Improving Transferability of Adversarial Patches on Face Recognition with Generative Models: The Appendix

A. Implementation details

A.1. Hyperparameters

TAP algorithms: We set the number of iterations $N = 400$, the learning rate $\alpha = 1$ and the decay factor $\mu = 1$. We use $\epsilon = 255$ for dodging and $\epsilon = 40$ for impersonation.

GenAP algorithms: We set the number of iterations $N = 100$, and the learning rate of Adam optimizer $\alpha = 0.01$.

A.2. Models

Face recognition models: All face recognition models are accessible from the Internet, including FaceNet[1], CosFace[2], ArcFace[3], Face++[4] and Aliyun[5].

Generative models: All generative models are accessible from the Internet, including ProGAN[6], StyleGAN[7], StyleGAN2[8].

B. Additional experiments on adversarial eyeglass frame

In the main text, we present the attack success rates on using adversarial eyeglass frames to perform dodging and impersonation attacks on the face verification task. In this section, we present the success rates on dodging and impersonation attacks on the face identification task in Tab. 1 and 2. The conclusions on these results are consistent.

Moreover, we visualize the adversarial examples generated by the TAP-TIDIM and the GenAP-DI methods for dodging attack (see Fig. 1) and impersonation attack (see Fig. 2), respectively. The proposed GAP methods generates face-like features as perturbations.

[1] https://github.com/timesler/facenet-pytorch
[2] https://github.com/MuggleWang/CosFace_pytorch
[3] https://github.com/TreB1eN/InsightFace_Pytorch
[4] https://www.faceplusplus.com/face-comparing/
[5] https://vision.aliyun.com/facebody
[6] https://github.com/tkarras/progressive_growing_of_gans/tree/master
[7] https://github.com/NVlabs/stylegan
[8] https://github.com/NVlabs/stylegan2

C. Experiments on adversarial respirators

In this section, we present the results on adversarial respirators to show the generalization of the proposed methods to different regions of the adversarial patches. Tab. 3, 4 show the results on dodging and impersonation attack, respectively, on the face verification task. Tab. 5, 6 show the results on face identification task. The conclusions are consistent with those drawn from the adversarial eyeglass frames. We visualize the adversarial examples in Fig. 3 and Fig. 4.

D. Experiments on SemanticAdv

SemanticAdv [3] uses StarGAN [1] to adversarially perturb the attributes in a face image. This is very similar with our idea of using face-like features to generate adversarial perturbations. Nevertheless, SemanticAdv proposed to generate the adversarial example by interpolation between two feature maps. Their algorithm is designed for the imperceptible setting, where the adversarial perturbation should be indistinguiable from the original image. However, the slight perturbation limits their performance in the patch setting, where the perturbation is allowed to be perceptible. Instead, the proposed GenAP algorithms can leverage the relaxation on perceptibility to improve the attack success rates. We perform experiments to verify this point in the following.

Specifically, we use the code from SemanticAdv[9] to reproduce their results. We use the first 100 image pairs from their dataset (Celeba) to compare the algorithms. The code performs targeted attack on face verification. We modify their code to 1) perturb only the eyeglass frame region, 2) use our substitute models. Since their method generates adversarial examples for each attributes for each attacker-target pair, we consider their attack for an image pair is successful if the attack from any attribute is successful. The results are shown in Tab. 7. Our methods significantly outperforms theirs in the patch setting. In this setting, the slight perturbation generated by their method has difficulty in even white-box attacks.

[9] https://github.com/AI-secure/SemanticAdv

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	0.9875*	0.3000	0.7550	0.9875*	0.2150	0.5950
	TAP-TIDIM	**1.0000***	0.3750	0.8050	**1.0000***	0.3250	0.6625
	GenAP	0.9950*	**0.7100**	**0.9650**	**1.0000***	**0.5650**	**0.9250**
	GenAP-DI	**1.0000***	0.5975	0.9200	**1.0000***	0.4750	0.8500
CosFace	TAP-MIM	0.1600	0.9950*	0.8175	0.0525	0.9975*	0.7075
	TAP-TIDIM	0.0425	**1.0000***	0.6700	0.0275	**1.0000***	0.5425
	GenAP	**0.6700**	**1.0000***	**0.9675**	**0.5700**	**1.0000***	**0.9850**
	GenAP-DI	0.5350	**1.0000***	0.9400	0.3700	**1.0000***	0.9550
FaceNet	TAP-MIM	0.1075	0.2750	0.9950*	0.0425	0.2100	0.9900*
	TAP-TIDIM	0.0350	0.1900	**1.0000***	0.0125	0.1475	**1.0000***
	GenAP	**0.4825**	**0.4000**	0.9975*	**0.3375**	**0.3250**	0.9975*
	GenAP-DI	0.3425	0.2650	**1.0000***	0.1875	0.1675	0.9950*

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

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	0.6450*	0.1000	0.0700	0.6850*	0.0925	0.0650
	TAP-TIDIM	0.9050*	0.1200	0.1000	0.9275*	0.1175	0.0625
	TAP-TIDIMv2	0.9375*	0.2575	0.1900	**0.9425***	0.2075	0.0950
	GenAP	0.8625*	**0.3425**	**0.2750**	0.8675*	**0.2425**	**0.2150**
	GenAP-DI	**0.9625***	0.3225	0.2425	0.9400*	0.2325	0.1550
CosFace	TAP-MIM	0.1150	0.7425*	0.1525	0.1175	0.7250*	0.0825
	TAP-TIDIM	0.1200	0.9625*	0.1650	0.1100	0.9750*	0.1075
	TAP-TIDIMv2	0.1700	0.9875*	0.2275	0.1650	0.9900*	0.1500
	GenAP	**0.2975**	0.9425*	**0.3625**	**0.2750**	0.9350*	**0.2875**
	GenAP-DI	0.2375	**0.9950***	0.3550	0.2500	**0.9925***	0.2675
FaceNet	TAP-MIM	0.0325	0.0625	0.5600*	0.0375	0.0600	0.4900*
	TAP-TIDIM	0.0325	0.0750	0.8975*	0.0275	0.0700	0.9075*
	TAP-TIDIMv2	0.0775	**0.1500**	**0.9125***	0.0550	0.0800	**0.9225***
	GenAP	**0.0900**	0.1225	0.7900*	**0.1200**	**0.1100**	0.7450*
	GenAP-DI	0.0650	0.1050	0.9025*	0.0825	0.0800	**0.9225***

