Multimedia Tools and Applications (2021) 80:855–875
https://doi.org/10.1007/s11042-020-09604-z

Attacks on state-of-the-art face recognition using attentional adversarial attack generative network

Lu Yang[1] · Qing Song[1] · Yingqi Wu[1]

Received: 10 February 2020 / Revised: 9 August 2020 / Accepted: 12 August 2020 /
Published online: 5 September 2020

Abstract
With the broad use of face recognition, its weakness gradually emerges that it is able to be attacked. Therefore, it is very important to study how face recognition networks are subject to attacks. Generating adversarial examples is an effective attack method, which misleads the face recognition system through obfuscation attack (rejecting a genuine subject) or impersonation attack (matching to an impostor). In this paper, we introduce a novel GAN, Attentional Adversarial Attack Generative Network (A^3GN), to generate adversarial examples that mislead the network to identify someone as the target person not misclassify inconspicuously. For capturing the geometric and context information of the target person, this work adds a conditional variational autoencoder and attention modules to learn the instance-level correspondences between faces. Unlike traditional two-player GAN, this work introduces a face recognition network as the third player to participate in the competition between generator and discriminator which allows the attacker to impersonate the target person better. The generated faces which are hard to arouse the notice of onlookers can evade recognition by state-of-the-art networks and most of them are recognized as the target person.

Keywords Face recognition · Generative adversarial networks · Adversarial attack

1 Introduction

Neural networks are widely used in different tasks in the society which is profoundly changing our life [15, 20, 61]. A good algorithm, adequate training data, and computing power

✉ Qing Song
priv@bupt.edu.cn

Lu Yang
soeaver@bupt.edu.cn

Yingqi Wu
wuyqq@bupt.edu.cn

[1] Pattern Recognition and Intelligence Vision Lab, Beijing University of Posts and Telecommunications, Beijing, China

make neural networks supersede humans in many tasks, especially face recognition tasks [9, 21, 61]. Face recognition can be used to determine which one the face images belong to or whether the two face images belong to the same one. Applications based on this technology are gradually adopted in some important tasks, such as identity authentication in a railway station and for payment. Unfortunately, it has been shown that face recognition networks can be deceived inconspicuously by mildly changing inputs maliciously. The changed inputs are named adversarial examples that implement adversarial attack on networks [50, 51]. Szegedy et al. [57] present that adversarial attack can be implemented by applying an imperceptible perturbation which is hard to be observed for human eyes for the first time. Following the work of Szegedy, many works focus on how to craft adversarial examples to attack neural networks [12, 29, 38, 53]. Neural networks are gradually under suspicion. The works on adversarial attack can promote the development of neural networks. Akhtar et al. [1] review these works' contributions in the real-world scenarios. Illuminated by predecessor's works, we also do some research about the adversarial attacks (Fig. 1).

In this work, we explore the feature representation of different faces. Most of the adversarial attacks aim at misleading classifier to a false label. Existing works produce perturbation on the images [16, 40, 57], do some makeup or add eyeglass / hat / occlusions [17, 50, 51] to faces. And their adversarial examples are fixed by the algorithms which are not flexible for attacks. Most of the adversarial attacks can not accept any images as inputs. The method manipulating the intensity of input images directly is intensity-based. Our work uses the geometry-based method to generate adversarial examples that adjust a tiny part of faces imperceptibly to do the attack. Our goal is to generate face images that are similar to the original images but can be classified as the target person by imitating the feature representation of the target person and which can accept any faces as inputs. In other words, we want to generate face samples, which are similar to the x sample in image space, similar to the y sample in feature space, and x and y can be arbitrarily specified. Due to the development of intensity-based attacks, many works present their corresponding defense methods. To promote the development of the adversarial attacks, we propose a more complex and novel way to generate adversarial examples. For these purposes, we present A^3GN to produce the fake image whose appearance is similar to the origin but is able to be classified as the target person.

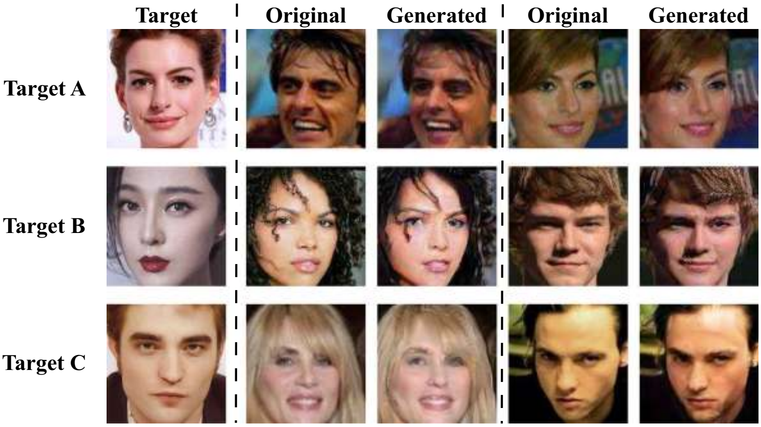

Fig. 1 Adversarial attack results in our work. The first column is the target face. The 2nd and 4th columns are the original images and the rest are the generated images. Given target images, our work is to generate images similar to the original faces but classified as the target person

In the face verification domain, whether the two faces belong to one person is based on the cosine distance between the feature map in the last layer not based on the probability for each category. So A^3GN pays more attention to the exploration of feature representation for faces. To get the instance information, we introduce an attentional variational autoencoder to get the latent code from the target face, and meanwhile, attentional modules are provided to capture more feature representation and facial dependencies of the target face. For adversarial examples, A^3GN adopts two discriminators – one for estimating whether the generated faces are real called *sample discriminator*, another for estimating whether the generated faces can be classified as the target person called *identity discriminator*. Meanwhile, cosine loss is introduced to promise that the fake images can be classified as the target person by the target model. In the black-box scenario, we introduce a substitute network as the identity discriminator to do the feature estimation of different target networks. Feature estimation is a crucial part of the black-box attack which helps our substitute network to learn the feature representation of multiple face recognition networks to do the black-box attack. Our main contributions can be summarized into four-fold:

- We focus on a novel way of attacking against state-of-the-art face recognition networks. They will be misled to identify someone as the target person not misclassify inconspicuously in face verification according to the feature map, not the probability.
- GAN is introduced to generate adversarial examples different from traditional intensity-based attacks. Meanwhile, this work presents a new GAN named A^3GN to generate adversarial examples that are similar to the origins but have the same feature representation as to the target face.
- We introduce a substitute network to do the feature estimation of face recognition networks which achieves the purpose of black-box attack in face verification.
- Good performance of A^3GN can be shown by a set of evaluation metrics in physical likeness, Similarity Score, and accuracy of recognition.

2 Related work

2.1 Face recognition

We witness the great development and success of convolutional neural networks in face recognition so far. With the development of advanced architectures and discriminative learning approaches, face recognition performance has been boosted to an unprecedented level [9, 35, 55, 56, 58, 61]. Face recognition can be categorized as face verification and face identification. In our work, we focus on face verification which determines whether a pair of faces belong to the same person and the latter classifies a face to a specific identity. To learn discriminative deep face representation, there are two major types of approaches are widely studied: softmax-free methods and softmax-based methods.

The main purpose of softmax-free methods is to distinguish identities in the feature space with the guidance of distances among samples. Siamese network [7] utilized the contrastive loss to learn contrastive representations. In Siamese networks, two facial images are successively fed into two networks to obtain their respective embeddings, and the contrastive loss penalizes the distance between two embeddings when the input images are paired. Florian et al. proposed the FaceNet with Triplet loss [48], which explicitly maximizes the inter-class distance and meanwhile minimizes the intra-class distance, where a margin term is used to determine the decision boundaries between positive and negative pairs. These methods and their improved versions have achieved good results in face recognition, and have also been applied to metric learning [26, 47, 67], fine-grained visual recognition [13, 32, 69] and person re-identification [22, 33, 36].

