Improving Transferability of Adversarial Patches on Face Recognition with Generative Models

Zihao Xiao[1*†] Xianfeng Gao[1,4*] Chilin Fu[2] Yinpeng Dong[1,3] Wei Gao[5‡]
Xiaolu Zhang[2] Jun Zhou[2] Jun Zhu[3†]
[1] RealAI [2] Ant Financial [3] Tsinghua University
[4] Beijing Institute of Technology [5] Nanyang Technological University
zihao.xiao@real.ai, ggxxff@bit.edu.cn, chilin.fcl@antgroup.com, dyp17@mails.tsinghua.edu.cn
gaow0007@ntu.edu.sg, {yueyin.zxl, jun.zhoujun}@antfin.com, dcszj@tsinghua.edu.cn

Abstract

Face recognition is greatly improved by deep convolutional neural networks (CNNs). Recently, these face recognition models have been used for identity authentication in security sensitive applications. However, deep CNNs are vulnerable to adversarial patches, which are physically realizable and stealthy, raising new security concerns on the real-world applications of these models. In this paper, we evaluate the robustness of face recognition models using adversarial patches based on transferability, where the attacker has limited accessibility to the target models. First, we extend the existing transfer-based attack techniques to generate transferable adversarial patches. However, we observe that the transferability is sensitive to initialization and degrades when the perturbation magnitude is large, indicating the overfitting to the substitute models. Second, we propose to regularize the adversarial patches on the low dimensional data manifold. The manifold is represented by generative models pre-trained on legitimate human face images. Using face-like features as adversarial perturbations through optimization on the manifold, we show that the gaps between the responses of substitute models and the target models dramatically decrease, exhibiting a better transferability. Extensive digital world experiments are conducted to demonstrate the superiority of the proposed method in the black-box setting. We apply the proposed method in the physical world as well.

1. Introduction

Deep convolutional neural networks (CNNs) have led to substantial performance improvements on many computer vision tasks. As an important task, face recognition is also greatly facilitated by deep CNNs [17, 23, 4]. Due to their excellent recognition performance, deep face recognition models have been used for identity authentication in security-sensitive applications, *e.g.*, finance/payment, public access, face unlock on smart phones, *etc*.

However, deep CNNs are shown to be vulnerable to adversarial examples at test time [21, 5]. Adversarial examples are elaborately perturbed images that can fool models to make wrong predictions. Early adversarial examples on deep CNNs are indistinguishable from legitimate ones for human observers by slightly perturbing every pixel in an image. Later, [18] proposes adversarial patches, which only perturb a small cluster of pixels. Several works have

(a) Attacker (c) Adv. patch and Adv. example by TAP-b
(b) Target identity (d) Adv. patch and Adv. example by GAP-b

Figure 1. Demonstration of adversarial patches against face recognition models. **(a)** The attacker who wants to impersonate the target identity. **(b)** An image of the target identity. **(c)** The adversarial patch and the corresponding adversarial example generated by the TAP-TIDIM algorithm. **(d)** The adversarial patch and the corresponding adversarial example generated by the proposed GenAP-DI algorithm. The proposed GenAP algorithms use face-like features as perturbations to improve the transferability.

*Equal contributions.
†Corresponding authors.
‡Work done as an intern at RealAI.

shown that the adversarial patches can be made into physical objects to fool deep CNNs in the wild. For example, [9, 19, 24] use adversarial stickers or T-shirts to fool special purpose object detectors. [18] proposes an adversarial eyeglass frame to impersonate another identity against face recognition models. These works show that adversarial patches are physically realizable and stealthy. Using the adversarial patches in the physical world, the attacker can fool a recognition model without accessing the digital input to it, making them an emerging threat to deep learning applications, especially to face recognition systems in security-sensitive scenarios.

Previous works on adversarial patches are developed under the white-box setting [9, 19, 3, 18], where the attacker knows the parameters of the target model, or under the query-based setting [18, 27], where the attacker can make many queries against the target model. But for a black-box model deployed in the wild, both the white-box information and the excessive queries are not easily attainable. In this paper, we focus on evaluating the robustness of face recognition models under the query-free black-box setting, which is a more severe and realistic threat model.

Under the query-free black-box setting, the adversarial attacks based on transferability are widely used. Transfer-based attacks [10] leverage that the adversarial examples for the white-box substitute models are also effective at the black-box target models. Specifically, most adversarial algorithms perform optimization on an adversarial objective specified by the substitute models as a surrogate, to approximate the true (but unknown) adversarial objective on the black-box target models. Existing techniques on improving the transferability of adversarial examples focus on using advanced non-convex optimization [6], data augmentations [26, 7], *etc*. These techniques are originally proposed to generate $\mathcal{L}_p$-norm $(p > 0)$ constrained adversarial examples, and we show that they can be extended to improve the transferability of adversarial patches as well.

However, even though these techniques are extended and applied in the patch setting, we still observe it easy for the optimization to be trapped into local optima with unsatisfactory transferability. First, the transferability is sensitive to initialization of the algorithms. Second, if the perturbation magnitude increases, the transferability first rises and then falls, exhibiting an overfitting phenomenon. The difficulties in escaping solutions of unsatisfactory transferability indicate that the optimization is prone to overfitting the substitute model and new regularization methods are required.

