Manifold Projection for Adversarial Defense on Face Recognition

Jianli Zhou[1,2], Chao Liang[1,2,*], and Jun Chen[1,2]

[1] National Engineering Research Center for Multimedia Software, School of Computer Science, Wuhan University, China
[2] Key Laboratory of Multimedia and Network Communication Engineering, Hubei Province

Abstract. Although deep convolutional neural network based face recognition system has achieved remarkable success, it is susceptible to adversarial images: carefully constructed imperceptible perturbations can easily mislead deep neural networks. A recent study has shown that in addition to regular off-manifold adversarial images, there are also adversarial images on the manifold. In this paper, we propose Adversarial Variational AutoEncoder (A-VAE), a novel framework to tackle both types of attacks. We hypothesize that both off-manifold and on-manifold attacks move the image away from the high probability region of image manifold. We utilize variational autoencoder (VAE) to estimate the lower bound of the log-likelihood of image and explore to project the input images back into the high probability regions of image manifold again. At inference time, our model synthesizes multiple similar realizations of a given image by random sampling, then the nearest neighbor of the given image is selected as the final input of the face recognition model. As a preprocessing operation, our method is attack-agnostic and can adapt to a wide range of resolutions. The experimental results on LFW demonstrate that our method achieves state-of-the-art defense success rate against conventional off-manifold attacks such as FGSM, PGD, and C&W under both grey-box and white-box settings, and even on-manifold attack.

Keywords: Face Recognition, Adversarial Defense.

1 Introduction

As one of the most popular real-world applications of computer vision tasks, such as multimedia analysis and surveillance [48], face recognition has been improved by a large margin with the aid of DNNs [41, 42, 44], and some models have even exceeded humans [36, 28]. Despite the prominent role, DNNs suffer from adversarial examples [43, 13], which constructed by maliciously adding some customized small perturbations to original input. Many recent works have shown

* Corresponding author

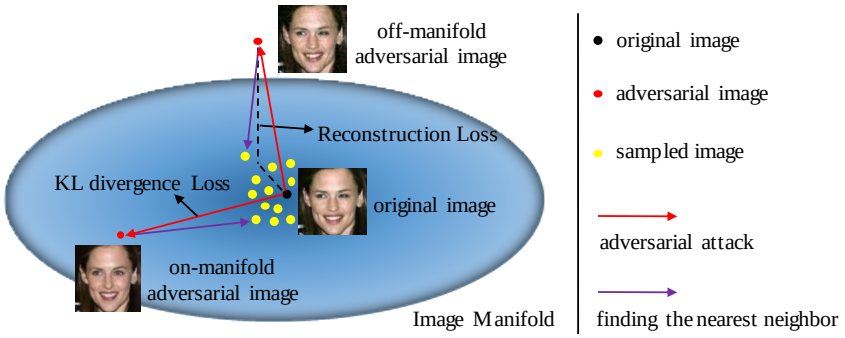

Fig. 1. The main idea of our model to defend adversarial images. Estimated by VAE, off-manifold and on-manifold adversarial images reduce the lower bound of log-likelihood by causing reconstruction loss and KL divergence loss, respectively (we explain it in Section 3.1). Our model samples images in high probability region and replaces the adversarial images with their nearest neighbors.

that adversarial examples can be easily found with a simple gradient method [8, 13, 25, 31]. In response, many defense strategies are proposed as a hedge against the adversarial examples. Several recent works are devoted to imposing transformations on input images due to the advantage of being model-agnostic. This type of method attempts to remove or alleviate hidden perturbations from input images. For example, Defense-GAN [35] employs generative adversarial networks (GANs) [12] to project the input images onto the range of the generator. In [40], Sun et al. design a sparse transformation layer to project the input images to a low dimensional quasi-natural space. However, these defense methods are limited to defend adversarial images which leave the manifold, but cannot handle on-manifold adversarial images.

In this work, we devise a novel *Adversarial Variational AutoEncoder* (A-VAE) to tackle both types of attacks. We assume that both types of adversarial images are moved away from the data distribution of high probability [38], and we focus on pulling them back onto their original regions, e.g. Figure 1. By taking advantage of VAE [24], we estimate the lower bound of the log-likelihood of a given image, and project it onto the high probability region of image manifold using a unified process. Specifically, given a downsampled image as input, the proposed model randomly sample various outputs on the high probability region of image manifold which are approximately similar to the input, then the nearest neighbor is selected to be classified by the face recognition model instead of the original image. Moreover, we notice there is a contradiction: the traditional reconstruction loss in training phase will lead to false retention of perturbation at inference time. To solve this problem, we utilize the well-known adversarial training as GANs to substitute the reconstruction loss. We experimentally demonstrate this property helps the model strip noise from images. In summary, our contributions are:

- We propose a generative defense method against diverse adversarial attacks on complex face dataset under both grey-box and white-box settings, which leverages the capacity of VAE to project input images onto the high probability region of image manifold.
- Except for the regular off-manifold attacks, we also consider novel on-manifold attack which is even more challenging on face dataset. We demonstrate the effectiveness of our defense method against both above attacks.
- We introduce a simple training mechanism that makes the goal of training and testing more consistent, which enhances the robustness of our model.