Based on the principle of image classification, softmax-based methods directly use identity labels as category information to supervise the face recognition networks. To directly enhance the feature discrimination, several softmax-based loss functions [9, 34, 35, 60, 63] have been proposed in recent years. To reduce intra-class variations, Center Loss [63] was proposed by Wen et al., which learns centers for each identity to emphasize the intra-class compactness in the embedding manifold. SphereFace [35] proposed the angular margin softmax loss (A-Softmax loss) which focuses on inter-class decision boundary to improve softmax loss. A-Softmax loss introduces an angular margin between the ground truth class and other classes and uses a multiplier to impose a multiplicative angular margin to the original decision boundaries during the training stage. However, A-Softmax loss is usually unstable and the optimal parameters are hard to determinate. To enhance the stability of A-Softmax loss, Wang et al. design an additive margin softmax loss (AM-Softmax) [60] which replaces angular margin by cosine margin to stabilize the optimization and have achieved promising performance. Deng et al. develop an additive angular margin softmax loss (Arc-Softmax) [63], which has a more clear geometric interpretation.

2.2 Generative adversarial networks

Generative Adversarial Networks (GANs), originally introduced by Goodfellow et al. [15], is one of the most promising methods for unsupervised learning in the complex distribution in recent years. GANs are a framework of artificial intelligence algorithms implemented by a system of two neural networks contesting with each other in a zero-sum game framework [2, 18]. This framework for estimating generative models via an adversarial process corresponds to a minimax two-player game [15]. GANs have achieved great performance and impressive results in image generation [10, 44], style transfer [14, 28, 59], image-to-image translation [27, 70, 71] and representation learning [37, 44, 45]. Despite this tremendous success, the training of GANs is known to be unstable and sensitive to the choices of hyper-parameters. Several studies have attempted to stabilize the GAN training dynamics and increase the sample diversity by modifying the learning objectives and dynamics [2, 46], designing new network architectures [30, 44], adding regularization methods [18, 39] and introducing heuristic tricks [43]. Some modified versions of GAN also demonstrate performance advantages in many scenarios [6, 43, 70, 71]. Most works utilize conditional variables such as attributes [6, 66]. CycleGAN [70] preserves key attributes between the input and the translated images by a cycle consistency loss which has received a good improvement in unpaired image-to-image translation. Conditional VAEs [52] have shown good performance for image-to-image translation which learns a mapping from input to output image. In [71], cVAE-GAN and cLR-GAN are used to learn a low-dimensional latent code and then map from a high-dimensional input to a high-dimensional output.

In our work, we use conditional variational autoencoder GAN to learn the feature representation of the target person for generating adversarial examples from any faces to attack face recognition networks.

2.3 Adversarial attack

With remarkable accuracy, neural networks get access to many important domains in society, the security problem of neural networks has become a critical problem. Szegedy et al. [57] reveal the perturbation which can fool DNN for the first time. Moosavi-Dezfooli et al. [40] introduced the Deepfool and demonstrate that 'universal perturbation' can fool the classifier by any image in most type of models, which could fool neural networks with a universal perturbation on images with high probability. Goodfellow et al. [16] indicates that the intrinsic reason for the adversarial attack is the linearity and high-dimensions of inputs, and propose a more time-saving practical method (FGSM) to generate adversarial examples by performing one-step gradient update along the direction of the sign of gradient at each pixel. Su et al. [54] present a method to generate one-pixel adversarial perturbations to attack models using differential evolution in an extremely specific scenario. Song et al. [53] propose unrestricted adversarial examples, a threat model where the attackers are not restricted to small norm bounded perturbations. Many works are proposed to explore more imperceptible adversarial examples to attack neural networks efficiently [4, 12, 25, 29, 38, 41, 64].

In literature, studies on generating adversarial examples in the face recognition domain are relatively limited. Bose et al. [3] craft adversarial examples by solving constrained optimization so that face detector can not detect faces. Sharif et al. [50] propose a method focusing on facial biometric systems which can be widely used in surveillance and access control. However, these adversarial attack methods rely on white-box manipulations of face recognition models, which is impractical in real-world scenarios. In [11], Dong et al. proposed an evolutionary optimization method for generating adversarial faces in black-box settings, which does not need to access the specific structure and parameters of the neural network. However, they require at least 1,000 queries to the target face recognition system before a realistic adversarial face can be synthesized. Whether a white-box attack or a black-box attack on the face recognition network, existing methods have many limitations, some of which need to change the appearance of the samples [50], and some of which can not arbitrarily control the results of the attack [11, 65].

In this paper, we focus on generating quasi-imperceptible adversarial examples to do white-box, black-box, and targeted attacks. In other words, we want to generate face samples, which are similar to the x sample in image space, similar to the y sample in feature space, and x and y can be arbitrarily specified.

3 Attentional adversarial attack generative network

In this section, we present the Attentional Adversarial Attack Generative Network (A^3GN), and introduce the overall structure (Section 3.1) and objective functions (Section 3.2). The pipeline of the A^3GN contains two different discriminators, *sample discriminator* and *identity discriminator* (Section 3.3) which help the A^3GN to generate the adversarial examples. In order to make the generated adversarial example similar to the target face in feature space and the original face in image space, we need a stronger feature extractor. For variational autoencoder, we hope to enhance its feature encoding ability for the target face and increase the similarity between the generated face and the original face in feature space. For the generator, we consider enhancing the quality of the generated images so that the generated faces are closer to the original ones visually, that is, to increase the similarity between the two in the image space. Based on the above considerations, we propose the Attentional VAE and Attentional Generator in Section 3.4 respectively. Experiments (Table 5) shows that the proposed blocks can significantly improve the quality of generated images, and can better simulate the distribution of targets in both image space and feature space. Section 3.5 will introduce the operation, *feature estimation*, we used in the black-box attack to improve the performance.

3.1 Overall

For exploring the feature distribution of different faces, we use VAE to capture the instance information of different faces for the generator to produce the adversarial examples. Given a target image y, using an encoding function E learns a latent code z of y, $E(y) \rightarrow z$. Generator G_1 combines z and an input image x to synthesize the output $\hat{x}$, $G_1(x, z) \rightarrow \hat{x}$, which is the adversarial example. x, y, $\hat{x}$ are all 112×112 RGB images. Here, z is the latent code of the target image y which contains the instance information of the target person and helps the generator G_1 to generate the adversarial example $\hat{x}$, thus $\hat{x}$ contains the instance information of the target image y and the geometric information such as the appearance of input image x. G_2 aims at reconstructing the original image x which guarantees that $\hat{x}$ has the same geometric appearance of x.

Sample discriminator D_1 determines whether $\hat{x}$ is real or not. Identity discriminator D_2 determines whether $\hat{x}$ can be recognized as the target person or not. The overview of A^3GN is shown in Fig. 2. The backbone of A^3GN is cVAE-GAN actually.

3.2 Objective functions

To make the generated images $\hat{x}$, $G_1(x, z)$, indistinguishable from real images, we adopt an adversarial loss [18]:

$$\begin{aligned}\mathcal{L}_{adv} &= \mathbb{E}_x[logD_1(x)] - \mathbb{E}_{x,z}[D_1(G_1(x,z))] \\ &-\lambda_{gp}\mathbb{E}_{x'}[(\left\|\nabla_{x'}D_1(x')\right\|_2 - 1)^2],\end{aligned} \tag{1}$$

where generated image $G_1(x, z)$ learns the latent code z from the target image y, while sample discriminator D_1 tries to distinguish $G_1(x, z)$ between real and fake image. Sample