We propose to regularize the adversarial patch by optimizing it on a low-dimensional manifold. Specifically, the manifold is represented by a generative model and the optimization is conducted in its latent space. The generative model is pre-trained on legitimate human face data and can generate diverse and unseen human face images by manipulating the latent vectors to assemble different face features. By optimizing the adversarial objective on this latent space, the adversarial perturbations resemble human face features (see Fig. 1, (d)), on which the predictions of the white-box substitute and the black-box target model are more related. Consequently, the overfitting problem is relieved and the transferability is improved.

Extensive experiments are conducted to show the superiority of the proposed method for black-box attacks on face recognition. We show its effectiveness in the physical world as well. Finally, we extend the proposed method to other tasks, *e.g*., image classification.

2. Related work

2.1. Adversarial patches

Most existing works on adversarial patches are designed for the white-box setting [18, 9, 19, 3, 24] or the query-based black-box setting [18, 27]. This paper focuses on the query-free black-box setting, a realistic assumption on the adversary's knowledge on the target models deployed in the wild [6]. Although some works demonstrate results on query-free attacks [3, 24], their methods are not optimized for this setting and not optimal.

2.2. Transferable adversarial examples

There are many works proposed for improving the transferability of adversarial examples, and most of them are developed under the $\mathcal{L}_p$-norm constrained setting [6, 26, 7]. In contrast, we focus on adversarial patches, a different condition on the adversary's capacity to perturb the visual inputs. Adversarial patches are physically realizable and stealthy, posing threats to target models deployed in the wild. In this paper, we show that while many methods proposed for the $\mathcal{L}_p$-norm constrained setting are useful for the patch setting, they are still prone to overfitting the substitute models and new regularization techniques are required.

2.3. Generative modeling for adversarial examples

Researchers have discovered that using generative models to generate adversarial examples has advantages. For example, efficient attack algorithms are proposed for white-box attacks [25] and query-based attacks [22, 28]. Emerging threat models are studied using generative models as well, *e.g*., unrestricted adversarial examples [20] and semantic adversarial examples [16]. Unrestricted adversarial examples are closely related to adversarial patches, but [20] does not show an improvement of transferability. Although SemanticAdv [16] claims an improvement of transferability in their setting[1], we show that it is sub-optimal in the patch setting. Our work shows how to adequately use generative models to improve the transferability of adversarial patches.

[1] They consider semantic perturbations.

3. Methodology

This section introduces our method of generating adversarial patches on face recognition models with generative models. Sec. 3.1 introduces the attack setting. Sec. 3.2 extends the existing transfer-based attack methods from the $\mathcal{L}_p$-norm constrained setting to the patch setting, and show their problems. Sec. 3.3 elaborates the proposed method.

3.1. Attack setting

Face recognition usually includes face verification and face identification. The former identifies whether two face images belong to the same identity, while the latter classifies an image to a specific identity. For face verification, the similarity between two faces are compared with a threshold to give the prediction. For face identification, the similarity between a face image and those of a gallery set of face images is compared, and the input image is recognized as the identity whose representation is most similar to its.

Let $f(\mathbf{x}) : \mathcal{X} \to \mathbb{R}^d$ denote a face recognition model that extracts a normalized feature representation vector for an input image $\mathbf{x} \in \mathcal{X} \subset \mathbb{R}^n$. Given a pair of face images $\{\mathbf{x}_s, \mathbf{x}_t\}$, the face recognition model estimates the similarity between the two faces by calculating the distance between the feature vectors extracted from the two images

$$\mathcal{D}_f(\mathbf{x}_s, \mathbf{x}_t) = ||f(\mathbf{x}_s) - f(\mathbf{x}_t)||_2^2. \tag{1}$$

And face verification and identification methods are done based on this similarity score $\mathcal{D}_f$ or its simple variants.

An adversary has generally two goals against the face recognition models — dodging and impersonation. Dodging attack aims to generate an adversarial face image that is recognized wrongly, which can be utilized to protect privacy against excessive surveillance. For face verification, the adversary can modify one image from a pair of images belonging to the same identity, to make the model recognize them as different identities. For face identification, the adversary generates an adversarial face image such that it is recognized as any other false identity.

Impersonation attack corresponds to generating an adversarial face image that is recognized as an adversary-specified target identity, which could be used to evade the face authentication systems. For face verification, the attacker aims to find an adversarial image that is recognized as the same identity of another image, while the original images are from different identities. For face identification, the generated adversarial image is expected to be classified as a specific identity.

3.2. Transferable adversarial patch

In the query-free black-box setting, the detailed information of the target model is unknown and an excessive amount of queries are not allowed. The adversarial attacks based on transferability [6, 26] show that, the adversarial examples for some white-box substitute model g can remain adversarial for the black-box target model f. We focus on generating transferable adversarial patches (TAPs).

Suppose g is a white-box face recognition model that is accessible to the attacker, and it can also define a similarity score $\mathcal{D}_g(\mathbf{x}_s, \mathbf{x}_t)$ for face recognition, similar to Eq. (1). An adversary solves the following optimization problem to generate the adversarial patch on the substitute model [18]:

$$\begin{aligned} &\max_{\mathbf{x}} \ \mathcal{L}_g(\mathbf{x}, \mathbf{x}_t), \\ &\text{s.t. } \mathbf{x} \odot (1 - \mathbf{M}) = \mathbf{x}_s \odot (1 - \mathbf{M}), \end{aligned} \tag{2}$$

where $\mathcal{L}_g$ is a differentiable adversarial objective, $\odot$ is the element-wise product, and $\mathbf{M} \in \{0, 1\}^n$ is a binary mask. The constrain emphasizes that only the pixels whose corresponding elements in $\mathbf{M}$ are 1 can be perturbed. Fig. 1 demonstrates how the masks $\mathbf{M}$ control the regions of the adversarial patches. We use $\mathcal{L}_g = \mathcal{D}_g$ for dodging attack and $\mathcal{L}_g = -\mathcal{D}_g$ for impersonation attack, respectively. In this paper, we fix the adversarial loss to fairly compare different techniques operated on the input $\mathbf{x}$.

