23] [17] [24] of the model. Here we give the definition of beneficial perturbation:

Definition 1 ***Beneficial Perturbation***
Let the optimization objective of the crafted adversarial example x^{adv} be to decrease the loss l. For a particular input x, the sample that is crafted using the following optimization objective is called beneficial example x^{ben}:

$$x^{ben} = \arg\max_{x^{ben}} \left(l \left(x^{ben} \right) \right) \tag{1}$$

The beneficial example can be obtained by adding the perturbation with value $x^{ben} - x$ on the input x, and the perturbation is called the beneficial perturbation Δx^{ben}.

In contrast, the optimization objective of the adversarial perturbation $\Delta x^{adv} = x^{adv} - x$ can be expressed as:

$$x^{adv} = \arg\min_{x^{adv}} \left(l \left(x^{adv} \right) \right) \tag{2}$$

where x^{adv} is the crafted adversarial example. The optimization objective for beneficial perturbations is the same as the optimization objective for the task on which the perturbations are applied and is not intended to fool the model. Therefore, this type of perturbations are *beneficial* and can be pitted against adversarial perturbations. A comparison of benign pictures, adversarial examples, and beneficial samples is shown in Fig. 1.

We argue that adding the beneficial perturbations on the features can make the optimization process of adversarial examples harder which can make the crafted adversarial examples escape poor local optima and improve the transferability of them. However, we cannot directly compute the beneficial perturbations before backpropagation since we cannot obtain the gradients on the features. To address this issue, we use the gradient calculated in the last backpropagation to simulate the gradient calculated in this backpropagation. We refer to the proposed attack method as **B**eneficial **P**erturbation **F**eature Augmentation **A**ttack (BPFA).

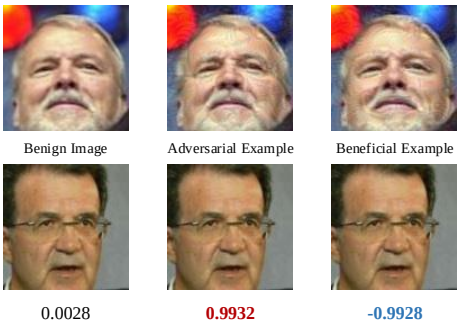

Figure 1. Comparison of benign images, adversarial examples, and beneficial examples. The three images in the first row are a benign image, an adversarial example, and a beneficial example, respectively. The images in the second row are three identical victim images. The number below the victim image is the cosine similarity between it and the image above it.

Our main contributions are summarized as follows:

- Based on the adversarial strategy, we propose a novel adversarial attack method on FR named BPFA. In the process of crafting adversarial examples using BPFA, the beneficial perturbations are added on the preselected features of the FR model to be pitted against the adversarial perturbation added on the input image to improve the transferability of the crafted adversarial examples.
- We explore the properties of beneficial perturbations added on the features of the FR model. We find that the beneficial perturbations added on the features may have some semantic information and analyze the causes for the semantic information.
- Extensive experiments show that the proposed BPFA method achieves better black-box attack performance than the existing gradient-based adversarial attacks on FR without harming the white-box attack performance.

2. Related Work

2.1. Face Recognition

The main process of the latest face recognition is to use the FR model to extract the features of two face images and

calculate the similarity (*e.g*., cosine similarity) between the extracted features of the two images. Next, the calculated similarity is compared to a specific threshold that is set in advance. If the similarity is greater than the threshold, the individuals in the two face images are judged to be the same identity, otherwise the different identities [29].

Recent works on FR have focused on reducing the intra-class distance and increasing the inter-class distance of features extracted by FR models [35]. Schroff *et al*. [25] proposed FaceNet and the triplet loss to map face images into a compact Euclidean space where distances directly represent the similarity of faces. Wang *et al*. [35] proposed CosFace, which uses the large margin cosine loss to supervise the training of FR models and further improves the performance of FR tasks. To obtain more discriminative features for FR, Deng *et al*. [6] proposed ArcFace that leverages the additive angular margin loss to train the FR models and achieves higher performance than CosFace.

2.2. Adversarial Attack on Face Recognition

Adversarial attacks on FR can be mainly classified into two categories: gradient-based attacks and generator-based attacks. We will introduce these two types of adversarial attacks on FR in the following.

Most gradient-based adversarial attacks on FR are improved based on I-FGSM [21]. The FR belongs to the open-set task for which feature-level loss is more suitable. Therefore, Zhong and Deng [44] proposed FIM. Compared with I-FGSM, the main improvement of FIM is to change the label-level loss to the feature-level loss. To improve the transferability of adversarial face examples, Zhong and Deng [45] proposed DFANet. DFANet performs dropout on the features of the convolutional layers during the forward propagation to obtain ensemble-like effects. For face encryption, Yang *et al*. [41] proposed TIP-IM. Different from FIM and DFANet, TIP-IM focuses on attacking face classification tasks. The main operation of TIP-IM is to use MMD [3] loss to improve the visual quality of the crafted adversarial examples and use the greedy insertion to select the optimal victim image from a predefined gallery set.