2 Related works

2.1 Adversarial attacks

The existence of the adversarial examples is first shown in [43]. Regular adversarial attacks fool a well trained classifier through adding a small perturbation δ to a real image X. ϵ_p limits the l_p norm of the perturbations. We introduce several state-of-the-art attacks.

Fast Gradient Sign Method (FGSM) FGSM [43] generates adversarial examples by minimizing the probability of true class y:

$$X^{adv} = X + \epsilon_p \cdot sign(\nabla_x J(X, y)), \tag{1}$$

where $J(X, y)$ is cross-entropy loss. $sign(\cdot)$ is the sign function. FGSM performs one-step update towards the direction of gradient ascent.

Projected Gradient Descent (PGD) Madry et al. [30] improve FGSM by applying it multiple times with a small step size α. The adversarial examples can be formally expressed as:

$$X_0^{adv} = X, X_{n+1}^{adv} = Clip_{\epsilon_p}^{X}(X_n^{adv} + \epsilon_p \cdot sign(\nabla_x J(X_n^{adv}, y))), \tag{2}$$

where $Clip_{\epsilon_p}^{X}(\cdot)$ clips updated images to constrain it within the ϵ-ball of X. They also set the initial perturbation to a random point within the allowed ϵ-ball and restart the search multiple times to avoid falling into a local minimum.

Carlini & Wagner (C&W) C&W [3] is a strong optimization-based attack method. They find the smallest perturbations by optimizing the L_p norm of perturbation δ iteratively:

$$\arg\min_{\delta} \quad \|\delta\|_p + c \cdot \mathcal{L}(X + \delta), \tag{3}$$

where the loss function $\mathcal{L}$ is chosen to make the examples to be misclassified and c is a hyperparameter.

2.2 Adversarial attacks on face recognition

Face recognition relies on distinguishing subtle differences, which often concentrated on small landmark locations, so it is more vulnerable to adversarial attacks. Different from the attacks described above, many attacks on face recognition

are more concerned with the semantic structure of the face. Sharif et al. [37] employ a pair of eyeglass frames to fool face recognition systems. In [4], the adversarial face images are created by manipulating landmark locations. With the help of GANs, AdvFaces [5] learns to generating minimal perturbations in the salient facial regions. While regular attacks usually produce adversarial images leave the manifold, semantic-based attacks can generate high-fidelity images on manifold, which is more challenging.

2.3 Adversarial defense

Adversarial defense methods can be roughly categorized into two groups:

Target model enhancement: Adversarial training [13] is the most popular defense method by injecting adversarial examples into training datasets. This approach works well for adversarial trained attacks but remains vulnerable to black-box attacks. He et al. [17] propose a trainable parametric noise injection technique to improve model regularization. Xu et al. [14] squeeze the images via color bit depth squeezing and spatial smoothing. In [46], Xie et al. propose to use feature denoising to increase adversarial robustness.

Input-transformation: As a pre-processing strategy, input-transformation is model-agnostic and can complement other defenses. HGD [27] trains a image denoiser using a loss function defined by the difference in the top layers of the target model. Liu et al. [29] propose a feature distillation method to effectively defend against adversarial examples. Recently, a variety of work assume that the adversarial images leaves the manifold and aim to project it back. Among them, PixelDefend [38] and Defense-GAN [35] leverage generative networks to transform adversarial images into clean images. The disadvantage to these approaches is that they are limited by the expressiveness of generative networks, and cannot be applied to large-scale datasets. Abhimanyu et al. [10] approximate the image manifold using a web-scale image datasets. The main idea here is to localize nearest neighbors of adversarial image in the image datasets and classify them instead of the adversarial image.

2.4 Generative Adversarial Networks(GANs)

GANs was first introduced by Goodfellow et al. [12], it has important applications in various fields, such as image generation [21, 33] and image-to-image translation [47, 32, 19]. The GANs framework has an excellent capacity to fit data distribution because of its min-max two-player game mechanism. Isola et al. [19] have shown that conditional generative adversarial networks are competent at image-to-image translation tasks. Variational Autoencoder (VAE) [24] consists of an encoder that represents the input as a distribution over the latent space and a decoder that reconstructs the input from the latent code. Donahue et al. [7] employ bidirectional GANs to learn an inverse mapping from data to latent representation. Analogously, ALI [11] proposes the same framework to learn mutually inference. In our work, to learn image distributions, we use a VAE-based architecture with adversarial training. Different from the prior work

on image-to-image translation tasks, we replace the explicit reconstruction loss with GANs to mitigate the effects of small disturbances in the input.

3 Method

3.1 Motivation

As mentioned in Section 2, we conclude two types of attacks: off-manifold attacks and on-manifold attacks. On-manifold attacks can create adversarial images which are perceptually close to the original images, e.g., advFaces [4], A^3GN [26] and GFLM [3]. Therefore, we cannot make the common assumptions that all the adversarial images are out of the image manifold, but we assume that they are moved away from data distribution of high probability [38], at least.

These observations inspire us to project the data back onto the region of high probability. We use VAEs estimate the log-likelihood $\log p(x)$ of data x with a posterior $p(z|x)$ and a conditional distribution $q(z|x)$:

$$\log p(x) \geq -D_{KL}(q(z|x)||p(z)) + \mathbb{E}_{q(z|x)}[log(p(x|z))], \tag{4}$$

where $p(z)$ is a prior distribution and D_{KL} is the Kullback-Leibler divergence. By learning a probabilistic encoder $E(x)$ to represent $q(z|x)$ and an probabilistic decoder $Dec(z)$ to represent $p(x|z)$, we can find a tight estimate of the lower bound on the likelihood of images by optimizing z.