Existing works on improving the transferability of adversarial examples focus on the $\mathcal{L}_p$-norm constrained setting. We can extend them to the patch setting. The vanilla algorithm is to use the momentum iterative method (MIM) [6] to solve the optimization problem (2). We denote this algorithm as TAP-MIM. Advanced techniques to improve the transferability can be applied, *e.g.* the data augmentations in TI-DIM [26, 8]. The overall algorithm is depicted in Alg. 1 and is denoted as TAP-TIDIM. In the experiment (Sec. 4.2), we show that TAP-TIDIM outperforms TAP-MIM, indicating that methods proposed for the $\mathcal{L}_p$-norm constrained setting might also be useful for the patch setting. Note that the TAP-TIDIM algorithm is similar with using the EoT technique [2] to generate universal and physical-world adversarial patches in the white-box setting [3, 9, 19]

However, even for the more advanced TAP-TIDIM algorithm, it is still difficult for the optimization to escape local optima with unsatisfactory transferability. Specifically, we observe the following two phenomena in our ablation studies (the details are in Sec. 4.3):

- **The transferability is sensitive to the initialization of the optimization.** Note that TAP-TIDIM uses the face image of the attacker to initialize the patch, *i.e.*, $\bar{\mathbf{x}}_0 = \mathbf{x}_s$ (see line 2 of Alg. 1). For the impersonation attack, a simple modification is to use the face image of the target identity to initialize the patch, *i.e.*, $\bar{\mathbf{x}}_0 = \mathbf{x}_t$. The modified algorithm is denoted as TAP-TIDIMv2. Experiments show that, TAP-TIDIMv2 finds solutions with significantly higher transferability than TAP-TIDIM by simply changing the initialization step (see Tab. 2).
- **The transferability degrades when the search space is large.** Specifically, we apply an additional $\mathcal{L}_\infty$-norm con-

Algorithm 1 Transferable Adversarial Patch: TAP-TIDIM

Input: The adversarial objective function $\mathcal{L}_g$; a real face image $\mathbf{x}_s$ of the attacker; a real face images $\mathbf{x}_t$ of the target identity; a binary mask matrix $\mathbf{M}$.

Input: A set of transformations $\mathcal{T}$; the size of perturbation ϵ; learning rate α; iterations N; decay factor μ.

Output: An adversarial image $\mathbf{x}^*$ by solving Eq. (2).

$\mathbf{g}_0 = 0$;

$\bar{\mathbf{x}}_0 = \mathbf{x}_s$;

for $n = 0$ to $N-1$ **do**

  Sample a transformation T from $\mathcal{T}$;

  Blend the adversarial patch to $\mathbf{x}_s$

$$\mathbf{x}_n^* = \mathbf{x}_s \odot (1 - \mathbf{M}) + \bar{\mathbf{x}}_n \odot \mathbf{M};$$

  Input $T(\mathbf{x}_n^*)$ and obtain the loss $\mathcal{L}_g(T(\mathbf{x}_n^*), \mathbf{x}_t)$

  Obtain the gradient $\nabla_{\mathbf{x}=\bar{\mathbf{x}}_n} \mathcal{L}_g(T(\mathbf{x}_n^*))$;

  Convolve the gradient as in [7]

$$\mathbf{W} * \nabla_{\mathbf{x}} \mathcal{L}_g(T(\mathbf{x}_n^*)),$$

  where $\mathbf{W}$ is the Gaussian kernel and $*$ is convolution;

  Update $\mathbf{g}_{t+1}$ as in [6]

$$\mathbf{g}_{n+1} = \mu \cdot \mathbf{g}_n + \frac{\mathbf{W} * \nabla_{\mathbf{x}} \mathcal{L}_g(T(\mathbf{x}_n^*))}{\|\mathbf{W} * \nabla_{\mathbf{x}} \mathcal{L}_g(T(\mathbf{x}_n^*))\|_1};$$

  Update $\bar{\mathbf{x}}_{n+1}$ by applying the sign gradient as

$$\bar{\mathbf{x}}_{n+1} = \mathrm{Clip}_{[\mathbf{x}_0^* - \epsilon, \mathbf{x}_0^* + \epsilon]}\big(\bar{\mathbf{x}}_n - \alpha \cdot \mathrm{sign}(\mathbf{g}_{n+1})\big);$$

end for

return $\mathbf{x}^* = \mathbf{x}_s \odot (1 - \mathbf{M}) + \bar{\mathbf{x}}_N \odot \mathbf{M}$.

strain on the optimization problem (2) to control the size of the search space, i.e., $|\mathbf{x} \odot \mathbf{M}|_\infty \leq \epsilon$. The $\mathcal{L}_\infty$-norm constrain bounds the maximum allowable perturbation magnitude [10]. Our ablation studies show that, when ϵ increases, the transferability first rises and then falls (see Fig. 3).