The generator-based adversarial attacks on FR consist of two stages: training and inference. In the training phase, the training method of GAN [10] is used. In the inference phase, only the generator is used to generate adversarial examples. Currently, the generator-based adversarial attacks on FR mainly focus on generating adversarial examples of a particular part of the face [39] and generating adversarial examples based on makeup transfer [42] [15].

3. Methodology

This section introduces our proposed BPFA method. Sec. 3.1 introduces the problem formulation. Sec. 3.2 focuses on the detailed construction of BPFA. To make the explanation of BPFA more clear, Sec. 3.3 introduces the crafting method of beneficial perturbations added on features, and Sec. 3.4 introduces the gradient recording method of BPFA.

3.1. Problem Formulation

In general, there are two categories of adversarial attacks on FR, *i.e*., impersonation (targeted) attacks and dodging (untargeted) attacks [42]. Impersonation attacks on FR aim to fool the FR model into recognizing adversarial examples as the face images of the target identity pre-selected by the attacker, while dodging attacks on FR aim to fool the FR model into recognizing adversarial examples as the face images of someone other than the attacker identity. In this paper, we mainly study the impersonation attack on FR, and we will introduce the impersonation attack on FR in the following. Let $f^{vct}(x)$ be an FR model used by the victim to extract the feature vector from the face image x. Let x^s and x^t be the attacker image and the victim image, respectively. The optimization objective of impersonation attacks on FR can be expressed as:

$$\begin{aligned} x^{adv} = \arg\min_{x^{adv}} \left(\mathcal{D}\left(f^{vct}\left(x^{adv}\right), f^{vct}\left(x^{t}\right)\right)\right) \\ \text{s.t.} \|\Delta x^{adv}\|_p \leq \epsilon \end{aligned} \tag{3}$$

where $\mathcal{D}$ is a pre-selected distance metric, and ϵ is the maximum allowable perturbation magnitude.

3.2. Beneficial Perturbation Feature Augmentation Attack

In most cases, the attacker cannot obtain the victim model f^{vct}. Therefore the optimization goal of Eq. (3) cannot be directly realized. A common method to realize the optimization objective of Eq. (3) is to use a surrogate model f^{sur} that the attacker can obtain to craft adversarial examples and transfer the crafted adversarial examples to the victim model to carry out the attack. This requires that the adversarial examples crafted using the surrogate model have strong transferability. For the improvement of the transferability of adversarial face examples, one of the typical methods is DFANet [45]. To the best of our knowledge, DFANet is also the latest adversarial attack on FR under the problem formulation of this paper.

DFANet can be regarded as an attack method that improves the transferability of adversarial face examples by feature augmentation. Although DFANet greatly improves the transferability of adversarial face examples, DFANet only performs *random* dropout on the features of convolutional layers of the FR model. We argue that elaborately designed feature augmentation methods can achieve better results than random feature augmentation methods. To some extent, the optimization process of DFANet is too easy, and

the crafted adversarial examples are easy to fall into poor local optima and "overfit" the surrogate models. If we could somehow make the optimization process of DFANet harder, adversarial examples crafted using the harder optimization process might be able to escape poor local optima, resulting in transferability improvements. If the optimization objective of adversarial examples is to minimize a specific loss, then to make the optimization process harder, we should be pitted against the optimization objective of adversarial examples, which is to maximize the loss. Therefore, in the process of crafting adversarial examples, we add the beneficial perturbations on the pre-selected features to be pitted against the optimization process of the adversarial examples to make the adversarial examples escape from poor local optima. We name this attack method as **B**eneficial **P**erturbation **F**eature Augmentation **A**ttack (BPFA).

The optimization objective of BPFA can be express as:

$$\min_{\Delta x^{adv}} \max_{\Delta \Omega^{ben}} \mathcal{L}\left(x^{adv}, x^{t}\right) \tag{4}$$

where Δx^{adv} is the crafted adversarial perturbation that need to be added on the attacker image x^s, and $\Delta\Omega^{ben}$ is the set of the crafted beneficial perturbations that need to be added on a pre-selected feature set Ω of the surrogate model f^{sur}. $\mathcal{L}$ is the loss function used to calculate the distance between the features extracted from f^{sur}:

$$\mathcal{L}\left(x^{adv}, x^{t}\right) = \|\phi\left(f_{\Omega}^{sur}\left(x^{adv}\right)\right) - \phi\left(f^{sur}\left(x^{t}\right)\right)\|_2^2 \tag{5}$$

where $\phi(x)$ denotes the operation that normalizes x, and f_{Ω}^{sur} denotes the model that adds beneficial perturbations on the feature set Ω during forward propagation. x^{adv} is the adversarial example that initiates with the same value of x^s.

The optimization process of BPFA corresponds to a minimax two-player game. Beneficial perturbations added on the features of the model tend to increase the loss, while adversarial perturbations added on the input images tend to decrease the loss. The pipeline of the proposed BPFA is depicted in Fig. 2.