The key to solving the problem is that, the two terms in Equation 4 correspond to the two types of attacks: For attacks leave the manifold, the adversarial images are outside of the distribution and it causes reconstruction loss inevitably (the second term). For on-manifold attacks, the adversarial images deviate from the high probability regions of the image distribution, it will increase the KL divergence of the posterior from the prior (the first term). Both types of adversarial images reduce the estimation of the likelihood, so the model can distinguish them from legal images and try to project them back.

Another intractable problem is how to accurately restore high fidelity faces. For general datasets, e.g. ImageNet [34], F-MNIST [45] and MNIST [26], the defense methods may only need to extract the coarse-grained information to satisfy the recognition conditions, such as shape and contour, while face recognition pays more attention to the distinction in local facial details. Therefore, face reconstruction quality is the premise of defense performance. Since the adversarial disturbance is imperceptible, we make a hypothesis that an image consists of low-level information and high-level information, and adversarial attacks ruin or manipulate with high-level information, but the low-level information is retained. Instead of hallucinating the whole image, we constrain the expression space of the model to the high-level part of the images, so the network search in a smaller image space and less burden of generation.

In the next section, by constructing a specific network architecture and training strategy, we try to subtly achieve the above theories to move the adversarial

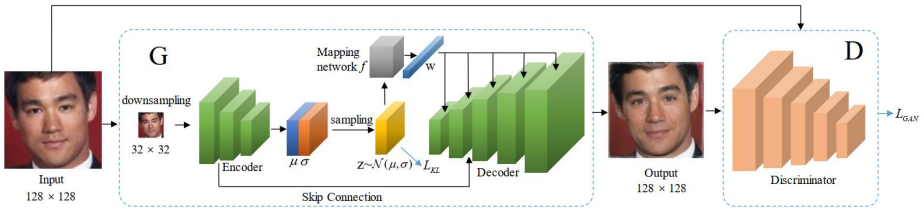

Fig. 2. The architecture of our network. Before passing the input layer, the input image is first downsampled to 32×32. The mapping network f transforms the latent code z to a new latent code w, which controls the variation of output images. The loss function consists only KL divergence loss and adversarial loss, does not include pixel-wise loss.

face images back to the high probability regions. The proposed model consists of a generator G and a discriminator D. The overall framework is shown in Figure 2.

3.2 Objective

To optimize the lower bound on the likelihood of datapoint x, the objective of a VAE can be expressed as

$$\mathcal{L}_{VAE} = D_{KL}(q(z|x)||p(z)) - \mathbb{E}_{q(z|x)}[log(p(x|z))]. \tag{5}$$

The first term tries to move the approximate posterior $q(z|x)$ closer to the prior $p(z)$, while the second term acts as an reconstruction error. Normally, the reconstruction loss will be Euclidean distance.

However, in our case, we *do not* adapt any explicit reconstruction loss. The reason is that, the pixel-based loss prompt the generator to faithfully preserve all the information from the input. During training, the training images are clean so there are no problems. However, in the testing phase, the input may be an adversarial image, and the generator will also tend to retain the perturbations, which is unreasonable. In this case, we use the adversarial loss to act as an implicit reconstruction loss:

$$\mathcal{L}_{GAN} = \mathbb{E}_x[logD(x)] + \mathbb{E}_x[log(1 - D(G(x)))], \tag{6}$$

where G is composed of the encoder E and the decoder Dec, and D is an discriminator. The generator G learns to synthesize images which cannot be distinguished by D, while D tries to classify the generated images and real images. Intuitively, during the training process, the generator will treat the input as a significant reference and try to extract useful components from it. Therefore, we expect the generator to generate images similar to the input in some ways. If this goal can be realized, the implicit loss will be better than the traditional loss. In this way, the reconstruction loss can be learned by the network during the training process, rather than being heuristically stipulated. With a large number

of experiments, we implement this using a specific network architecture, which will covered in Section 3.3.

Hence, the final objective function for A-VAE is

$$\arg\min_{G}\max_{D}\mathcal{L}_{GAN} + KL(q(z|x)||p(z)), \tag{7}$$

where $p(z)$ is assumed to be $\mathcal{N}(\mathbf{0}, \mathbf{I})$. Ablation study of loss function is shown in Appendix A.

3.3 Architecture of generator

We employ variational autoencoder networks as the basic framework of the generator. Since the adversarial perturbations only changes the high-level information in the images, we then design the network around this core.

As mentioned in Section 3.2, We expect the network to extract semantic information from the input. However, we found that when the input size is consistent with the output, the network just copies the input as it is, so that the network does not have to understand the image content. To solve this problem, we perform a downsampling operation on the input images. This strategy motivates the generator to focus the attention on low-level information and create some vivid details to complement the input faces. The prerequisite of this is that the generator must comprehend the semantic information of the input. Downsampling also mitigates the impact of the adversarial perturbations on the generator.