The aforementioned two phenomena are indicators that the optimization problem (2) has many local optima of unsatisfactory transferability and the adversarial patches are overfitting the substitute model. It is hard to escape from these solutions even though many existing regularization techniques [6, 26, 7] have been applied in TAP-TIDIM. Therefore, we resort to new regularization methods for the patch setting in the following section.

3.3. Generative adversarial patch

We propose to optimize the adversarial patch on a low-dimensional manifold as a regularization to escape from the local optima of unsatisfactory transferability in the optimization problem (2). The manifold poses a specific structure on the optimization dynamics. We consider a good manifold should have the following properties:

1. Sufficient capacity. The manifold should have a sufficient capacity so that the white-box attack on the substitute model is successful.

2. Well regularized. The manifold should be well regularized so that the responses of the substitute models and the target models are effectively related to avoid overfitting the substitute models.

To balance the demands for capacity and regularity, we use the manifold learnt by a generative model, where the generative model is pre-trained on natural human face data. Specifically, let $h(\mathbf{s}) : \mathcal{S} \to \mathbb{R}^n$ denote a pre-trained generative model and $\mathcal{S}$ is its latent space. The generative model can generate diverse and unseen human faces by manipulating the latent vector to assemble different face features, *e.g.*, the color of eyeballs, the thickness of eyebrows, *etc*. We propose to use the generative model to generate the adversarial patch, and optimize the patch through the latent vector. The optimization problem (2) becomes:

$$\begin{aligned} &\max_{\mathbf{s} \in \mathcal{S}} \ \mathcal{L}_g(\mathbf{x}, \mathbf{x}_t), \\ &\text{s.t. } \mathbf{x} \odot (1 - \mathbf{M}) = \mathbf{x}_s \odot (1 - \mathbf{M}), \\ &\qquad \mathbf{x} \odot \mathbf{M} = h(\mathbf{s}) \odot \mathbf{M} \end{aligned} \tag{3}$$

where the second constrain restricts the adversarial patch on the low-dimensional manifold represented by the generative model. When constrained on this manifold, the adversarial perturbations resemble face-like features. We expect that the responses to the face-like features are effectively related for different face recognition models, which improves the transferability of the adversarial patches. This hypothesis will be confirmed in experiments.

The performance of the algorithms depends on the generative model h and the latent space $\mathcal{S}$ that define the manifold. In Sec. 4.4, we perform ablation studies on the architectures, parameters of the generative models h, as well as the latent spaces $\mathcal{S}$. First, the **capacity** of the latent space influences the white-box attack on the substitute model. The latent space of the generative model should have sufficient capacity so that the optimization can find effective adversarial examples on the white-box substitute model. Second, we observe that a generative model, which can generate features semantically related to the adversarial task at hand (*i.e.* face features in our case), can effectively relate the responses from the substitute models and the target models and provide better **regularity**.

A straightforward algorithm to solve the optimization problem (3) is to use the Adam optimizer [15]. We denote this algorithm as GenAP. Similar with TAP-TIDIM, existing techniques [26] can be incorporated. This algorithm is depicted in Alg. 2 and is denoted as GenAP-DI[2].

4. Experiments

In the experiments, we demonstrate the superiority of the proposed GenAP methods in black-box attacks. Sec. 4.1

[2]GenAP-DI is Generative Adversarial Patch with Diversified Inputs

Algorithm 2 Transferable Adversarial Patch: GenAP-DI

Input: The adversarial objective function $\mathcal{L}_g$; a real face image $\mathbf{x}_s$ of the attacker; a real face images $\mathbf{x}_t$ of the target identity; a binary mask matrix $\mathbf{M}$.
Input: A generative model h.
Input: A set of transformations $\mathcal{T}$; iterations N; a gradient-based optimizer, *e.g.*, Adam [15].
Output: An adversarial image $\mathbf{x}^*$ by solving Eq. (3).

Randomly initialize the latent vector $\mathbf{s}_0^* \sim N(0, \mathbb{I})$;
for $n = 0$ to $N - 1$ **do**
  Sample a transformation T from $\mathcal{T}$;
  Blend the adversarial patch to $\mathbf{x}_s$

$$\mathbf{x}_n^* = \mathbf{x}_s \odot (1 - \mathbf{M}) + h(\mathbf{s}_n^*) \odot \mathbf{M};$$

  Input $T(\mathbf{x}_n^*)$ and obtain the loss $\mathcal{L}_g(T(\mathbf{x}_n^*), \mathbf{x}_t)$
  Obtain the gradient $\nabla_{\mathbf{s}=\mathbf{s}_n^*} \mathcal{L}_g(T(\mathbf{x}_n^*))$;
  Update $\mathbf{s}_{n+1}^*$ using the optimizer
end for
return $\mathbf{x}^* = \mathbf{x}_s \odot (1 - \mathbf{M}) + h(\mathbf{s}_N^*) \odot \mathbf{M}$.

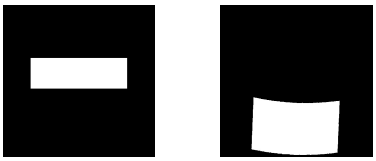
introduces the experimental setting. Sec. 4.2 presents the results in the digital-world attack setting. Sec. 4.3 and 4.4 perform ablation studies on the TAP and the GenAP algorithms respectively. Sec. 4.5 presents the physical-world results.

4.1. Experimental setting

Datasets. Two face image datasets are used for evaluation: LFW [11] and CelebA-HQ [12]. LFW is a dataset for unconstrained face recognition. CelebA-HQ is a human face dataset of high quality. The datasets are used to test the generalization of our methods on both low quality and high quality face images, as the generative models we used are essentially trained on high quality images.