3.3. Crafting Method of Beneficial Perturbation Added on Features

In this section, we will introduce the crafting method of beneficial perturbations added on features in detail. If we use the loss of Eq. (5) to craft adversarial examples, we can use the following formula:

$$x^{adv} = x^{s} - \eta \text{sign}\left(\nabla_{x^s}\mathcal{L}\left(x^{adv}, x^{t}\right)\right) \tag{6}$$

where η is a nonnegative step size. A larger η indicates a larger magnitude of added perturbations. Let a specific feature in the attacker FR model be ω, then the formula for adding an adversarial perturbation $\Delta\omega^{adv} = \omega^{adv} - \omega$ to ω can be expressed as:

$$\omega^{adv} = \omega - \eta \text{sign}\left(\nabla_{\omega}\mathcal{L}\left(x^{adv}, x^{t}\right)\right) \tag{7}$$

Eq. (7) is improved based on FGSM [11], and the linearity assumption of FGSM is still applicable to Eq. (7). Based on the linearity assumption of the model, the objective of the adversarial perturbation crafted by Eq. (7) is to minimize the loss $\mathcal{L}\left(x^{adv}, x^{t}\right)$. Therefore, if we want to craft a beneficial perturbation $\Delta\omega^{ben} = \omega^{ben} - \omega$ that maximizes the loss, we can use the following formula:

$$\omega^{ben} = \omega + \eta \text{sign}\left(\nabla_{\omega}\mathcal{L}\left(x^{adv}, x^{t}\right)\right) \tag{8}$$

The main difference between Eq. (8) and Eq. (7) is that Eq. (8) has a positive sign before η, which is opposite to the sign before η in Eq. (7).

What we have described above is the method of adding a beneficial perturbation on a single feature. In real cases, we can add beneficial perturbations on multiple features. The process of adding beneficial perturbations on multiple features can be viewed as the combination of repeatedly executing the process of adding a beneficial perturbation on a single feature. In practice, we can use backpropagation of computational graphs to calculate the gradients on all the features in the network at once and pick the gradients we need to craft beneficial perturbations.

3.4. Record gradient with a memory bank

If we want to use beneficial perturbations for feature augmentation, we need to add the pre-computed beneficial perturbation $\Delta\omega^{ben}$ on the feature ω of the model before backpropagation. However, we cannot compute the gradients on features before backpropagation. Therefore, we cannot use Eq. (8) to compute the beneficial perturbation that should be added on the feature ω directly. If we use one backpropagation to compute the beneficial perturbations and another backpropagation to compute the adversarial perturbation, the amount of computation to craft adversarial examples will increase a lot. Inspired by the memory bank [38] in the unsupervised feature learning [12] [4], we record the gradients computed during the last backpropagation and use them to calculate the beneficial perturbations we need to add on features to reduce the amount of computation. Specifically, we add a gradient memory bank to each layer corresponding to the feature for which we plan to add the beneficial perturbation, to record the gradient on the feature computed in the last backpropagation. In the forward propagation, we simulate the gradient in the backpropagation of this time with the gradient pre-recorded in the gradient memory bank, and use the simulated gradient to calculate the beneficial perturbation that needs to be added on the feature corresponding to the gradient memory bank. The experimental results show that the beneficial perturbations that we calculated using the simulated gradients can still increase the loss well, and be pitted against the optimization process of the adversarial perturbation that need to be added on the input image.