Given the latent code z generated by the encoder E, we aim to make it control the high-level variations of the synthesized images. A traditional decoder receives z at the input layer and transmits the required information of the upper layers by consuming the capacity of the network, which greatly limits the expressiveness power. Inspired by the recent success of style-based generator [22], we map z to an intermediate latent space $\mathcal{W}$ using a mapping network f, and inject it into each convolution layer of the decoder Dec through AdaIN operation [22]. The mapping network f is composed of 4 fully connected layers. Furthermore, we add a skip connection between two layers in the network to help the network transmit information more easily and reduce the burden on the bottleneck layer. Architecture details of A-VAE can be found in Appendix D.

3.4 Inference

At inference time, given an image x, the downsampled version is fed into the encoder E, then the encoder randomly sample M latent code z to generate different output images with diverse details, where z is clipped to meet a prerequisite that $KL(\mathcal{N}(z, \sigma^2\mathbf{I})||\mathcal{N}(\mathbf{0}, \mathbf{I})) \leq \tau$, and τ is a threshold. Finally, we calculate Euclidean distance in pixel-level between these images and the original input image, and take the smallest one as the input of the face recognition model.

3.5 Discussion

Relevance to existing methods As a pre-processing defense using generative models, our work is most similar to Defense-GAN [35]. It also finds a closest clean output which exists in image manifold to a given image. However, there are several fundamental differences between our method and Defense-GAN. First, we utilize a whole encoding-decoding network to generate the output, which has an input image as a valuable prior. Yet, Defense-GAN can only rely on random latent code to generate the image. Therefore, no matter which image is given, the potential generation space is always the entire image space, it needs better expressive power to recover the same quality image as ours, which is very tricky on face datasets. Second, at inference time, we find the nearest neighbor of input in a large number of randomly sampled output images, but Defense-GAN approximates the input by optimizing the latent code. Under on-manifold attacks, optimizing the latent code will make the generator reconstruct the image which is the same as the input, thus making the defense meaningless.

In [9, 50], nearest neighbor search is also used to defend against adversarial images, equipped with a web-scale image dataset. The dataset needs to collect large-scale images and register in advance with images having same identities as test images, both of which are unrealistic in practical applications, especially for face recognition. Furthermore, both of them show their defense methods are more easily compromised by white-box attacks, we believe that one main reason is because they use the extracted features for nearest neighbor retrieval, but the feature extractor itself is vulnerable to adversarial attacks. On the contrary, we perform the nearest neighbor retrieval directly at the pixel-level.

A-VAE resembles ALI [11] and BIGAN [7], which have three components in training process: an encoder, a decoder and a discriminator. They ask the discriminator to distinguish generated data from real images, and between latent code z from the posterior distribution $q(z|x)$ and prior distribution $p(z)$. In our case, the KL-divergence term is employed to supervise the approximate posterior, and we replace the explicit reconstruction loss with GANs to learn a similarity metric. The main difference is that they expect the model to learn meaningful hidden features, whereas we are looking for a robust reconstruction process. It is worth mentioning that although ALI's inference network also samples stochastic latent code, unlike our method, it does not get different reconstructions of a same input.

4 Experiments

4.1 Experimental settings

Training. We train our models on the publicly available CASIA-WebFace dataset [49] consists of 494,414 face images belonging to 10,575 different individuals. We use aligned CASIA-WebFace and crop to 128 $\times$ 128 image size. The hyperparameters λ of the loss function is empirically set to 1. We use ADAM optimizers [23] with β_1=0 and $\beta_2 = 0.99$. We train A-VAE for 140,000 steps with the batch

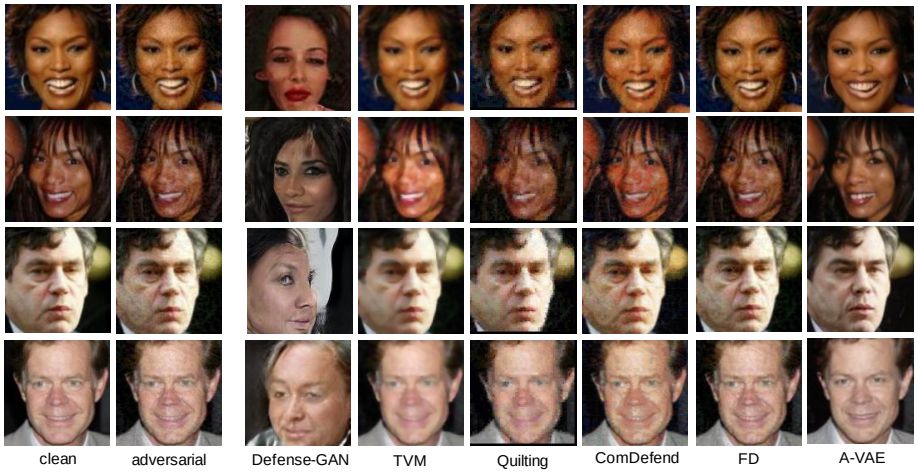

Fig. 3. Qualitative comparison results of image reconstruction results on LFW. The adversarial images is generated by FGSM grey-box attacks with $\epsilon = 8$. FD refers to Feature Distillation [29].

size of 16 on a single NVIDIA GTX 2080Ti. The learning rate is set to 0.001 during the training process.