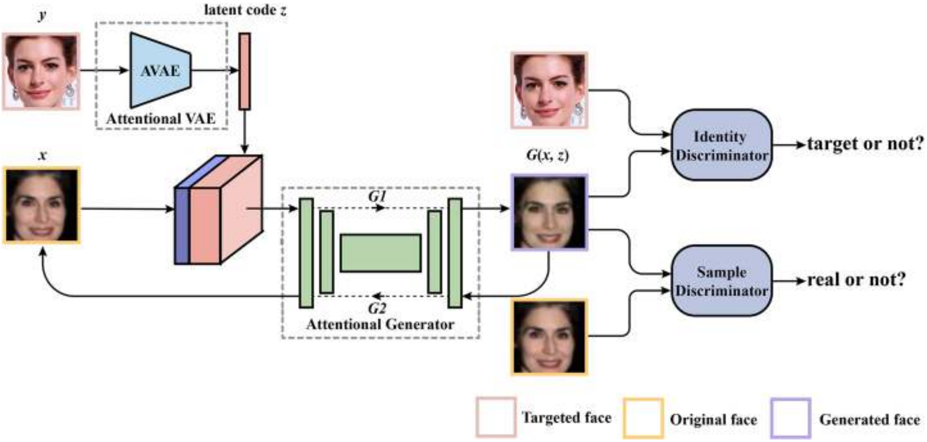

Fig. 2 Overview of A^3GN. *Attentional AVE (AVAE)* captures the latent code z from target face y. *Attentional Generator* is a cycle generator consisting of G_1 and G_2. G_1 is sent into *sample discriminator* to determine whether it is a real image or not with x, and sent into *identity discriminator* to determine whether it can be classified as the identity of target person or not with y. G_2 aims at reconstructing the original image x which guarantees that $\hat{x}$ has the same geometric appearance of x

discriminator D_1 tries to maximize $D_1(x)$ which is opposite to the generator G_1. And x' is sampled between a pair of a real and a generated images. λ_{gp} is set to 10.

Further, the latent code is encouraged to be close to a random Gaussian [71]:

$$\mathcal{L}_{KL}(E) = \mathbb{E}_y[\mathcal{D}_{KL}(E(y)\|\mathcal{N}(0, I))], \tag{2}$$

where $\mathcal{D}_{KL}(p\|q) = -\int p(z)log\frac{p(z)}{q(z)}dz$.

To preserve the content of the input images x, while changing instance-level information and a part of feature representation of the inputs, we introduce a cycle-consistency loss [70] to the generator as reconstruction loss:

$$\mathcal{L}_{rec} = \mathbb{E}_{x,z}[\|x - G_2(G_1(x, z))\|_1], \tag{3}$$

where G_2 is used to take in the generated image $G_1(x, z)$ as input and reconstruct the original image x. The reconstruct loss adopts the ℓ_1 norm. Here, G_1 and G_2 are two different generators with inputs of different dimensions.

To guarantee that $\hat{x}$ can be classified as y by D_2, we adopt a cosine loss, defined as:

$$\begin{aligned}\mathcal{L}_{cos} &= 1 - SIM(y, G_1(x, z)) = 1 - \cos\theta \\ &= 1 - \mathbb{E}_{x,y,z}[\frac{D_2(y) \cdot D_2(G_1(x, z))}{\|D_2(y)\| \cdot \|D_2(G_1(x, z))\|}],\end{aligned} \tag{4}$$

where D_2 is the identity discriminator, and $D_2(y)$ and $D_2(G_1(x, z))$ mean the feature representation of y and $G_1(x, z)$. Minimizing cosine loss can minimize the difference between generated image $G_1(x, z)$ and target image y in space which brings benefit to generating adversarial examples. The objective functions are defined as,

$$\mathcal{L}_{D_1} = -\mathcal{L}_{adv}, \tag{5}$$

$$\mathcal{L}_G = \mathcal{L}_{adv} + \lambda_{rec}\mathcal{L}_{rec} + \lambda_{cos}\mathcal{L}_{cos}, \tag{6}$$

where λ_{rec} and λ_{cos} are hyper-parameters that control the relative importance of reconstruction loss and cosine loss respectively compared to the adversarial loss. In our work, we use λ_{rec} = 10 and λ_{cos} = 10.

3.3 Identity discriminator

In this work, we propose an identity discriminator D_2 as third-player to participate in the generative adversarial competition which brings about impersonating target faces better. For generating images with similar feature representations to the target image, we adopt a face recognition network as the identity discriminator directly. In the white-box scenario, we adopt the target face recognition network as the identity discriminator. But in the black-box scenario, we introduce a substitute network to imitate the feature representation of the target face recognition networks so that it can attack multiple networks without any access to these networks which will be elaborated in Section 4.3.

For a given input image x, and a latent code z from the target image y, $E(y) \rightarrow z$, our goal is to translate x into $\hat{x}$, $G_1(x, z) \rightarrow \hat{x}$, which can be classified as y by D_2.

3.4 Attentional VAE and attentional generator

Our aim is to control the distribution of image space and feature space of generated faces, so we need a stronger feature extractor. In the process of encoding the target face image, in order to enhance feature encoding ability for the target face, and increase the similarity between the generated face and the original face in feature space, we plug the geometric attentional block into VAE to constitute *attentional variational autoencoder*. In addition, in order to enhance the generator, we adopt a channel-wise attentional block into the generator to model interdependencies between the channel to capture feature representation of faces named *attentional generator*.

The overview of Attentional VAE is shown in Fig. 3. VAE in our work is to learn the feature representation of the target person whose facial dependency is significant for capturing the latent code. It is related to the self-attention method which computes the response at one point in a sequence by attending to all points. For this purpose, we introduce non-local block [62] to capture facial dependency. For instance-level learning, we combine basic variational autoencoder residual block and non-local to propose *Attentional VAE (AVAE)* in our A^3GN in Fig. 3. As shown in Fig. 2, AVAE can encode the geometric information of target face and learn the facial dependency from different parts of the human face effectively.

We concatenate the original face x (3-dimension) with the latent code z as the input of the attentional generator in Fig. 4. After two subsampling convolution layers in the generator, we introduce squeeze-and-excitation operations [23] to emphasize informative features and suppress less useful ones in the channel. SE operations propose to squeeze global spatial information into a channel descriptor by using global average pooling to generate channel-wise statistics. In excitation operation, a gating mechanism with a sigmoid activation is employed to capture channel-wise dependencies. Finally, we employ scaling to rescale the transformation output. Owing to squeeze-and-excitation, we can maintain informative features from the latent code more and suppress the useless information in channels which contribute to capturing feature representation of the target person.

3.5 Feature estimation for black-box attack

Feature estimation is a crucial thought in black-box attack. The generalization of the generator trained in the white-box scenario can not satisfy the performance of the black-box attack. It can not attack networks in black-box well shown in Table 7. Thus we present feature estimation in the black-box scenario which can estimate the feature representation of black-box networks. For feature estimation, we adopt a substitute model to estimate the feature representation of target networks. The substitute model is used to fit the black-box face recognition networks. It does not provide the probability of the input face but provides the feature representation. The substitute model S is a pre-trained simple face recognition

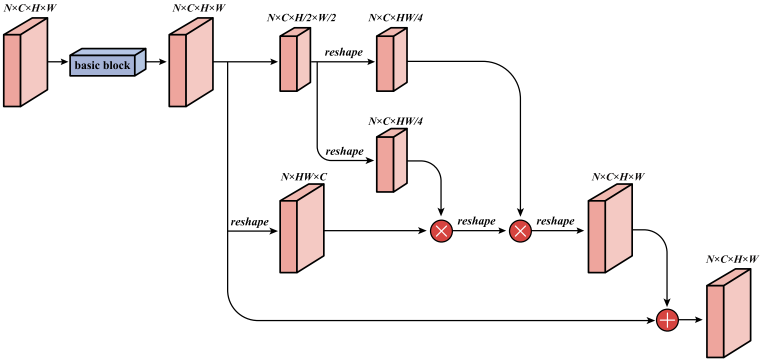

Fig. 3 Overview of the geometric attentional block in AVAE. Basic block is a basic residual block, which is introduced in [20]. ⊕ means element-wise sum and ⊗ means matrix multiplication

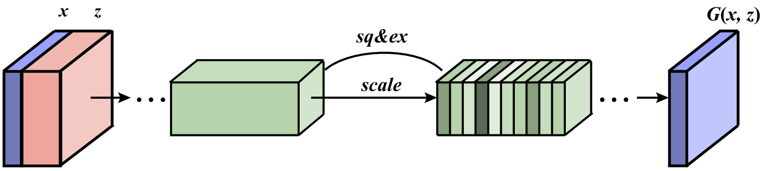

Fig. 4 Overview of a channel-wise attentional block in the Attentional Generator. "*sq&ex*" is the squeeze operation (global average pooling) and the excitation operation (gating mechanism with a sigmoid activation). "*scale*" is the operation to rescale the transformation output with activations after squeeze and excitation to get the channels with different weights of importance

network that can imitate the feature in the training phase. And it will be the identity discriminator D_2 in the black-box scenario to replace the target face recognition network in the white-box scenario which will be updated during the training phase to imitate the feature representation.