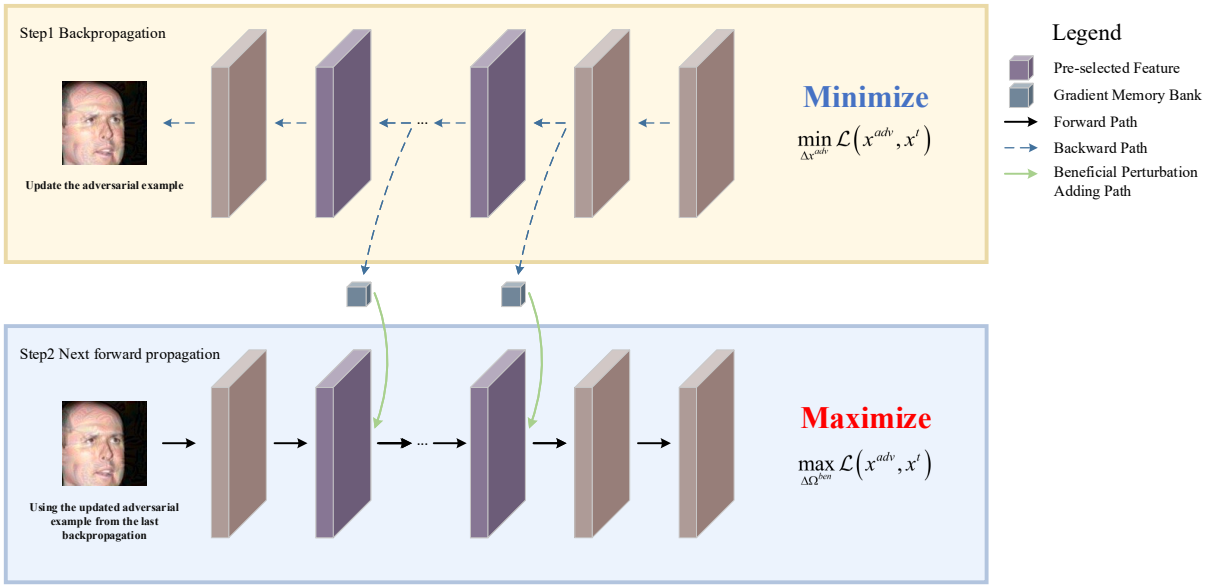

Figure 2. The pipeline of the proposed BPFA method. The initial forward propagation is omitted. In the process of backpropagation (step 1), the gradients on the pre-selected features that need to be added beneficial perturbations are recorded in the gradient memory banks, and the gradient on the input image is used to craft adversarial perturbation Δx^{adv} that need to be added on the input image to minimize the loss. Then, in the next forward propagation (step 2), we input the input image after adding the adversarial perturbation Δx^{adv} into the model, and use the gradients recorded in the gradient memory banks to craft beneficial perturbations $\Delta\Omega^{ben}$ and add them on their corresponding features to maximize the loss, thus being pitting against the optimization process of adversarial perturbation. The procedures of step 1 and step 2 above are executed repeatedly until the last backpropagation before the maximum number of iterations is reached

4. Experiments

Sec. 4.1 introduces the experimental setting. Sec. 4.2 presents the experiment results of the proposed BPFA and other baseline methods. Sec. 4.3 perform ablation experiments on the proposed BPFA. Sec. 4.4 studies the effectiveness of simulated gradients. Sec. 4.5 perform experiments on the interpretation of beneficial perturbations added on the features.

4.1. Experimental Setting

Datasets. We choose LFW [16] and CelebA-HQ [18] as the datasets for our experiments. LFW is a face dataset for unconstrained FR. CelebA-HQ is a dataset with high quality images. These two datasets are commonly used in the research of adversarial attack on FR [39] [42] [15].

Following [45] [41] [15] [39] [8], we also select a part of the data from the selected dataset as the dataset to test the performance of the attack methods. Specifically, we randomly select 1000 face pairs from the LFW dataset as the dataset for LFW, and 1000 face pairs from the CelebA-HQ dataset as the dataset for CelebA-HQ. All the pairs we select are negative pairs, where one image in a pair is used as the attacker image and the other as the victim image. Without special emphasis, we will use LFW and CelebA-HQ to refer to our selected LFW and CelebA-HQ datasets, respectively.

Face Recognition Models. The FR models used in our experiments are FaceNet [25], MobileFace [6](we denote it as MF in the following), IRSE50 [14], and IR152 [13] that are identical with the FR models used in [42] and [15]. We choose the threshold when the FAR tested on the complete LFW dataset is 0.001 as the threshold to calculate the attack success rate.

Attack Setting. All the attacks in the experiments are conducted under the setting of impersonation attacks, because impersonation attacks are more difficult compared to dodging attacks [45] [22]. Following [45], we set the maximum allowable perturbation magnitude ϵ to 10 based on L_∞ norm bound with respect to pixel values in $[0, 255]$, and the maximum iterative step to 1500.

Evaluation Metrics. We use *attack success rate* (ASR) to evaluate the performance of different attacks. In impersonation attacks, ASR is the ratio of adversarial examples identified by the FR model as victim images (*i.e.*, the ratio of adversarial examples that attack successfully).

Baseline methods Since Adv-Makeup [42] and AMT-GAN [15] are attacks based on makeup transfer and GenAP is patch-based attack, it is unreasonable to use L_∞ norm to limit their maximum allowable perturbation magnitude on the whole face images. TIP-IM [41] is an attack used for face encryption that requires high visual quality. Our experiments demonstrate that if TIP-IM is directly applied

under the setting of this paper, the ASR of the combination of FIM [44], MI [7], DI [40], and TIP-IM is lower than the combination of FIM, MI, DI which we denote it as FMD. Therefore, Adv-Makeup, AMT-GAN, GenAP and TIP-IM are not used as baselines in this paper.

For the setting of this paper, the state-of-the-art attack method is the combination of FIM [44], MI [7], DI [40], and DFANet [45] which we denote it as FMDN. Therefore, we mainly use FMDN as our baseline.

4.2. Comparison Study

We study the performance of FIM, FMD, FMDN, and FMDN-BPFA on the LFW dataset and CelebA-HQ dataset in comparison experiments. The locations where BPFA adds beneficial perturbations are on the features of the convolutional layers in the comparison experiments. The results of different attacks on LFW and CelebA-HQ datasets are shown in Tab. 1. Tab. 1 shows that FMDN-BPFA achieves performance improvement compared with FMDN. Examples of adversarial examples crafted by FIM, FMD, FMDN, and FMDN-BPFA are illustrated in Fig. 3. After adding DFANet to FMD, FMDN achieves higher performance than FMD in most cases. However, In the case of using MF as the attacker model and crafting adversarial examples on the CelebA-HQ dataset, FMDN does not achieve performance improvement compared with FMD. In the same case, after adding our proposed BPFA, FMDN-BPFA achieves better performance than FMDN, thus indicating that BPFA can compensate for the deficiency of DFANet to some extent.