Testing. We test our models on LFW [18]. LFW is a standard face verification testing dataset which includes 13,233 web-collected face images from 5,749 identities. We evaluate the verification accuracy on the 6,000 face pairs. Among them, 3,000 pairs are from the same identity and another 3,000 pairs represent the different identities. sampling times M is 1000, and threshold τ is set to 0.03.

Adversarial attacks. We evaluate the performance of A-VAE against the state-of-the-art attacks including FGSM [43], PGD [30] and C&W [3]. C&W is constrained by L_2 norm with an allowed maximum value ϵ, and others are constrained by L_∞ norm. We implement the C&W attack with learning rate 10.0. We consider grey-box attacks and white-box attacks. All attack methods performs obfuscation attacks on same identity pairs, and impersonation attacks on different identities pairs. We perform attacks for images in the probe set.

Evaluation metrics. We evaluate our method on a ResNet-50 [16] trained from VGG-Face2 [2]. More results on ArcFace [6] are given in Appendix E. The target model verifies whether the two images belong to one person by cosine distance between them. We measure the percentage of pairs which are successfully verified. The cosine distance threshold is set at 1% FAR.

4.2 Qualitative results

We present our image reconstruction results against FGSM black-box attacks with $\epsilon = 8$ on LFW and compare to other input-transformation defense methods in Figure 3, including Defense-GAN [35], TVM [15], Quilting [15], ComDefend

Fig. 4. stochastic generated results on LFW. Each line presents input image(leftmost) and different output of latent code.

[20] and Feature Distillation [29]. As can be seen, limited by expressive power, the images reconstructed by Defense-GAN [35] have a huge gap with the original images, so that the identity information is completely lost. TV minimization [15] almost completely eliminates adversarial perturbations, but makes the images blurry at the same time, which may lead to the loss of some crucial details. Compared to other methods, A-VAE is the only one that generates high-fidelity face images and captures the true identity information under large perturbations. The qualitative results show robustness of A-VAE to adversarial attacks on face images. Since it is just constrained by implicit reconstruction loss, our model can determine the important components of the image and focus on reconstructing it. This property relies on the model to understand the image from the high-level semantic features, instead of the pixels.

Figure 4 shows stochastic generated results of the same input on LFW, using different latent code. We can find that the variation in latent code has a significant impact on the appearance of the eyebrows, eyes, nose and mouth. These factors combine to express a wide variety of different identities, which allows our model to mine appropriate image that match the original identity well. More qualitative results are shown in Appendix F.

4.3 Gray-box attacks

In this section, we evaluate the performance of A-VAE against the gray-box attacks. In this setting, the attacker knows the details of the classifier but does not have access to the details of the defense strategy. The accuracy comparison results on the LFW dataset is shown in Table 1. Adversarial training [13], feture denoising [46] and HGD [27] obtain the FGSM adversarial images with $\epsilon = 8$. The evaluations involve the state-of-the-art attack methods including FGSM,

Table 1. Verification accuracies of different defense methods on the LFW dataset, under FGSM, PGD, C&W grey-box attacks.

LFW (Same identity pairs/Different identities pairs/Average)					
Defense	clean	FGSM $\epsilon = 4$	FGSM $\epsilon = 8$	PGD $\epsilon = 8$	C&W
No Defense	0.992/0.992 /0.992	0.487/0.417 /0.452	0.190/0.300 /0.245	0.000/0.007 /0.003	0.000/0.017 /0.008
Adversarial Training [13]	0.981/0.993 /0.987	0.513/0.787 /0.650	0.177/0.737 /0.457	0.023/0.190 /0.107	0.000/0.417 /0.208
Feature Denoising [46]	0.950/0.953 /0.952	0.647/0.717 /0.682	0.213/0.730 /0.472	0.073/0.260 /0.167	0.020/0.570 /0.295
TVM [15]	**0.990**/0.991 /0.991	0.737/0.683 /0.710	0.343/0.357 /0.350	0.307/0.353 /0.330	0.007/0.020 /0.013
Quilting [15]	0.980/0.993 /0.987	0.813/0.890 /0.852	0.593/0.680 /0.637	0.667/0.783 /0.725	0.230/0.037 /0.134
ComDefend [20]	0.989/0.990 /0.990	0.523/0.637 /0.630	0.281/0.389 /0.335	0.022/0.148 /0.085	0.000/0.017 /0.008
Feature Distillation [29]	**0.990**/0.993 **/0.992**	0.667/0.580 /0.623	0.383/0.380 /0.382	0.143/0.190 /0.167	0.003/0.027 /0.015
HGD [27]	0.943/**1.000** /0.972	0.650/0.860 /0.755	0.387/0.790 /0.588	0.150/0.647 /0.398	0.040/0.597 /0.318
A-VAE	0.927/**1.000** /0.963	**0.830/0.957 /0.893**	**0.637/0.863 /0.753**	**0.697/0.960 /0.828**	**0.423/0.797 /0.610**

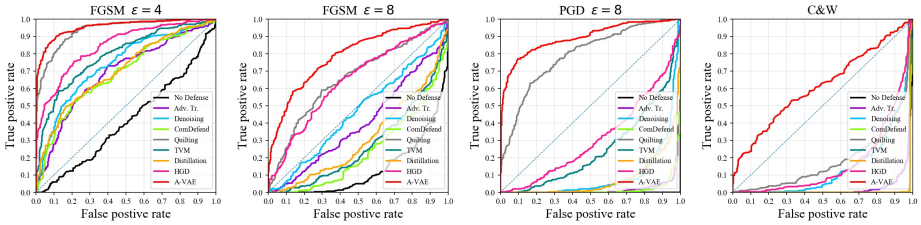

Fig. 5. ROC curves of different defense methods under FGSM, PGD and C&W attacks (Setting: gray-box, LFW).