The substitute model S takes $\hat{x}$ as input. To estimate the feature representation of target networks T_N (N means the number of target networks which will be 3 in our following experiments), we introduce *estimation loss* to promise that the substitute model can get the similar feature representation of target networks:

$$\mathcal{L}_{est} = \mathcal{L}_{est\_target} + \mathcal{L}_{est\_fake}, \tag{7}$$

where

$$\mathcal{L}_{est\_target} = \mathbb{E}_x[\left\| \frac{\sum_{i=1}^{N} \|T_i(x)\|_2}{N} - \|S(x)\|_2 \right\|_1], \tag{8}$$

$$\mathcal{L}_{est\_fake} = \mathbb{E}_{x,z}[\left\| \frac{\sum_{i=1}^{N} \|T_i(G_1(x,z))\|_2}{N} - \|S(G_1(x,z))\|_2 \right\|_1], \tag{9}$$

where N means the total amount of target networks.

Considering that different face recognition networks obtain various feature representation, we first normalize the features of different networks, and then average the normalized feature for estimation loss. The substitute model aims at obtaining a similar normalized feature representation for the same input. The substitute model will output the feature of adversarial examples to fit the feature of the target image. Meanwhile, $D_2(G_1(x,z))$ used in $\mathcal{L}_{cos}$ will be replaced by $S(G_1(x,z))$ in the black-box scenario.

4 Experiments

4.1 Evaluation metrics

In our work, we define a set of specific evaluation metrics to measure the effectiveness of the attacks:

- **Real Accuracy & Fake Accuracy & mAP.** When cosine distance between examples and target faces is more than 0.45, we consider examples as target faces with a true predicting label. Real accuracy shows the percentage of original images that can be classified as the target person, which is usually 0%, while fake accuracy shows the percentage of generated images that can be classified as the target person. mAP is the mean average precision with different thresholds in a range from 0 to 1 whose step is 0.01.

- **Similarity Score.** Cosine distance between original faces/generated faces and target face faces is seen as a Similarity Score. Cosine distance is a significant metric in face recognition for verifying whether the two images belong to one person. In our results, we show the Similarity Scores before / after the attack and the improvement (Δ) to exhibit the effectiveness of A^3GN for attacks. Meanwhile, the Similarity Scores between the real image and the fake image (denotes as R - F) exhibit the ability that the generated images can be recognized as their real identities by face recognition networks. The Similarity Scores between the real image and the fake image are less, the attack is more successful.
- **SSIM.** SSIM means the percentage of structural similarity index between original faces and generating faces higher than a threshold. SSIM is a quantization metric to determine whether generating faces are perturbed slightly compared with original faces. In our work, we set 0.9 as the threshold to evaluate the quality of generated images compared to original images.

4.2 Datasets

The state-of-the-art face recognition networks are trained in CASIA-WebFace dataset [68] and refined MS-Celeb-1M [9, 19]. Meanwhile, our A^3GN is also trained on CASIA-WebFace. And in the inference time, we perform A^3GN on LFW [24], CFP-FP [49] and AgeDB-30 [42] by generating adversarial examples paired with target faces to verify whether they belong to one person.

CASIA-WebFace CASIA-WebFace dataset [68] is a web-collected dataset which has 494,414 face images belonging to 10,575 different individuals. In our experiments, we use aligned CASIA-WebFace which has images with size of 112×112 after alignment.

MS-Celeb-1M The original MS-Celeb-1M dataset [19] contains about 100k identities with 10 million images. In [9], the noise of MS-Celeb-1M is decreased, and finally, refined MS-Celeb-1M contains 3.8M images of 85k unique identities.

LFW Labelled Faces in the Wild (LFW) dataset [24] contains 13,233 web-collected images from 5,749 different identities, with large variations in pose, expression and illuminations. In face verification, the verification accuracy is usually measured on 6,000 face pairs. But in our work, we pair all the images in LFW with target face image.

CFP-FP Celebrities in Frontal Profile (CFP-FP) dataset [49] contains about 7,000 multi-angle images from 500 different identities. For each identity, it has 10 frontal images and 4 lateral images, which are very suitable for evaluating side face recognition algorithms.

AgeDB-30 Age Database (AgeDB-30) [42] contains 16,488 images of celebrities belonging to 568 identities, such as actors, writers, scientists and politicians, each with identity, age and gender attributes. The minimum and maximum ages are 1 and 101, respectively. The average age range for each subject is 50.3 years.

4.3 Implementation of A^{3}GN

In this section, we do some experiments to verify the feasibility and effectiveness of attentional blocks. We train A^3GN on CASIA-WebFace and utilize it to generate adversarial examples on LFW to attack the target model in the inference time. We employ ArcFace which has an accuracy of 99.53% [9] on LFW as the target model in these experiments about the architecture of A^3GN. In the white-box scenario, the target face recognition network will be ResNet50 with softmax, SphereFace and ArcFace successively. In the black-box scenario, we use a pre-trained model ResNet50 as the substitute network to do the feature estimation of different target networks.

Network Architecture We design A^3GN based on cVAE-GAN. For the encoder, we use a classifier with 4 residual basic blocks for the latent code. Adapted from [6] [70], the generator in our work is composed of two convolution layers for downsampling, 6 residual blocks, and two convolution layers for upsampling. In the generator, we use instance normalization which is not used in discriminator. In our work, we have two discriminators. One is the target face recognition network for classifying whether the image patches belong to the target person or not called *identity discriminator* and another is PatchGAN discriminator [27] for classifying whether the image patched are real or not called *sample discriminator*.

Training Details In the training process, the target person contains 7 different face images for capturing the latent code. All the input images are resized and cropped to 112×112. Because our goal is to generate images for fooling the face recognition network, all the images should do the alignment similar to the operation in face verification. We update generator once by $\mathcal{L}_G$ after five sample discriminator updates and one generator update by $\mathcal{L}_{cos}$ while the identity discriminator is fixed all the time in the white-box scenario. And in the black-box scenario, the substitute model is updated by $\mathcal{L}_{est}$ in each iteration. All the models are trained for 200,000 iterations and use Adam [31] with $\beta_1 = 0.5$ and $\beta_2 = 0.999$. The batch size is set to 32 in all experiments. We set the learning rate to 0.0001 for the first 100,000 iterations and linearly decay the learning rate to 0 over the next 100,000 iterations.

4.4 Quantitative evaluation

In this section, we do some ablation experiments in the white-box scenario to optimize A^3GN. The effectiveness of geometric attention and channel-wise attention will be shown by quantitative evaluation as follows. The target network used in this section is ArcFace.

Baseline We perform a quantitative analysis of the mAP, the difference of Similarity Score and SSIM on our baseline. All the results are calculated on average among 5 target faces to eliminate the occasionality. The performance of baseline is shown in Table 1. We design two groups of experiments with different conditions to verify the effectiveness of A^3GN. One is to encode image A to attack the same image A. Another is to encode image A to attack image A', A and A' belong to the same identity. Neither A and A' is in target image datasets for the latent code in the training process. In the experiment of baseline, we choose one target person randomly to test the performance. The threshold of cosine distance for fake accuracy is set to 0.45. The experiment in $A \to A$ can get higher accuracy than the experiment in $A \to A'$ because it can learn the feature representation of A in the encoder

Table 1 Performance of baseline with two conditions. A and A' are different samples, but belong to the same identity

	Fake Acc.(%)	mAP(%)	Similarity Score(Δ)	SSIM(%)
$A \to A$	97.52	53.74	0.506	3.47
$A \to A'$	93.82	51.72	0.490	3.57

for attacking A. As shown in the Table 1, our baseline can fool the target model generally. Most of them can be classified as the target person at a threshold of 0.45. SSIM is a metric to evaluate the quality of generated images in some similar works. But we think it does not an objective metric to evaluate the similarity between the generated images and the original images for human eyes.