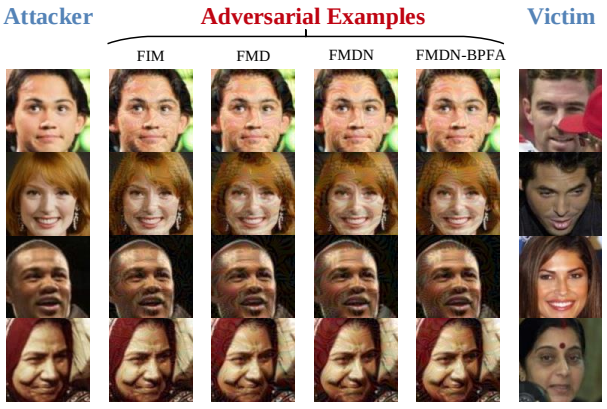

Figure 3. Illustration of the adversarial examples crafted by FIM, FMD, FMDN, and FMDN-BPFA. The images in the first and last columns are the attacker images and victim images, respectively. The second to fifth columns demonstrate the adversarial examples crafted by FIM, FMD, FMDN, and FMDN-BPFA, respectively.

4.3. Ablation Studies

An important hyperparameter of our proposed BPFA is the step size η of the beneficial perturbations to be added, and another important hyperparameter is the layers where the beneficial perturbations are added. In the following, we will conduct ablation studies on them.

4.3.1 Effect of the Step Size of Beneficial Perturbation

We test the black-box attack performance of FMDN-BPFA under different η on the LFW dataset with MF as the attacker model, where the beneficial perturbations were added on the features of all convolutional layers. The experiment results are illustrated in Fig. 4.

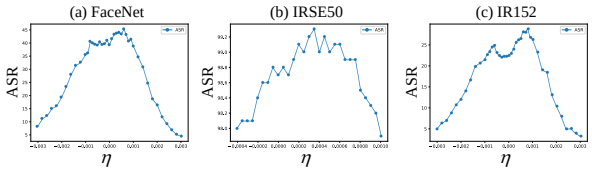

Figure 4. ASR of FMDN-BPFA under different η with MF as the attacker model on the LFW dataset under the black-box attack setting.

Fig. 4 shows that with the increase of η, the black-box attack performance of the adversarial examples crafted by FMDN-BPFA generally shows a trend of first increasing and then decreasing. When η is 0, FMDN-BPFA degenerates to FMDN. When η is less than 0, the attack method adds adversarial perturbations instead of beneficial perturbations on the features. To analyze the reasons behind Fig. 4, we should consider how BPFA works. BPFA adds beneficial perturbations on the features to be pitted against the optimization process of adversarial perturbations added on the input image, thereby improving the transferability of the adversarial examples. When η is greater than 0, if the value of η is too small, the strength of the beneficial perturbations added on the feature can not be pitted against the adversarial examples well. If the value of η is too large, the important feature will be destroyed by the beneficial perturbations. Therefore, the curve first increases and then decreases when η is greater than 0. When η is less than 0, we add adversarial perturbations on the features. We name this attack method as **A**dversarial **P**erturbation **F**eature Augmentation **A**ttack (APFA). Note that APFA can also improve the transferability of the crafted adversarial examples in some cases. We argue that the improvement of the transferability is due to the important information contained in the adversarial perturbations crafted by gradients (see Sec. 4.5). However, Fig. 4 shows that the performance improvement brought by APFA is much lower than BPFA.

	attack	LFW				CelebA-HQ			
		FaceNet	MF	IRSE50	IR152	FaceNet	MF	IRSE50	IR152
FaceNet	FIM	**100.0***	4.9	11.9	7.2	**100.0***	8.6	14.8	9.2
	FMD	**100.0***	25.0	42.9	29.9	**100.0***	27.8	38.7	29.1
	FMDN	**100.0***	28.9	46.5	33.1	**100.0***	31.6	42.2	30.9
	FMDN-BPFA	**100.0***	**34.7**	**53.9**	**39.1**	**100.0***	**35.8**	**46.4**	**34.9**
MF	FIM	6.7	**100.0***	65.4	4.4	6.9	**100.0***	60.6	6.5
	FMD	40.5	**100.0***	98.5	22.0	36.8	**100.0***	98.1	28.2
	FMDN	40.8	**100.0***	98.8	23.5	36.8	**100.0***	98.1	28.1
	FMDN-BPFA	**45.4**	**100.0***	**99.0**	**27.1**	**38.3**	**100.0***	**98.3**	**31.3**
IRSE50	FIM	12.7	74.4	**100.0***	27.3	14.5	76.2	**100.0***	31.8
	FMD	55.1	98.8	**100.0***	64.8	48.1	98.7	**100.0***	63.4
	FMDN	58.5	99.0	**100.0***	68.2	49.3	98.7	**100.0***	63.6
	FMDN-BPFA	**59.3**	**99.1**	**100.0***	**71.5**	**51.2**	**98.8**	**100.0***	**66.9**
IR152	FIM	8.8	5.2	28.8	**100.0***	12.4	11.3	37.6	**100.0***
	FMD	32.3	23.0	53.0	**100.0***	36.5	34.0	61.5	**100.0***
	FMDN	37.7	28.6	58.3	**100.0***	44.0	43.8	76.5	**100.0***
	FMDN-BPFA	**39.6**	**29.6**	**64.2**	**100.0***	**46.5**	**44.4**	**77.1**	**100.0***