PGD, C&W. It has been shown that A-VAE significantly improves the accuracy against off-manifold attacks and also outperforms other defense methods prominently. For example, although adversarial training defense with FGSM images achieves an average accuracy of 0.457 against FGSM attacks with $\epsilon = 8$, our method achieves an accuracy of 0.753 without any knowledge of adversarial images. Moreover, adversarial training has a poor generalization across different attacks, however, our method does not over-fit a specific attack. Figure 5 shows the Receiver Operating Characteristic (ROC) curves for different defense methods. We further show that our method also achieves better performance at different resolutions (Appendix B).

4.4 White-box attacks

We present experimental results on white-box attacks using FGSM, PGD and C&W. Since our method is non-differentiable, to perform white-box attacks, we apply Backward Pass Differentiable Approximation (BPDA) [1] to estimate the

Table 2. Verification accuracies of different defense methods on the LFW dataset, under FGSM, PGD, C&W white-box attacks.

LFW (Same identity pairs/Different identities pairs/Average)				
Defense	clean	FGSM $\epsilon = 4$	FGSM $\epsilon = 8$	PGD $\epsilon = 8$
No Defense	0.992/0.992/0.992	0.487/0.417/0.452	0.190/0.300/0.245	0.000/0.007/0.003
adversarial Training [13]	0.981/0.993/0.987	0.363/0.650/0.507	0.177/0.610/0.393	0.000/0.010/0.005
Feature Denoising [46]	0.950/0.953/0.952	0.401/0.440/0.423	0.170/0.450/0.310	0.000/0.020/0.010
ComDefend [20]	**0.989**/0.990/**0.990**	0.467/0.543/0.507	0.303/0.513/0.408	0.187/0.277/0.232
HGD [27]	0.943/**1.000**/0.972	0.243/0.677/0.460	0.103/ 0.627/0.365	0.323/0.713/0.518
A-VAE	0.927/**1.000**/0.963	**0.720/0.743/0.732**	**0.468/0.637/0.552**	**0.557/0.594/0.573**

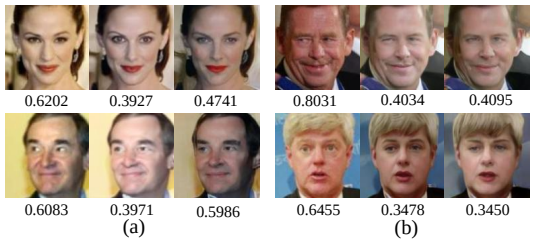

Fig. 6. Reconstruction results on the LFW dataset, under the on-manifold attack. Each pair presents clean face (left), adversarial face (middle) and reconstruction face (right). Cosine similarity score is calculated by comparing to gallery image.

gradient of the output of the classifier with respect to the input. We compare our defense with adversarial training [13], feature denoising [46], ComDefend [20] and HGD [27]. ComDefend causes gradient masking by adding random Gaussian noise in the image compression process, hence we employ Expectation over Transformation (EOT) [1] to correctly compute the gradient. Table 2 shows the accuracy comparison results on LFW. By comparison with Table 1, We can see that A-VAE does not suffer seriously when the attacker knows the defense strategy. Through this phenomenon, we conclude that our method does not entirely rely on gradient masking to improve robustness.

4.5 On-manifold attacks

On-manifold attacks generate high-quality adversarial images which are perceptually similar to the original images, e.g., GFLM [4] and advFaces [5]. However, in practice, we found when they deceive the target model successfully, the real label of the adversarial images is also changed. In other words, neither of them achieve legitimate adversarial face images that only causing imperceptible perturbation. In our case, similar to [39], we directly compute the latent code z generated by the encoder to realize onmanifold attack, instead of original images. The latent code z is optimized n times with a step size ϵ:

$$z_0^{adv} = z, z_n^{adv} = z_{n-1}^{adv} + \epsilon \cdot \bigtriangledown_z J(F(Dec(z_{n-1}^{adv})), y), \tag{8}$$

where J is the cosine similarity loss. The on-manifold adversarial image is obtained by $Dec(z_n^{adv})$. We use $n = 5$ and $\epsilon = 20$. Figure 6 shows the reconstruction

Table 3. Verification accuracies of different defense methods on the LFW dataset, under on-manifold attack.

LFW (Same identity pairs/Different identities pairs/Average)	
Defense	On-manifold attack
No Defense	0.080/0.451/0.321
adversarial Training [13]	0.237/0.867/0.537
Feature Denoising [46]	0.273/**0.867**/0.570
TVM [15]	0.337/0.623/0.480
Quilting [15]	0.393/0.664/0.529
Feature Distillation [29]	0.327/0.593/0.460
ComDefend [20]	0.407/0.553/0.480
HGD [27]	0.301/0.543/0.422
A-VAE	**0.644**/0.687/**0.667**

results on LFW. As can be seen in Figure 6 (b), this attack still inevitably produces some images that are notably different from the original images, and our method cannot handle this situation. The verification performance is shown in Table 3.