Attentional VAE In Section 3.4, we propose AVAE to obtain global geometric information of the target face image, but there are many types of pairwise functions in non-local module [62], such as Embedded Gaussian and Dot Product. In Table 2, we analyze the impact of different pairwise functions on A3GN. Obviously, each metric gets improvement compared with the baseline (without non-local module), which indicates that It is effective to capture the facial dependencies for encoding the latent code. In addition, we can also see that the AVAE module with Embedded Gaussian is better than that with Dot Product. However, no similar results have been observed in [62]. We infer that this is because compared with Dot Product, Embedded Gaussian pairwise function has the characteristics of self-attention mechanism, which makes AVAE module pay more attention to important areas of face. This improves the quality of generated face, especially the SSIM metric.

Attentional Generator The geometric attention in encoder can capture the global geometric information effectively. We conjecture that introducing attentional blocks in the generator may also get better performance. During the process of generating images, the generator forces the fake images more similar to the original images which result in the loss of feature representation of the target person due to $\mathcal{L}_{rec}$. Thus, we consider introducing a channel-wise attentional block into the generator to focus on the information of the latent code. As explained in [23], we can control the reduction ratio r of SE module to get a geometric attentional block with different capacities, the performance at different reduction ratios is shown in Table 3. Attentional generators with different reduction ratios have similar performance and far exceed the baseline (without SE module). This indicates that channel attention mechanism can retain more information features from potential code and suppress useless information in channels. Channel attention mechanisms introduce additional parameters, which are inversely proportional to the reduction ratio. However, the experiment in Table 3 shows that the performance of attentional generator is almost independent of the parameters of SE module, which further proves that the effectiveness of our method comes from the screening and enhancement of effective information of the latent code, not only the increase of model capacity.

Dimensions of the Latent Code For the dimension of the latent code z, we also do an experiment to find the best dimension for encoding the target image. Considering the computation cost, we set the dimension of 3, 5, 7, and 10. Different dimensions of the latent code mean different amounts of instance information of the target person. In our work, we need to explore the most suitable dimension of the latent code to represent the target person best

Table 2 Ablation study on geometric attentional block with different pairwise functions when $A \rightarrow A$

	Pairwise Function	Fake Acc.(%)	mAP(%)	Similarity Score(Δ)	SSIM(%)
$A \rightarrow A$	Baseline	97.52	53.74	0.532	3.47
	Embedded Gaussian	98.12	54.62	0.543	**3.59**
	Dot Product	97.92	54.04	0.539	**3.39**

Table 3 Ablation study on channel-wise attentional block with different reduction ratios when $A \rightarrow A$

	Ratio r	MParams (Δ)	Fake Acc.(%)	mAP(%)	Similarity Score(Δ)	SSIM(%)
$A \rightarrow A$	Baseline	0.0	97.52	53.74	0.532	3.47
	2	12.6	99.61	54.98	0.552	4.59
	4	6.1	99.67	56.34	0.557	**4.70**
	8	2.4	**99.68**	**56.27**	**0.561**	4.63
	16 (ours)	**0.8**	99.66	**56.27**	0.558	4.61

for the input of our generator. From Table 4, we can see that the experiments with different dimensions will get similar results with a little difference. So, we can draw a conclusion that the fake accuracy and the Similarity Score can benefit from the latent codes with different dimensions preliminarily in quantitative evaluation. These four dimensions can complete our attack task. By synthesizing previous research work [71] and our experimental results, the dimensions of the latent code in above experiments are all set to 7.

Following the three aforementioned ablation studies, we combine geometric attention and channel-wise attention to improve the performance. The results are shown in Table 5. And the curves of accuracies in $A \rightarrow A$ are shown in Fig. 5. As we can see, most of the generated images can get more than 0.4 of cosine distance which far surpasses the result between real images and the target image. A^3GN can fool the face recognition network successfully.

4.5 Qualitative evaluation

In addition to the quantitative evaluation, we exhibit the effectiveness of 4 different models by showing the qualitative comparison results in Fig. 6. All the generated images in Fig. 6 can be classified as the target person in the threshold of 0.45 and they are similar to the original images just with quasi-imperceptible perturbation. We observe that our model can provide a higher visual quality of attack results on LFW even in baseline. However, the generated images are similar to the target image in physical likeness slightly such as the nose and eyes. We conjecture that it is because that the generator hammers at making the cosine distance between generated images and target image higher. It shows that face recognition network recognizes people by focusing on their noses and eyes more, and the contours of their faces and mouths less.

About the dimension experiments, we also generate adversarial examples to estimate the visual effects. In quantitative evaluation, we observe that the experiments with different dimensions can get similar results. But in qualitative evaluation, we observe that there are

Table 4 Ablation study on AVAE with different dimensions of the latent code when $A \rightarrow A$

	Dim.	Fake Acc.(%)	mAP(%)	Similarity Score(Δ)	SSIM(%)
$A \rightarrow A$	3	99.36	55.54	0.512	5.05
	5	99.64	55.87	0.519	5.19
	7 (ours)	99.59	56.28	0.533	**5.19**
	10	**99.71**	**56.29**	**0.543**	5.13

Table 5 Ablation study on A^3GN. Baseline: Conditional GAN baseline

		Attack Acc. on LFW(%)			Similarity Score			SSIM(%)
		Real Acc.	Fake Acc.	mAP	Before	After	R - F	
$A \rightarrow A$	Baseline	0.0	97.52	53.74	0.026	$0.532_{(0.506)}$	0.163	3.47
	Geometric Att.	0.0	98.12	54.62	0.026	$0.543_{(0.517)}$	0.161	3.59
	Channel-wise Att.	0.0	**99.66**	56.27	0.026	$0.558_{(0.532)}$	**0.146**	4.61
	Both	0.0	99.59	**56.28**	0.026	$\mathbf{0.559}_{(0.533)}$	0.161	**5.19**
$A \rightarrow A'$	Baseline	0.0	93.82	51.72	0.022	$0.512_{(0.490)}$	0.163	3.57
	Geometric Att.	0.0	95.08	52.53	0.022	$0.523_{(0.500)}$	0.162	3.69
	Channel-wise Att.	0.0	**98.96**	54.99	0.022	$0.545_{(0.523)}$	**0.146**	4.48
	Both	0.0	98.92	**55.14**	0.022	$\mathbf{0.546}_{(0.525)}$	0.160	**5.16**

Geometric Att.: Conditional GAN with geometric attentional block. Channel-wise Att.: Conditional GAN with channel-wise attentional block. Both: Conditional GAN with geometric attentional and channel-wise attentional blocks. The threshold of cosine distance is set to 0.45

some differences among the images generated from the different experiments. The inapparent outline of the target person emerges on the images generated from the experiment with low-dim shown in Fig. 7. But in the experiment with high-dim, the generated images are clearer. So, we can draw a conclusion that though the experiments with different dimensions can get similar results in quantitative evaluation their performances on the visual effects are different. A^3GN focuses on the geometric information more in low-dim experiment due to the less information of the latent code sent into the generator. Meanwhile, it focuses on the semantic information more in high-dim experiment. But considering the training speed, too