Table 1. The success rates of impersonation attack on the face verification task. The first column represents the attacker model. The third to the tenth columns in the second row represent the victim model. $*$ indicates white-box attacks.

4.3.2 Effect of the Layers for Adding Beneficial Perturbation on

To conduct studies on the performance of beneficial perturbations added on different layers, we first count the number of different types of layers in FaceNet, MobileFace, IRSE50 and IR152. We find that convolutional layers, BN layers, and activation layers are the majority of all layers in these models. Therefore, we conduct our experiments on these layers. Specifically, in the process of crafting adversarial examples using FMDN-BPFA, we only add beneficial perturbations to the convolutional layers, BN layers and activation layers, respectively. The experimental results are illustrated in Tab. 2.

Tab. 2 shows that the performance of FMDN-DFANet is only slightly different whether the beneficial perturbations are added on the convolutional layers, BN layers, or activation layers. To figure out the reason for this, we should consider the architectures of the attacker models. In the attacker models, the convolutional layers, BN layer, and activation layers tend to be next to each other. When beneficial perturbations are added on these layers, they have similar effects on the entire model due to the strong linearity of single-layer networks. Therefore, the performance of FMDN-BPFA with beneficial perturbations added on the convolutional layers, BN layers and activation layers is similar.

	layer	FaceNet	MF	IRSE50	IR152
FaceNet	*conv*	**100.0***	34.7	53.9	39.1
	bn	**100.0***	**35.3**	**54.1**	**39.5**
	act	**100.0***	32.7	52.9	37.3
MF	*conv*	**45.4**	**100.0***	99	27.1
	bn	44.4	**100.0***	**99.1**	**28.4**
	act	44.5	**100.0***	**99.1**	27
IRSE50	*conv*	59.3	**99.1**	**100.0***	71.5
	bn	59.1	**99.1**	**100.0***	**72.2**
	act	**59.7**	**99.1**	**100.0***	71.2
IR152	*conv*	**39.6**	**29.6**	**64.2**	**100.0***
	bn	38.3	28.8	63.5	**100.0***
	act	38.4	27.7	63.2	**100.0***

Table 2. The ASR of BPFA when beneficial perturbations are added on different layers. The models in the first column represent the attacker model. The third to the sixth columns in the first row represent the victim models. *conv*, *bn* and *act* in column 2 represent the layers where the beneficial perturbations are added are on the features of the convolutional layers, BN layers and activation layers, respectively. $*$ indicates white-box attacks.

4.4. Evaluation on the Effectiveness of Simulated Gradients

The purpose of adding beneficial perturbations on the features is to be pitted against the change of the loss caused by the adversarial perturbation added on the input image. Therefore, we need to verify whether the beneficial perturbations crafted by simulated gradients can be pitted against the change of loss caused by adversarial perturbations, that

is, whether these beneficial perturbations can increase the loss. To this end, we conduct experiments on the loss of the model with or without adding beneficial perturbations crafted by simulated gradients on the features and the loss of the model after adding beneficial perturbations crafted by simulated gradients under different η. The experimental results are shown in Fig. 5.

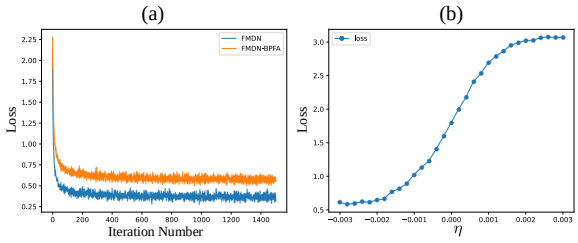

Figure 5. The loss of MF on the LFW dataset. (a) The loss of FMDN and FMDN-BPFA under different iterations. (b) The loss of FMDN-BPFA under different η at the second iteration.

Fig. 5 shows that adding beneficial perturbations crafted by simulated gradients on the features can effectively increase the loss. In addition, subplot (b) of Fig. 5 demonstrates that the loss increases as η increases.