For each dataset, we select face image pairs to evaluate the adversarial algorithms on that dataset. For face verification, we select 400 pairs in dodging attack, where each pair belongs to the same identity, and another 400 pairs in impersonation attack, where the images from the same pair are from different identities. For face identification, we select 400 images of 400 different identities as the gallery set, and the corresponding 400 images of the same identities to form the probe set. Both dodging and impersonation are performed on the probe set. This setting follows [8].

Face recognition models. We study three face recognition models, including FaceNet [17], CosFace [23] and ArcFace [4], which all achieve over 99% accuracies on the LFW validation set. In testing, the feature representation for each input face image is extracted. Then, the cosine similarity between a pair of face images is calculated and compared with a threshold. We first calculate the threshold of each face recognition model by the LFW validation set. It contains $6,000$ pairs of images from same identities $(3,000)$ and different identities $(3,000)$. We choose the threshold of each model that gives the highest accuracy on this validation set. In addition, we also evaluate the performance of our method on commercial face recognition systems—Face++ and Aliyun. Given a pair of face images, a system returns a score indicating their similarity.



(a) Eyeglass frame (b) Respirator

Figure 2. The binary masks $\mathbf{M}$ indicating the regions of the designed patches. (a) An eyeglass frame. (b) A respirator.

Generative models. We study three pre-trained generative models, including ProGAN [12], StyleGAN [13] and StyleGAN2 [14], which can generate face images of high quality. ProGAN has only one latent space. For StyleGANs, we use the $\mathbf{Z}$, $\mathbf{W}$, $\mathbf{W}^+$ and the noise latent spaces [1, 13].

Regions of patches. We use two different regions to generate the patches, an eyeglass frame and a respirator, to show the generalization of the proposed methods to different face regions. The binary masks indicating the regions of these patches are displayed in Fig. 2.

Evaluate Metrics. We use the thresholding strategy and nearest neighbor classifier for face verification and identification, respectively. To evaluate the attack performance, we report the success rate (higher is better) as the fraction of adversarial images that are not classified to the attacker himself by the model in dodging attack, and are misclassified to the desired target identity in impersonation attack.

4.2. Experimental results

In this section, we present the experimental results of adversarial patches for black-box attack in the digital world. We generate adversarial patches using the TAP and the GenAP algorithms, respectively. In impersonation attack, we also use a vanilla baseline of pasting the corresponding face region from the target identity to the attacker (PASTE). We then feed the generated adversarial examples to our local models and commercial APIs to test the success rates of attacks. For the TAP algorithms, we use $\epsilon = 40$, which achieves the best transferability as shown by the ablation study in Sec. 4.3. For the GenAP algorithms, we use StyleGAN2 and its W^+ plus the noise space as a representative, where the results of other generative models and latent spaces are left to ablation studies in Sec. 4.4. We show the results on the face verification task using the eyeglass frame in Tab. 1 (dodging) and 2 (impersonation). Results on face identification, the respirator mask and SemanticAdv [16] are in the supplementary materials, which are qualitatively similar with the results in Tab. 1 and 2.

The results show that the adversarial patches achieve high success rates on the black-box models. First, TAP-TIDIM outperforms TAP-MIM. This shows that applying the existing techniques [6, 26, 7] originally proposed to improve the transferability of $\mathcal{L}_p$-norm constrained adversarial examples against image classification models can be helpful for improving the transferability of adversarial patches against face recognition models as well. Second, the vanilla GenAP significantly outperforms TAP algorithms in most cases (except when using FaceNet as the substitute model for impersonation attack) without bells and whistles. These results show the effectiveness of the proposed regularization method to improve the transferability of the patches. Third, the vanilla GenAP and the more sophisticated GenAP-DI performs similarly, showing that applying additional techniques [26] do not necessarily significantly improve the performance of the GenAP algorithms. Forth, the GenAP algorithms significantly outperform PASTE. This shows that the GenAP algorithms do not naively generating the face features of the target identity, but search the optimal adversarial face features fitting the attacker's own face features. Fifth, our results also show the insecurity of the commercial systems (Face++ and Aliyun) against adversarial patches.

4.3. Ablation study on TAP-TIDIM

This section presents the ablation studies on the TAP algorithms to support the discussions in Sec. 3.2. These ablation studies show that TAP-TIDIM has trouble escaping local optima of unsatisfactory transferability, though many regularization methods have been used [26, 7]. This motivates us to develop new regularization in this paper.

4.3.1 Sensitivity to initialization

The initialization step is the only difference between TAP-TIDIM and TAP-TIDIMv2 when solving problem (2). But TAP-TIDIMv2 shows significantly higher success rates in black-box impersonation attack, as shown in Tab. 2. While the solution of TAP-TIDIMv2 is within the search space of TAP-TIDIM[3], TAP-TIDIM cannot find it and is trapped into local optima with significantly worse transferability.

4.3.2 Sensitivity to ϵ

The hyperparameter ϵ in TAP-TIDIM can control the upper bound for the perturbation magnitudes of the adversarial patches, which is an indicator of the size of the search space. The larger the ϵ, the larger the search space. Fig. 3 shows that, as the upper bound ϵ increases, the success rates on the black-box models first rise and then fall. When ϵ is small, the transferability benefits from the larger search space by finding more effective adversarial examples against the substitute model. But when ϵ is large, the adversarial patches overfit the substitute model and are trapped into poor local optima. The transferability of TAP-TIDIM reaches the optimality around $\epsilon = 40$, which is used in Sec. 4.2.