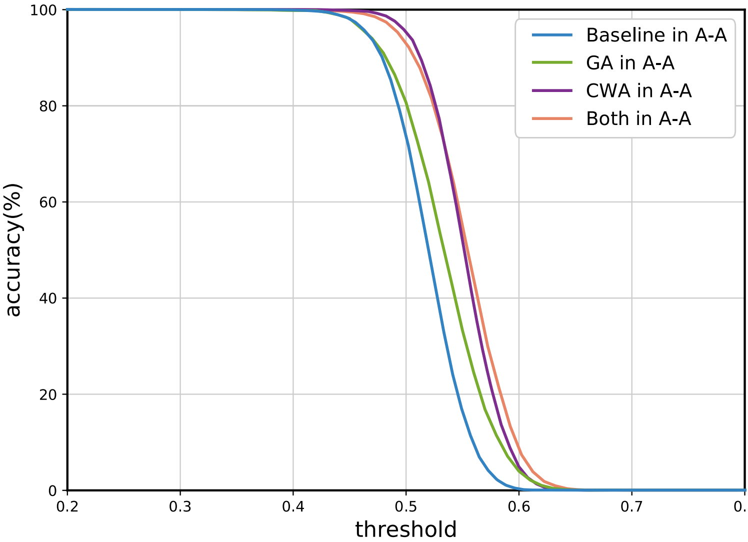

Fig. 5 Accuracy curve in different thresholds. The horizontal axis represents the different thresholds and the vertical axis represents the accuracy in different thresholds. GA means geometric attention. CWA means channel-wise attention

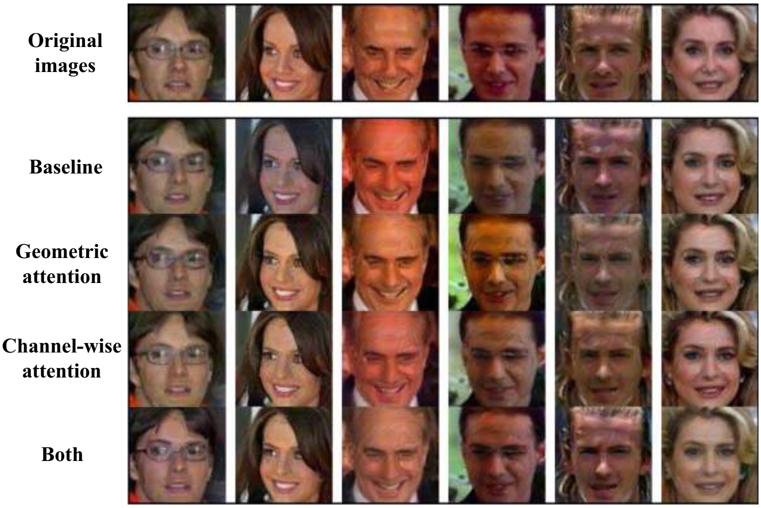

Fig. 6 Generated images by A^3GN with 4 different models. The first row is the original images and the rest is the generated images by A^3GN with 4 different models. The target person is Target A in Fig. 8

high-dim is not suitable. So in the following experiments, we choose 7-dim as the dimension of the latent code.

Furthermore, we choose 5 different target face images to exhibit the results of attacks in Fig. 8. Most of the generated images are prone to the target face image slightly. It would seem that most face recognition network focuses on recognizing people by their facial feature and a slight change on the facial feature can fool the face recognition network to recognize as another person which is imperceptible for observers. Meanwhile, a mask learned from the target person can also fool the network.

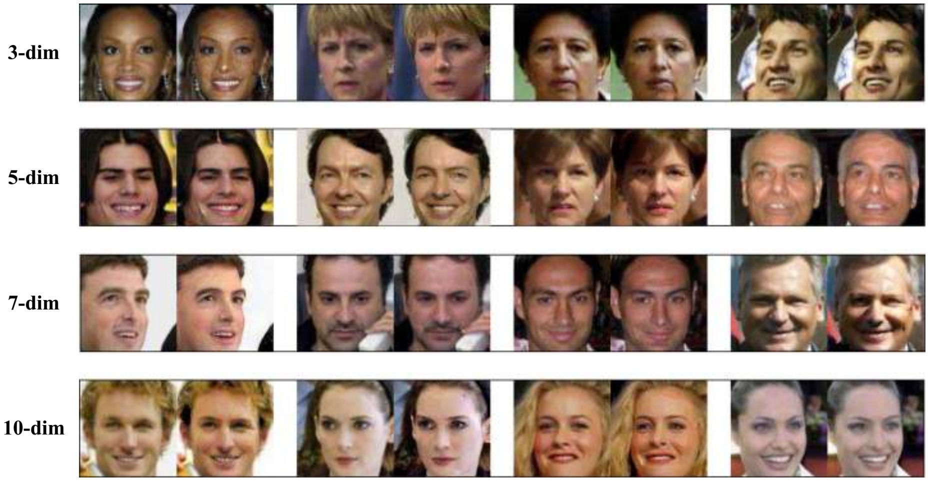

Fig. 7 Generated images by A^3GN in the experiments with 3-dim, 5-dim, 7-dim and 10-dim of the latent codes. The left is original image and the right is generated image. All the generated images can be classified as target person by ArcFace in the threshold of 0.45. The target person is Target A in Fig. 8

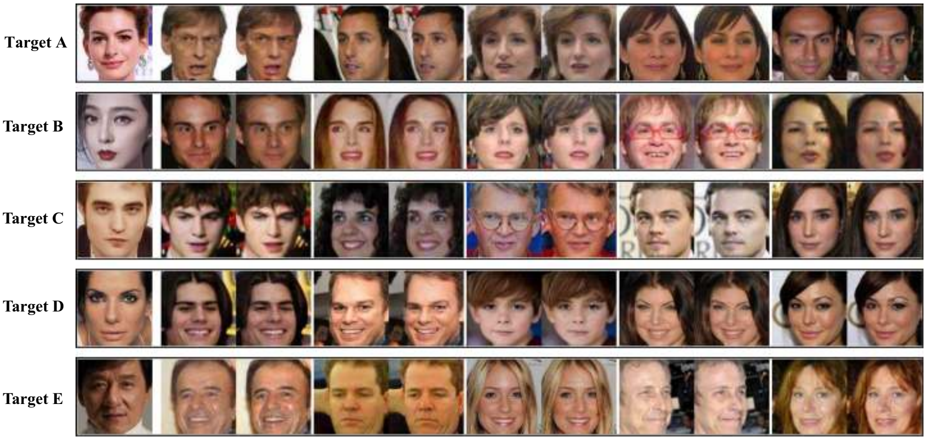

Fig. 8 Generated images by A^3GN for 5 target faces. The left is original image and the right is generated image

4.6 White-box and black-box

In this section, we do some experiments to verify the performance of A^3GN in the white-box scenario and the black-box scenario with different target models.

White-box Attack In the white-box scenario, we choose 4 different state-of-the-art face recognition networks to verify the feasibility of our model A^3GN and the generated images are evaluated on LFW, CFP-FP and AgeDB-30 datasets, respectively. The performance on different target models in the white-box scenario is shown in Table 6. In the white-box scenario, the parameters, architectures and the feature space of target models are obtained in the training process. Thus, the generator can generate images directionally. The metrics of evaluation in this section are mAP, the difference of Similarity Score, the Similarity Score between real images and fake images and SSIM. And in Table 6, we also show the accuracy of the target network on three datasets in face verification. mAP and the difference of Similarity Score indicate the ability to fool the networks to recognize as the target person and the Similarity Score between real images and fake images can indicate the ability to fool the networks to be mistaken. All of them can prove that our model A^3GN can be applied to fool different state-of-the-art networks.

In the experiment on attacking ResNet with softmax, Center Loss and SphereFace and evaluating on LFW dataset, mAPs are lower than that in the experiment on attacking ArcFace. But the attack on Center Loss and SphereFace are more effective in reducing the Similarity Score between real images and fake images. And we conjecture that different training data and different accuracy on LFW result in the different performances on generated images. For ResNet with softmax, Center Loss and SphereFace, the training data is CASIA-WebFace. But for ArcFace, the training data is MS-Celeb-1M. Different training data bring about different feature representations.