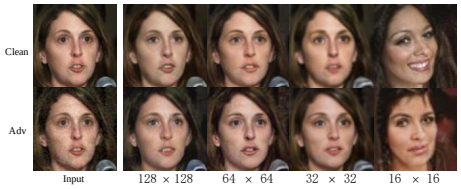

Fig. 7. Reconstruction results of different models.

4.6 Tradeoff between quality and robustness

In fact, A-VAE achieves the tradeoff between quality and robustness with a very simple strategy: We resize the scale of the input images. The more image information given, the model can supplement details in a smaller search space, the final resconstructed image will be more similar, but the adversarial perturbations will tend to be preserved. Conversely, the less image information given, the adversarial perturbations remains less, which makes the model more robust but requires a better expressiveness. To explore the effects of different scaling weights, we construct multiple generators with input size of 128×128, 64×64, 32×32 and 16×16, and correspondingly increse or delete the number of downsampling layers in the encoder block.

The performance of different models on clean images and adversarial images is shown in Figure 7. Interestingly, when the input size drops to 16×16, the model seems to ignore the input and behaves like a standard GAN. The performance of different models on clean images and adversarial images is shown in Figure 8. As can be seen, when the input size increases, the gap between the two

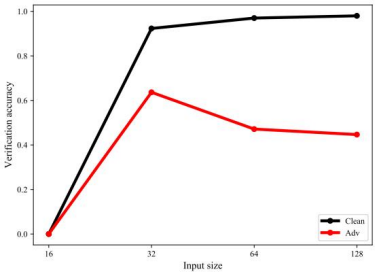

Fig. 8. Verification accuracies of different models on same identity pairs, using clean and adversarial images. (Setting: gray-box, FGSM with $\epsilon = 8$).

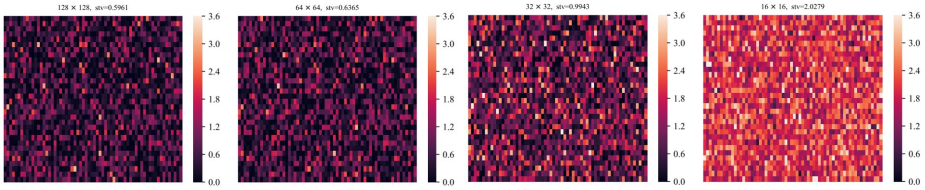

Fig. 9. Standard deviation of feature over 128 different realizations of the same input image, using different models. The feature of 2048 dimensions is shown as 32×64.

accuracies enlarge, which means that the robustness declines. The accuracy of clean images continues to rise, which means that the quality increases. Figure 9 further illustrates standard deviation of extracted feature over 128 different realizations of the same image. We find it interesting that the expression space of the model is negatively related to the input size, although we do not explicitly constrain the model to generate images that are consistent with the input. This proves that our implicit reconstruction loss is effective. We set the input size as 32×32 so that the model retains enough capacity to reconstruct the original image well.

5 Conclusion

In this paper, we have proposed an adversarial defense method which projects the image into the high probability regions of image manifold. By constructing a special architecture and training mechanism, we enhance the robustness against both off-manifold and on-manifold attacks. The evaluation results on LFW show the superiority of our method.

Acknowledgements This work is supported by National Nature Science Foundation of China (No. U1611461, U1903214, 61876135, 61862015), National Key R&D Program of China (No. 2017YFC0803700), National Nature Science Foundation of Hubei Province (2019CFB472) and Hubei Province Technological Innovation Major Project (2018AAA062, 2018CFA024, 2017AAA123).

References

1. Athalye, A., Carlini, N., Wagner, D.A.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: Dy, J.G., Krause, A. (eds.) Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018. Proceedings of Machine Learning Research, vol. 80, pp. 274–283. PMLR (2018), http://proceedings.mlr.press/v80/athalye18a.html
2. Cao, Q., Shen, L., Xie, W., Parkhi, O.M., Zisserman, A.: Vggface2: A dataset for recognising faces across pose and age. In: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018). pp. 67–74. IEEE (2018)
3. Carlini, N., Wagner, D.A.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017. pp. 39–57. IEEE Computer Society (2017). https://doi.org/10.1109/SP.2017.49, https://doi.org/10.1109/SP.2017.49
4. Dabouei, A., Soleymani, S., Dawson, J., Nasrabadi, N.: Fast geometrically-perturbed adversarial faces. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1979–1988. IEEE (2019)
5. Deb, D., Zhang, J., Jain, A.K.: Advfaces: Adversarial face synthesis. arXiv preprint arXiv:1908.05008 (2019)
6. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: Additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4690–4699 (2019)
7. Donahue, J., Krähenbühl, P., Darrell, T.: Adversarial feature learning. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017)
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
9. Dubey, A., van der Maaten, L., Yalniz, Z., Li, Y., Mahajan, D.: Defense against adversarial images using web-scale nearest-neighbor search. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 8767–8776. Computer Vision Foundation / IEEE (2019)
10. Dubey, A., Maaten, L.v.d., Yalniz, Z., Li, Y., Mahajan, D.: Defense against adversarial images using web-scale nearest-neighbor search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 8767–8776 (2019)
11. Dumoulin, V., Belghazi, I., Poole, B., Lamb, A., Arjovsky, M., Mastropietro, O., Courville, A.C.: Adversarially learned inference. In: 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings. OpenReview.net (2017)
12. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., Bengio, Y.: Generative adversarial nets. In: Advances in neural information processing systems. pp. 2672–2680 (2014)
13. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
14. Gu, S., Yi, P., Zhu, T., Yao, Y., Wang, W.: Detecting adversarial examples in deep neural networks using normalizing filters. UMBC Student Collection (2019)
15. Guo, C., Rana, M., Cissé, M., van der Maaten, L.: Countering adversarial images using input transformations. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 -