[3]Strictly speaking, the solution of TAP-TIDIMv2 is within the search space of TAP-TIDIM ($\epsilon = 255$) and the results in Tab. 2 are for TAP-TIDIM ($\epsilon = 40$). Nevertheless, the TAP-TIDIM ($\epsilon = 40$) outperforms TAP-TIDIM ($\epsilon = 255$) as shown in Fig. 3 and our conclusion holds.

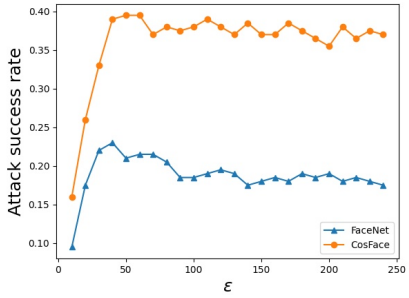


Figure 3. The success rates of TAP-TIDIM on the black-box models first rise and then fall when the maximal perturbation magnitude ϵ increases. This indicates that the adversarial patches are overfitting the substitute model. The results are black-box impersonation attack on FaceNet and CosFace under the face verification task. The adversarial examples are generated against ArcFace by restricting the adversarial patches to an eyeglass frame region. 200 image pairs from the LFW dataset are used.

4.4. Ablation study on GenAP

In this section, we present the ablation studies on the GenAP algorithms, which show how the regularity and the capacity of the manifold influence the transferability of the generative adversarial patches. This section use the GenAP algorithm and the eyeglass frame mask for experiments.

4.4.1 Parameters of the generative models

We use the StyleGAN2 with different parameters, including the parameters that are randomly initialized (Rand), trained on the LSUN car dataset (CAR) and trained on the FFHQ face image dataset (FFHQ). We use $\mathbf{W}^+$ plus the noise space [1, 13] as the latent space. Tab. 3(a) shows that using the randomly initialized model cannot find effective adversarial examples even in the white-box case. While the StyleGAN2 trained on CAR is effective at white-box attack, their transferability to black-box models is poor. The transferability of GenAP is better than the TAP algorithms (c.f., Tab. 2) only when the generative model is trained on human face dataset. This phenomenon indicates that using face-like features as perturbations is important for bridging the gap between the substitute and the target face recognition models to improve transferability in the GenAP methods.

	Attack	CelebA-HQ					LFW				
		ArcFace	CosFace	FaceNet	Face++	Aliyun	ArcFace	CosFace	FaceNet	Face++	Aliyun
ArcFace	TAP-MIM	0.9875*	0.1800	0.6475	0.0000	0.1800	0.9850*	0.1475	0.4275	0.0075	0.1850
	TAP-TIDIM	**1.0000***	0.2975	0.7050	0.1625	0.7350	**1.0000***	0.2500	0.5200	0.1550	0.5850
	GenAP	0.9975*	**0.6100**	**0.9375**	**0.7975**	**0.9900**	0.9975*	**0.4850**	**0.8725**	**0.7450**	**0.9850**
	GenAP-DI	**1.0000***	0.5050	0.8600	0.6600	0.9800	**1.0000***	0.4050	0.7725	0.6250	**0.9850**
CosFace	TAP-MIM	0.0475	0.9925*	0.6950	0.0025	0.4250	0.0125	0.9975*	0.4950	0.0075	0.5350
	TAP-TIDIM	0.0025	**1.0000***	0.4925	0.0350	0.6550	0.0100	**1.0000***	0.3550	0.0425	0.5650
	GenAP	**0.5375**	0.9975*	**0.9550**	**0.5675**	**1.0000**	**0.3650**	**1.0000***	**0.9275**	**0.5600**	**0.9900**
	GenAP-DI	0.3650	**1.0000***	0.9150	0.3675	0.9950	0.2175	**1.0000***	0.9025	0.3900	0.9800
FaceNet	TAP-MIM	0.0250	0.2100	0.9900*	0.0025	0.2350	0.0075	0.1475	0.9675*	0.0050	0.3400
	TAP-TIDIM	0.0025	0.1525	**0.9975***	0.0775	0.7050	0.0025	0.1025	**1.0000***	0.0850	0.6300
	GenAP	**0.3575**	**0.3475**	**0.9975***	**0.5675**	**1.0000**	**0.1950**	**0.2450**	0.9950*	**0.5550**	**0.9950**
	GenAP-DI	0.2100	0.1950	**0.9975***	0.4400	0.9800	0.0950	0.1500	0.9925*	0.3625	0.9700

Table 1. The success rates of black-box dodging attack on FaceNet, CosFace, ArcFace, Face++ and Aliyun in the digital world under the face verification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to an eyeglass frame region. * indicates white-box attacks.