4.5. Interpretability of Beneficial Perturbations Added on the Features

Our experiments show that adding beneficial perturbations on features for feature augmentation can improve the transferability of the crafted adversarial examples. However, the exact form of beneficial perturbations is unknown to us. To address this issue, we study the form of beneficial perturbations added on features to improve the interpretability of them. Specifically, we use FaceNet as the attacker model to craft adversarial examples using the FMDN-BPFA method. At the second iteration, we visualize the beneficial perturbation added on the features of conv2d_2a.conv layer of the model, as shown in Fig. 6 for some channels. We

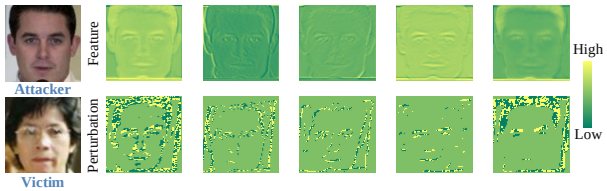

Figure 6. A visualization of the beneficial perturbations added on the features. The two images in the first column are the attacker image and the victim image, respectively. The second to the sixth columns in the first row demonstrate the extracted features, and the second to the sixth columns in the second row demonstrate the beneficial perturbations added on the corresponding features above them.

can see the contours of the person from the images of the beneficial perturbations in Fig. 6, thus suggesting that the beneficial perturbations may have some semantic information. To analyze the reason behind this, we should consider how beneficial perturbations are crafted. Beneficial perturbations are derived by performing a sign operation on the gradients. If an element of a gradient tensor has a nonzero value, the sign operation will give it a value whose absolute value is 1 (corresponding to the dark green and light yellow parts in Fig. 6). If an element of a gradient tensor has a value of 0, the absolute value of this element will be 0 after the sign operation (corresponding to the light green parts in Fig. 6). Therefore, the value of a beneficial perturbation can be seen as a reaction of the magnitude of the gradient to some extent. According to the attribution methods [30] [28], the gradient can be used to assign attribution value (sometimes also called "contribution") to each input feature of a network [2]. The beneficial perturbations added on the features can be seen as the saliency maps [1] of the features. We argue that for the FR model, certain regions in the features (*e.g.*, contour features) are important, leading to the beneficial perturbations added on the features having some semantic information. Similarly, adversarial perturbations added on the features can also be seen as saliency maps containing important information.

5. Conclusion

In this paper, we study the improvement of transferability of adversarial examples on FR. To improve the transferability of adversarial examples on FR, we propose BPFA. BPFA adds beneficial perturbations on the pre-selected features of the FR model to be pitted against the adversarial perturbations crafted in the input images, thereby alleviating the overfitting of the crafted adversarial examples to the surrogate model. We explore the beneficial perturbations added on the features and find that the beneficial perturbations may contain some semantic information. Extensive experiments show that our proposed BPFA can improve the transferability of adversarial face examples.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian J. Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 9525–9536, 2018. 8

[2] Marco Ancona, Enea Ceolini, Cengiz Öztireli, and Markus Gross. Towards better understanding of gradient-based attribution methods for deep neural networks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 8

[3] Karsten M. Borgwardt, Arthur Gretton, Malte J. Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alexander J. Smola. Integrating structured biological data by kernel maximum mean discrepancy. In *Proceedings 14th International Conference on Intelligent Systems for Molecular Biology 2006, Fortaleza, Brazil, August 6-10, 2006*, pages 49–57, 2006. 3

[4] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 1597–1607. PMLR, 2020. 4

[5] Valeriia Cherepanova, Micah Goldblum, Harrison Foley, Shiyuan Duan, John P Dickerson, Gavin Taylor, and Tom Goldstein. Lowkey: Leveraging adversarial attacks to protect social media users from facial recognition. In *International Conference on Learning Representations*, 2021. 1

[6] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 1, 3, 5

[7] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9185–9193, 2018. 6

[8] Yinpeng Dong, Hang Su, Baoyuan Wu, Zhifeng Li, Wei Liu, Tong Zhang, and Jun Zhu. Efficient decision-based black-box adversarial attacks on face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 7714–7722. Computer Vision Foundation / IEEE, 2019. 5

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. 1

[10] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron C. Courville, and Yoshua Bengio. Generative adversarial nets. In Zoubin Ghahramani, Max Welling, Corinna Cortes, Neil D. Lawrence, and Kilian Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2672–2680, 2014. 3

[11] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations (ICLR)*, 2015. 4

[12] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*, pages 9726–9735. Computer Vision Foundation / IEEE, 2020. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 5

[14] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-excitation networks. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7132–7141, 2018. 5

[15] Shengshan Hu, Xiaogeng Liu, Yechao Zhang, Minghui Li, Leo Yu Zhang, Hai Jin, and Libing Wu. Protecting facial privacy: Generating adversarial identity masks via style-robust makeup transfer. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14994–15003, 2022. 1, 3, 5

[16] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 1, 5

[17] Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. LAS-AT: adversarial training with learnable attack strategy. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 13388–13398. IEEE, 2022. 2

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. 5

[19] Ira Kemelmacher-Shlizerman, Steven M. Seitz, Daniel Miller, and Evan Brossard. The megaface benchmark: 1 million faces for recognition at scale. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 4873–4882. IEEE Computer Society, 2016. 1