May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=SyJ7ClWCb
16. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
17. He, Z., Rakin, A.S., Fan, D.: Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 588–597 (2019)
18. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments (2008)
19. Isola, P., Zhu, J.Y., Zhou, T., Efros, A.A.: Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1125–1134 (2017)
20. Jia, X., Wei, X., Cao, X., Foroosh, H.: Comdefend: An efficient image compression model to defend adversarial examples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 6084–6092. Computer Vision Foundation / IEEE (2019)
21. Karras, T., Aila, T., Laine, S., Lehtinen, J.: Progressive growing of gans for improved quality, stability, and variation. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018)
22. Karras, T., Laine, S., Aila, T.: A style-based generator architecture for generative adversarial networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 4401–4410 (2019)
23. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings (2015), http://arxiv.org/abs/1412.6980
24. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), http://arxiv.org/abs/1312.6114
25. Kurakin, A., Goodfellow, I., Bengio, S.: Adversarial examples in the physical world. arXiv preprint arXiv:1607.02533 (2016)
26. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proceedings of the IEEE **86**(11), 2278–2324 (1998)
27. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018. pp. 1778–1787. IEEE Computer Society (2018)
28. Liu, J., Deng, Y., Bai, T., Wei, Z., Huang, C.: Targeting ultimate accuracy: Face recognition via deep embedding. arXiv preprint arXiv:1506.07310 (2015)
29. Liu, Z., Liu, Q., Liu, T., Xu, N., Lin, X., Wang, Y., Wen, W.: Feature distillation: Dnn-oriented JPEG compression against adversarial examples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 860–868. Computer Vision Foundation / IEEE (2019)
30. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April

30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=rJzIBfZAb
31. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
32. Park, T., Liu, M., Wang, T., Zhu, J.: Semantic image synthesis with spatially-adaptive normalization. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 2337–2346. Computer Vision Foundation / IEEE (2019)
33. Radford, A., Metz, L., Chintala, S.: Unsupervised representation learning with deep convolutional generative adversarial networks. In: Bengio, Y., LeCun, Y. (eds.) 4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings (2016)
34. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al.: Imagenet large scale visual recognition challenge. International journal of computer vision **115**(3), 211–252 (2015)
35. Samangouei, P., Kabkab, M., Chellappa, R.: Defense-gan: Protecting classifiers against adversarial attacks using generative models. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=BkJ3ibb0-
36. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
37. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: 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. pp. 1528–1540 (2016)
38. Song, Y., Kim, T., Nowozin, S., Ermon, S., Kushman, N.: Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In: 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings. OpenReview.net (2018), https://openreview.net/forum?id=rJUYGxbCW
39. Stutz, D., Hein, M., Schiele, B.: Disentangling adversarial robustness and generalization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 6976–6987 (2019)
40. Sun, B., Tsai, N.h., Liu, F., Yu, R., Su, H.: Adversarial defense by stratified convolutional sparse coding. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 11447–11456 (2019)
41. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in neural information processing systems. pp. 1988–1996 (2014)
42. Sun, Y., Liang, D., Wang, X., Tang, X.: Deepid3: Face recognition with very deep neural networks. arXiv preprint arXiv:1502.00873 (2015)
43. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I.J., Fergus, R.: Intriguing properties of neural networks. In: Bengio, Y., LeCun, Y. (eds.) 2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings (2014), http://arxiv.org/abs/1312.6199

44. Taigman, Y., Yang, M., Ranzato, M., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1701–1708 (2014)
45. Xiao, H., Rasul, K., Vollgraf, R.: Fashion-mnist: a novel image dataset for benchmarking machine learning algorithms. arXiv preprint arXiv:1708.07747 (2017)
46. Xie, C., Wu, Y., Maaten, L.v.d., Yuille, A.L., He, K.: Feature denoising for improving adversarial robustness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 501–509 (2019)
47. Xue, Y., Xu, T., Zhang, H., Long, L.R., Huang, X.: Segan: Adversarial network with multi-scale L 1 loss for medical image segmentation. Neuroinformatics **16**(3-4), 383–392 (2018)
48. Ye, M., Liang, C., Yu, Y., Wang, Z., Leng, Q., Xiao, C., Chen, J., Hu, R.: Person reidentification via ranking aggregation of similarity pulling and dissimilarity pushing. IEEE Transactions on Multimedia **18**(12), 2553–2566 (2016)
49. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. CoRR **abs/1411.7923** (2014), http://arxiv.org/abs/1411.7923
50. Zhao, J., Cho, K.: Retrieval-augmented convolutional neural networks against adversarial examples. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019. pp. 11563–11571. Computer Vision Foundation / IEEE (2019)