To evaluate the effectiveness of generated adversarial samples in pose and age variations, we also test the results of white-box attacks on CFP-FP and AgeDB-30 datasets. For CFP-FP, we can see that the Similarity Score of the generated images has been significantly improved, which shows that A^3GN can effectively attack CFP-FP dataset. However,

Table 6 A^3GN performance on different target models and evaluation datasets in the white-box scenario

	Target Model	Verification Acc.(%)	mAP(%)	Similarity Score		SSIM(%)
				Δ	R - F	
LFW	Softmax	97.02	45.32	0.421	0.232	5.52
	Center Loss [63]	99.28	51.03	0.491	0.107	6.08
	SphereFace [60]	99.20	49.04	0.478	0.090	6.44
	ArcFace [9]	99.53	56.28	0.533	0.161	5.19
CFP-FP	Softmax	92.17	40.28	0.390	0.301	8.03
	SphereFace [60]	94.38	43.74	0.449	0.195	8.71
	ArcFace [9]	95.56	50.11	0.501	0.226	6.75
AgeDB-30	Softmax	90.85	39.72	0.400	0.250	6.37
	SphereFace [60]	91.70	42.37	0.481	0.089	6.11
	ArcFace [9]	95.15	50.09	0.509	0.186	5.69

according to the SSIM metric, we can observe that the generated images are quite different from the original images in the image space, which implies that the diversity of face angles poses a greater challenge to reconstructing face samples. For AgeDB-30, A^3GN can still effectively deceive it, and the effect of attack is better than CFP-FP, slightly less than LFW. Although the accuracy of ResNet with softmax, Center Loss and SphereFace in face verification on AgeDB-30 is lower than that of LFW and CFP-FP , the attack effect is not affected much, which indicates that A^3GN has better generalization ability for age variation. Especially, the SSIM metric is obviously improved compared with CFP-FP, which shows that the generated adversarial samples are closer to the original image in image space.

Black-box Attack In this section, we explore whether fooling one face recognition network leads to successful fooling other networks. In the black-box scenario, the parameters, architectures and the feature space of target models are not obtained in the training process. The identity discriminator in the black-box scenario is only ArcFace [9] in Table 7. And we have no access of target networks, ResNet [20] with softmax, SphereFace [35] and MobileFaceNet [5] in the training process. In the inference time, we just obtain the feature map of images from the last layers to test the performance. The performance on different target networks in the black-box scenario is shown in Table 7. Obviously, each result in Table 7 is lower than that in Table 6. But we also observe that the generated images can disturb the target networks slightly.

Table 7 A^3GN performance on different target models in the black-box scenario. All the metrics are evaluated on LFW dataset

Target Model	Verification Acc.(%)	mAP(%)	Similarity Score	
			Δ	R - F
Softmax	97.02	17.70	0.082	0.581
SphereFace [60]	99.20	12.32	0.110	0.593
MobileFaceNet [5]	99.18	9.07	0.297	0.407

Table 8 A^3GN performance on different target models in the black-box scenario with feature estimation

Target Model	Verification Acc.(%)	mAP(%)	Similarity Score	
			Δ	R - F
Softmax	97.02	29.05	0.517	0.277
SphereFace [60]	99.20	28.86	0.523	0.265
MobileFaceNet [5]	99.18	27.96	0.477	0.278

To improve the performance of the black-box attack, we introduce a substitute model to imitate the feature representation of the target models which is named *feature estimation* in our work. Certainly, we do an experiment to verify the effectiveness of feature estimation in the black-box scenario. Table 8 shows the performance of feature estimation. In this experiment, we adopt a pre-trained ResNet50 with softmax loss as the substitute model and ArcFace, SphereFace, and MobileFaceNet as the target models. We only get the outputs of the target models for feature estimation in the black-box scenario without the parameters and the gradients. Results in Table 8 are obtained by training one substitute model which means that A^3GN can attack several black-box models once. As we can see, compared to Table 7, the fake accuracy and the Similarity Score are improved a lot after feature estimation operation which can verify the effectiveness of feature estimation. Though it can not exceed the performance of the white-box attack, it can attack the target models roughly. The generated images with better visual effects and better attack effects are shown in Fig. 9. Due to the attack on multiple models, the generated images have a more serious outline of target person and some masks. To eliminate the outline for the generated images will be our future work.

Comparison with Previous Works We compare our A^3GN with previous attack models in face recognition on CASIA-WebFace dataset. Because they focus on fool the classifier to a false label, we compare our performance on this way in Table 9. If the cosine distance between the original image and the generated image is lower than 0.45, it is seen as a success for an attack. As we can see, the success rate of fool the face recognition network to a false label for A^3GN is 99.94%. It almost fools the network totally. Though it is 0.02% lower than GFLM, A^3GN can force the target model to recognize as the target person well.

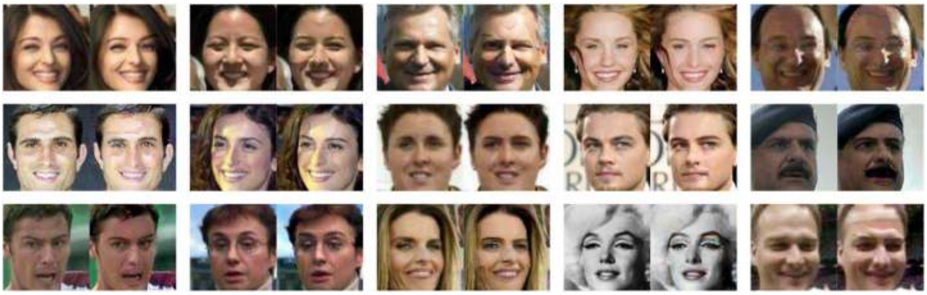

Fig. 9 Generated images by A^3GN in the experiment in the black-box scenario with feature estimation. The left is original image and the right is generated image. All of them can attack 3 target models with the threshold of 0.45. The substitute model is ResNet50 and the target person is Target A in Fig. 8

Table 9 Comparison with other attack models in face recognition. 'SR' means the success rate of fooling the network to a false label. 'Attack acc. on CASIA' means the accuracy of fooling the network to a target label

	SR(%)	Attack Acc. on CASIA(%)
stAdv [65]	99.18	–
GFLM [8]	99.96	–
A^3GN (ours)	**99.98**	**98.23**

5 Conclusion

Face recognition is a compelling task in deep learning. It is necessary to learn how face recognition networks are subject to attacks. In this paper, we focus on a novel way of attacking target models by fooling them to a specific label. For this purpose, we present A^3GN to generate adversarial examples similar to the original images but which can be classified as the target person. To learn the feature representation of target images, we introduce geometric attention and channel-wise attention into A^3GN to get good performance. Finally, we show the results of experiments on different target faces, white-box attack, and black-box attack. However, our model is limited to attacking one target person. It will be a future work that one model can attack different target faces.

In addition, we believe that the value of A^3GN is not limited to attack face recognition. Arbitrary manipulating the image space and feature space distributions of generated images has great research value. It can be used to study the interpretability of convolutional neural networks, universal representations and domain adaptation, etc. We have reason to believe that more work will pay attention to this issue in the future.