[20] Stepan Komkov and Aleksandr Petiushko. Advhat: Real-world adversarial attack on arcface face id system. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 819–826, 2021. 1

[21] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*. OpenReview.net, 2017. 3

[22] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net, 2017. 5

[23] Ye Liu, Yaya Cheng, Lianli Gao, Xianglong Liu, Qilong Zhang, and Jingkuan Song. Practical evaluation of adversarial robustness via adaptive auto attack. In *IEEE/CVF*

Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022, pages 15084–15093. IEEE, 2022. 2

[24] Anindya Sarkar, Anirban Sarkar, Sowrya Gali, and Vineeth N. Balasubramanian. Adversarial robustness without adversarial training: A teacher-guided curriculum learning approach. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12836–12848, 2021. 2

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 815–823, 2015. 1, 3, 5

[26] Mahmood Sharif, Sruti Bhagavatula, Lujo Bauer, and Michael K. Reiter. Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In *Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security*, CCS '16, page 1528–1540, New York, NY, USA, 2016. Association for Computing Machinery. 1

[27] Meng Shen, Hao Yu, Liehuang Zhu, Ke Xu, Qi Li, and Jiankun Hu. Effective and robust physical-world attacks on deep learning face recognition systems. *IEEE Trans. Inf. Forensics Secur.*, 16:4063–4077, 2021. 1

[28] Avanti Shrikumar, Peyton Greenside, Anna Shcherbina, and Anshul Kundaje. Not just a black box: Learning important features through propagating activation differences. *CoRR*, abs/1605.01713, 2016. 8

[29] Yi Sun, Xiaogang Wang, and Xiaoou Tang. Deep learning face representation from predicting 10, 000 classes. In *2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, June 23-28, 2014*, pages 1891–1898. IEEE Computer Society, 2014. 3

[30] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks. In Doina Precup and Yee Whye Teh, editors, *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 3319–3328. PMLR, 2017. 8

[31] Teppei Suzuki. Teachaugment: Data augmentation optimization using teacher knowledge. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10894–10904. IEEE, 2022. 2

[32] Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. In *2014 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1701–1708, 2014. 1

[33] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 6105–6114. PMLR, 2019. 1

[34] Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Andreas Steiner, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. Mlp-mixer: An all-mlp architecture for vision. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 24261–24272, 2021. 1

[35] Hao Wang, Yitong Wang, Zheng Zhou, Xing Ji, Dihong Gong, Jingchao Zhou, Zhifeng Li, and Wei Liu. Cosface: Large margin cosine loss for deep face recognition. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1, 3

[36] Shixian Wen and Laurent Itti. Beneficial perturbations network for defending adversarial examples. *CoRR*, abs/2009.12724, 2020. 2

[37] Cho-Ying Wu, Chin-Cheng Hsu, and Ulrich Neumann. Cross-modal perceptionist: Can face geometry be gleaned from voices? In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 10442–10451. IEEE, 2022. 1

[38] Zhirong Wu, Yuanjun Xiong, Stella X. Yu, and Dahua Lin. Unsupervised feature learning via non-parametric instance discrimination. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pages 3733–3742. Computer Vision Foundation / IEEE Computer Society, 2018. 4

[39] Zihao Xiao, Xianfeng Gao, Chilin Fu, Yinpeng Dong, Wei Gao, Xiaolu Zhang, Jun Zhou, and Jun Zhu. Improving transferability of adversarial patches on face recognition with generative models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11840–11849, 2021. 1, 3, 5

[40] Cihang Xie, Zhishuai Zhang, Yuyin Zhou, Song Bai, Jianyu Wang, Zhou Ren, and Alan L. Yuille. Improving transferability of adversarial examples with input diversity. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 2730–2739. Computer Vision Foundation / IEEE, 2019. 6

[41] Xiao Yang, Yinpeng Dong, Tianyu Pang, Hang Su, Jun Zhu, Yuefeng Chen, and Hui Xue. Towards face encryption by generating adversarial identity masks. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3877–3887, 2021. 1, 3, 5

[42] Bangjie Yin, Wenxuan Wang, Taiping Yao, Junfeng Guo, Zelun Kong, Shouhong Ding, Jilin Li, and Cong Liu. Adv-makeup: A new imperceptible and transferable attack on face recognition. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI-21*, pages 1252–1258. International Joint Conferences on Artificial Intelligence Organization, 8 2021. Main Track. 1, 3, 5

[43] Xinyu Zhang, Qiang Wang, Jian Zhang, and Zhao Zhong. Adversarial autoaugment. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. 2

[44] Yaoyao Zhong and Weihong Deng. Adversarial learning with margin-based triplet embedding regularization. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 6548–6557. IEEE, 2019. 3, 6

[45] Yaoyao Zhong and Weihong Deng. Towards transferable adversarial attack against deep face recognition. *IEEE Transactions on Information Forensics and Security*, 16:1452–1466, 2021. 1, 3, 5, 6