	Attack	CelebA-HQ					LFW				
		ArcFace	CosFace	FaceNet	Face++	Aliyun	ArcFace	CosFace	FaceNet	Face++	Aliyun
	PASTE	0.4725	0.3700	0.3000	0.2425	0.0900	0.4150	0.3100	0.1825	0.1775	0.0250
ArcFace	TAP-MIM	0.9900*	0.3100	0.2325	0.1250	0.0400	**1.0000***	0.2600	0.1875	0.0425	0.0100
	TAP-TIDIM	**1.0000***	0.3675	0.2725	0.1900	0.0650	**1.0000***	0.3125	0.2025	0.0800	0.0150
	TAP-TIDIMv2	**1.0000***	0.4975	0.3425	0.2525	0.0750	**1.0000***	0.4225	0.2350	0.1250	0.0050
	GenAP	**1.0000***	**0.5825**	**0.4625**	0.3425	**0.1700**	**1.0000***	**0.5000**	**0.4000**	**0.2125**	**0.1000**
	GenAP-DI	**1.0000***	0.5300	0.4100	**0.3500**	0.1450	**1.0000***	0.4325	0.3275	0.1825	0.0550
CosFace	TAP-MIM	0.4275	0.9900*	0.3125	0.1425	0.0450	0.3250	0.9850*	0.2525	0.0550	0.0200
	TAP-TIDIM	0.4550	**1.0000***	0.3725	0.2000	0.0750	0.2725	**1.0000***	0.2700	0.0750	0.0150
	TAP-TIDIMv2	0.5250	**1.0000***	0.4175	0.2650	0.0950	0.3625	**1.0000***	0.3225	0.1325	0.0100
	GenAP	**0.6575**	**1.0000***	0.5250	**0.3500**	**0.2000**	**0.5350**	**1.0000***	0.4600	**0.2100**	0.0700
	GenAP-DI	0.6325	**1.0000***	**0.5325**	0.3275	0.1900	0.4975	**1.0000***	**0.4650**	0.2000	**0.1000**
FaceNet	TAP-MIM	0.2400	0.2025	0.8300*	0.1150	0.0450	0.1425	0.1850	0.8375*	0.0300	0.0100
	TAP-TIDIM	0.1800	0.2200	**0.9775***	0.1175	0.0450	0.0925	0.1925	0.9800*	0.0300	0.0050
	TAP-TIDIMv2	0.3000	0.3300	**0.9775***	0.1725	0.0500	0.1650	0.2450	0.9825*	0.0650	0.0150
	GenAP	0.2750	0.2450	0.9025*	0.1250	0.0600	0.2450	0.2425	0.9200*	0.0900	**0.0350**
	GenAP-DI	0.2175	0.2025	0.9650*	0.1150	0.0500	0.1675	0.1850	**0.9850***	0.0550	**0.0350**

Table 2. The success rates of black-box impersonation attack on FaceNet, CosFace, ArcFace, Face++ and Aliyun in the digital world under the face verification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to an eyeglass frame region. * indicates white-box attacks.

4.4.2 Architectures of the generative models

We use three different generative models, including ProGAN [12], StyleGAN [13] and StyleGAN2 [14]. These generative models differ in their network architectures and can generate human face images with higher and higher quality. For the StyleGANs, we use the $\mathbf{W}^+$ latent plus the noise space [1, 13]. The results are shown in Tab. 3(b). The performance of GenAP depends on the architectures of the generative models. Even though ProGAN is trained on human face images, it is difficult to find effective adversarial examples in its latent space, even in the white-box case. Both StyleGANs achieve high success rates against the black-box models. These phenomena indicate that the style-based decoder in StyleGANs might be important for the GenAP algorithms to find effective adversarial examples.

4.4.3 Latent spaces of the generative models

We use different latent spaces for the StyleGAN2, including the $\mathbf{Z}$, the $\mathbf{W}$, the $\mathbf{W}^+$ and the noise spaces [1, 13]. The $\mathbf{W}^+$ is more flexible than the $\mathbf{Z}$, the $\mathbf{W}$ and the noise spaces with much more degrees of freedom. Tab. 3(c) shows that, the performance on the $\mathbf{W}$ and the $\mathbf{W}^+$ spaces is substantially higher than that on the $\mathbf{Z}$ and the noise spaces. The optimizations in the $\mathbf{Z}$ and the noise spaces cannot find effective adversarial patches even in the white-box case.

		CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
(a) Parameters							
ArcFace	RAND	0.2625*	0.0100	0.0100	0.2575*	0.0050	0.0200
	CAR	0.9850*	0.0950	0.1075	0.9825*	0.0625	0.0800
	FFHQ	**1.0000***	**0.5828**	**0.4625**	**1.0000***	**0.5000**	**0.4000**
(b) Architectures							
ArcFace	ProGAN	0.6750*	0.2375	0.2125	0.6475*	0.1650	0.1450
	StyleGAN	**1.0000***	0.5500	0.4250	**1.0000***	0.4750	0.3475
	StyleGAN2	**1.0000***	**0.5825**	**0.4625**	**1.0000***	**0.5000**	**0.4000**
(c) Latent spaces							
ArcFace	$\mathbf{Z}$	0.4175*	0.1125	0.0800	0.4050*	0.1200	0.0900
	$\mathbf{W}$	0.9575*	0.4775	0.3825	0.9425*	0.4300	0.3900
	$\mathbf{W}^+$	**1.0000***	0.5750	**0.4625**	**1.0000***	0.4925	0.3825
	Noise	0.2075*	0.0500	0.0450	0.1250*	0.0425	0.0250
	$\mathbf{W}^+$ + Noise	**1.0000***	**0.5825**	**0.4625**	**1.0000***	**0.5000**	**0.4000**

Table 3. The success rates of black-box impersonation attack when the architectures, the parameter and the latent space are changed in the proposed GenAP algorithm. The adversarial examples are generated against ArcFace by restricting the adversarial patches to an eyeglass frame region, and are tested on FaceNet, CosFace and ArcFace in the digital world under the face verification task. * indicates white-box attacks. The ablation studies are on (a) the parameters (RAND, CAR and FFHQ) of the StyleGAN2, (b) the architectures of the generative model (ProGAN, StyleGAN, StyleGAN2) and (c) the latent space ($\mathbf{Z}$, $\mathbf{W}$, $\mathbf{W}^+$ and noise) used by StyleGAN2 trained on FFHQ.