References

1. Akhtar N, Mian A (2018) Threat of adversarial attacks on deep learning in computer vision: A survey. arXiv:1801.00553
2. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein generative adversarial networks. In: ICML
3. Bose A, Aarabi P (2018) Adversarial attacks on face detectors using neural net based constrained optimization. arXiv:1805.12302
4. Carlini N, Wagner D (2017) Towards evaluating the robustness of neural networks. In: IEEE Symposium on Security and Privacy
5. Chen S, Liu Y, Gao X, Han Z (2018) Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In: CCBR
6. Choi Y, Choi M, Kim M, Ha J, Kim S, Choo J (2018) Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In: CVPR
7. Chopra S, Hadsell R, LeCun Y (2005) Learning a similarity metric discriminatively with application to face verification. In: CVPR
8. Dabouei A, Soleymani S, Dawson J, Nasrabadi NM (2018) Fast geometrically-perturbed adversarial faces. arXiv:1809.08999
9. Deng J, Guo J, Zafeiriou S (2018) Arcface: Additive angular margin loss for deep face recognition. arXiv:1801.07698

10. Denton E, Chintala S, Fergus R et al (2015) Deep generative image models using a laplacian pyramid of adversarial networks. In: NIPS
11. Dong Y, Su H, Wu B, Li Z, Liu W, Zhang T, Zhu J (2019) Efficient decision-based black-box adversarial attacks on face recognition. In: CVPR
12. Engstrom L, Tsipras D, Schmidt L, Madry A (2017) A rotation and a translation suffice: Fooling cnns with simple transformations. arXiv:1712.02779
13. Gao Z (2017) Wu Y, Jia Y, Learning a robust representation via a deep network on symmetric positive definite manifolds. Pattern Recognit
14. Gatys LA, Ecker AS, Bethge M (2016) Image style transfer using convolutional neural networks. In: CVPR
15. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets, pp 2672–2680
16. Goodfellow I, Shlens J, Szegedy C (2015) Explaining and harnessing adversarial examples. In: ICLR
17. Goswami G, Ratha N, Agarwal A, Singh R, Vatsa M (2018) Unravelling robustness of deep learning based face recognition against adversarial attacks. arXiv:1803.00401
18. Gulrajani I, Ahmed F, Arjovsky M, Dumoulin V, Courville A (2017) Improved training of wasserstein gans. arXiv:1704.00028
19. Guo Y, Zhang L, Hu Y, He X, Gao J (2016) Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV
20. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: CVPR
21. He Q, He B, Zhang Y (2019) Multimedia based fast face recognition algorithm of speed up robust features. Multimed Tools Appl
22. Hou R, Ma B, Chang H, Gu X, Shan S, Chen X (2019) Interaction-and-aggregation network for person re-identification. In: CVPR
23. Hu J, Shen L, Sun G (2017) Squeeze-and-excitation networks. arXiv:1709.01507
24. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical report
25. Huang R, Xu B, Schuurmans D, Szepesvari C (2015) Learning with a strong adversary. arXiv:1511.03034
26. Huang Z, Wang R, Shan S, Gool L, Chen X (2016) Cross euclidean-to-riemannian metric learning with application to face recognition from video. In: TPAMI
27. Isola P, Zhu J, Zhou T, Efros A (2017) Image-to-image translation with conditional adversarial networks. In: CVPR
28. Johnson J, Alahi A (2016) Fei-Fei L. In: ECCV. Perceptual losses for real-time style transfer and super-resolution
29. Kanbak C, Moosavi-Dezfooli SM, Frossard P (2017) Geometric robustness of deep networks: analysis and improvement. arXiv:1711.09115
30. Karras T, Laine S, Aila T (2019) A style-based generator architecture for generative adversarial networks. In: CVPR
31. Kingma D, Ba J (2014) Adam: A method for stochastic optimization. In: ICLR
32. Lin TY, RoyChowdhury A, Maji S (2015) Bilinear cnns for fine-grained visual recognition. In: ICCV
33. Liu J, Zha Z, Tian QI, Liu D, Yao T, Ling Q, Mei T (2016a) Multi-scale triplet cnn for person re-identification. In: ACM MM
34. Liu W, Wen Y, Yu Z, Yang M (2016b) Large-margin softmax loss for convolutional neural networks. In: ICML
35. Liu W, Wen Y, Yu Z, Li M, Raj B, Song L (2017) Sphereface: Deep hypersphere embedding for face recognition. In: CVPR
36. Mao S, Zhang S, Yang M (2019) Resolution-invariant person re-identification. In: IJCAI
37. Mathieu M, Zhao J, Ramesh A, Sprechmann P, LeCun Y (2016) Disentangling factors of variation in deep representation using adversarial training. In: NIPS
38. Miyato T, i Maeda S, Koyama M, Nakae K, Ishii S (2016) Distributional smoothing with virtual adversarial training. In: ICLR
39. Miyato T, Kataoka T, Koyama M, Yoshida Y (2018) Spectral normalization for generative adversarial networks. In: ICLR
40. Moosavi-Dezfooli S, Fawzi A, Fawzi O (2017) Universal adversarial perturbations. In: CVPR
41. Moosavi-Dezfooli SM, Fawzi A, Frossard P (2016) Deepfool: a simple and accurate method to fool deep neural networks. In: Proc CVPR
42. Moschoglou S, Papaioannou A, Sagonas C, Deng J, Kotsia I, Zafeiriou S (2017) Agedb: The first manually collected in-the-wild age database. In: CVPR Workshop
43. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: ICML

44. Radford A, Metz L, Chintala S (2016) Unsupervised representation learning with deep convolutional generative adversarial networks. In: ICLR
45. Salimans T, Goodfellow I, Zaremba W, Cheung V, Radford A, Chen X (2016) Improved techniques for training gans. arXiv:1606.03498
46. Salimans T, Zhang H, Radford A, Metaxas D (2018) Improving gans using optimal transport. In: ICLR
47. Sanakoyeu A, Tschernezki V, Büchler U, Ommer B (2019) Divide and conquer the embedding space for metric learning. In: CVPR
48. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: A unified embedding for face recognition and clustering. In: CVPR
49. Sengupta S, Chen J, Castillo C, Patel V, Chellappa R, Jacobs D (2016) Frontal to profile face verification in the wild. In: WACV
50. Sharif M, Bhagavatula S, Bauer L, Reiter MK (2016) Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. ACM SIGSAC, In
51. Sharif M, Bhagavatula S, Bauer L, Reiter MK (2018) Adversarial generative nets: Neural network attacks on state-of-the-art face recognition. arXiv:1801.00349
52. Sohn K, Yan X, Lee H (2015) Learning structured output representation using deep conditional generative models. In: NIPS
53. Song Y, Shu R, Kushman N, Ermon S (2018) Constructing unrestricted adversarial examples with generative models. In: NIPS
54. Su J, Vargas DV, Sakurai K (2017) One pixel attack for fooling deep neural networks. arXiv:1710.08864
55. Sun Y, Wang X, Tang X (2014) Deep learning face representation from predicting 10,000 classes. In: CVPR
56. Sun Y, Liang D, Wang X, Tan X (2015) Deepid3: Face recognition with very deep neural networks. arXiv:1502.00873
57. Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R (2013) Intriguing properties of neural networks. In: ICLR
58. Taigman Y, Yang M, Ranzato M, Wolf L (2014) Deepface: Closing the gap to human-level performance in face verification. In: CVPR
59. Ulyanov D, Lebedev V, Vedaldi A, Lempitsky V (2016) Texture networks: Feed-forward synthesis of textures and stylized images. In: ICML
60. Wang F, Liu W, Liu H, Cheng J (2018) Additive margin softmax for face verification. arXiv:1801.05599
61. (2018) Deep face recognition: A survey. arXiv:1804.06655
62. Wang X, Girshick R, Gupta A, He K (2017) Non-local neural networks. arXiv:1711.07971
63. Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: ECCV
64. Xiao C, Li B, Zhu J, He W, Liu M, Xiao D, Song D (2018a) Generating adversarial examples with adversarial networks. In: IJCAI
65. Xiao C, Zhu J, Li B, He W, Liu M, Song D (2018b) Spatially transformed adversarial examples. arXiv:1801.02612
66. Yan X, Yang J, Sohn K, Lee H (2016) Attribute2image: Conditional image generation from visual attributes. arXiv:1512.00570
67. Yao H, Zhang S, Zhang Y, Li J, Tian Q (2017) One-shot fine-grained instance retrieval. In: ACM MM
68. Yi D, Lei Z, Liao S, Li SZ (2014) Learning face representation from scratch. arXiv:1411.7923
69. Zhang X, Xiong H, Lin W, Tian Q (2017) Picking neural activations for fine-grained recognition. In: TOMM
70. Zhu J, Park T, Isola P, Efros A (2017a) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: ICCV
71. Zhu J, Zhang R, Pathak D, Darrell T, Efros A, Wang O, Shechtman E (2017b) Toward multimodal image-to-image translation. In: NIPS

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.