4.5. Physical-world experiment

In this section, we verify that the adversarial patches generated by the proposed GenAP algorithms are physically realizable, and their superiority is retained after printing and photographing. Specifically, we select a volunteer as the attacker and 3 target identities (one male and two females) from the CelebA-HQ dataset. For each target identity, we generate an eyeglass frame for the attacker to impersonate that identity. After the attacker wears the adversarial eyeglass frame, we take a video of him from the front and randomly select 100 video frames. The video frames are used for face verification. We evaluate the transferability of the patches using the cosine similarities. The higher the similarity, the better the transferability. Results in Fig.4 show that the patches generated by the proposed GenAP-DI retain high transferability even after printing and photographing.

4.6. Extra experiments

In the supplementary material, we also compare the proposed GenAP methods with an existing method of using face-like features as adversarial perturbations, SemanticAdv [16], and extend the GenAP methods to other tasks, *e.g.*, image classification. First, we explain why SemanticAdv is sub-optimal in the patch setting. Second, we show the generalizability of the proposed GenAP methods to other recognition tasks.

5. Conclusion

In this paper, we evaluate the robustness of face recognition models against adversarial patches in the query-free black-box setting. Firstly, we extend existing techniques from the $\mathcal{L}_p$-constrained ($p > 0$) setting to the patch setting, yielding TAP algorithms to generate transferable adversarial patches. However, several experimental phenomena indicate that it is hard for the TAP algorithms to escape from local optima with unsatisfactory transferability. Therefore, we propose to regularize the adversarial patches on the manifold learnt by generative models pre-trained on human face images. The perturbations in the proposed GenAP algorithms resemble face-like features, which is important for reducing the gap between the substitute and the target face recognition models. Experiments confirm the superiority of the proposed methods.

	target 1	target 2	target 3
CosFace			
TAP-TIDIMv2	16.8($\pm$1.6)	18.2($\pm$1.3)	4.1($\pm$2.0)
GenAP-DI	**27.2**($\pm$2.2)	**21.4**($\pm$2.3)	**12.0**($\pm$2.4)
FaceNet			
TAP-TIDIMv2	**23.3**($\pm$2.4)	$-$3.9($\pm$3.0	32.2($\pm$2.1)
GenAP-DI	22.8($\pm$2.6)	**24.6**($\pm$2.0)	**33.4**($\pm$1.8)

Table 4. The cosine similarties between the attacker wearing the adversarial eyeglass frame and three different target identities in the physical-world. The target identities are randomly drawn from CelebA-HQ. The adversarial eyeglass frame is crafted by the TAP-TIDIMv2 and the proposed GenAP-DI algorithms on ArcFace, and is tested on CosFace and FaceNet.

References

[1] Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent

space? In *Proceedings of the IEEE international conference on computer vision*, pages 4432–4441, 2019. 5, 6, 7

[2] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 3

[3] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 2, 3

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019. 1, 5

[5] Yinpeng Dong, Qi-An Fu, Xiao Yang, Tianyu Pang, Hang Su, Zihao Xiao, and Jun Zhu. Benchmarking adversarial robustness on image classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 321–331, 2020. 1

[6] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018. 2, 3, 4, 6

[7] Yinpeng Dong, Tianyu Pang, Hang Su, and Jun Zhu. Evading defenses to transferable adversarial examples by translation-invariant attacks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4312–4321, 2019. 2, 4, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7714–7722, 2019. 3, 5

[9] Ivan Evtimov, Kevin Eykholt, Earlence Fernandes, Tadayoshi Kohno, Bo Li, Atul Prakash, Amir Rahmati, and Dawn Song. Robust physical-world attacks on machine learning models. *arXiv preprint arXiv:1707.08945*, 2(3):4, 2017. 2, 3

[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 4

[11] 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. 5

[12] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017. 5, 7

[13] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 5, 6, 7

[14] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020. 5, 7

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4, 5

[16] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020. 2, 5, 8

[17] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 1, 5

[18] 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*, pages 1528–1540, 2016. 1, 2, 3

[19] Dawn Song, Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Florian Tramer, Atul Prakash, and Tadayoshi Kohno. Physical adversarial examples for object detectors. In *12th {USENIX} Workshop on Offensive Technologies ({WOOT} 18)*, 2018. 2, 3

[20] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018. 2

[21] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[22] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 742–749, 2019. 2

[23] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5265–5274, 2018. 1, 5

[24] Zuxuan Wu, Ser-Nam Lim, Larry Davis, and Tom Goldstein. Making an invisibility cloak: Real world adversarial attacks on object detectors. *arXiv preprint arXiv:1910.14667*, 2019. 2

[25] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 2

[26] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2730–2739, 2019. 2, 3, 4, 6

[27] Chenglin Yang, Adam Kortylewski, Cihang Xie, Yinzhi Cao, and Alan Yuille. Patchattack: A black-box texture-based attack with reinforcement learning. In *European Conference on Computer Vision*, pages 681–698. Springer, 2020. 2

[28] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. *arXiv preprint arXiv:1710.11342*, 2017. 2