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

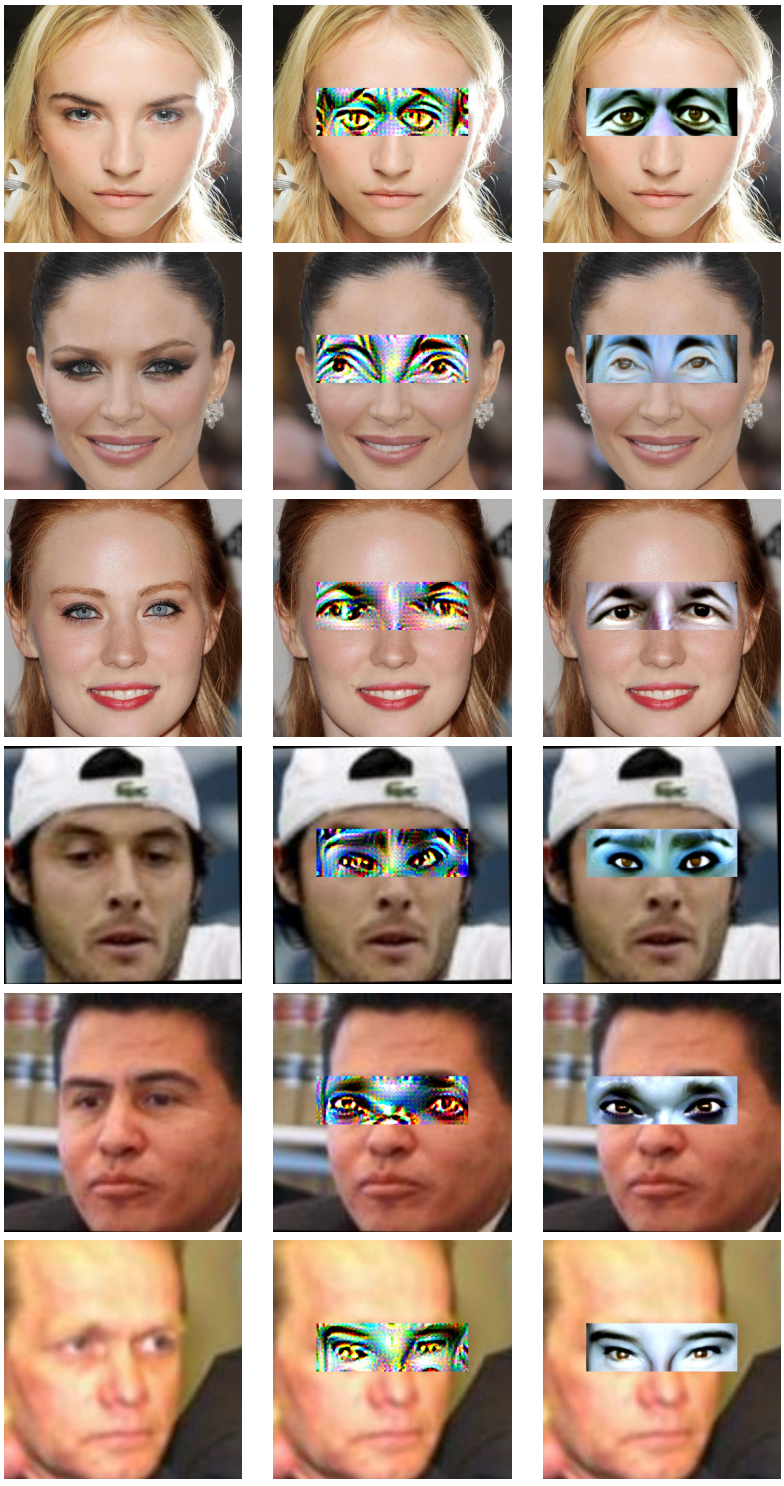
E. Experiments on image classification

The proposed GenAP methods can be easily extended to the image classification task by replacing the adversarial losses for face recognition (Eq. (2) in the main text) by the cross-entropy loss [2] widely used in image classification. This section shows the effectiveness of the GenAP algorithms on the image classification tasks.

We use two datasets, CIFAR10 and ImageNet. The images from these two datasets are 32×32 and 224×224, respectively. The adversarial patch region is designed to be a square at the center of the image. We observe the experimental results by changing the length of the square. The detailed information of the recognition models, generative models and lengths of the square patches are listed in Tab. 8. All models are accessible from the Internet[10] To evaluate the attack performance, we report the success rate (higher is better) as the fraction of adversarial images that are misclassified to the desired target class (*i.e.*, targeted attack). For each dataset, we randomly sample 1000 images. For each image, we randomly sample a distinct class as the target.

For the TAP algorithms, we set $\epsilon = 255$. For the GenAP algorithms on CIFAR10, we introduce several GenAP algorithms that optimize in different latent spaces. StyleGAN2-ADA is a conditional generative model. First, the GenAP-

[10]CIFAR recognition models: https://github.com/huyvnphan/PyTorch_CIFAR10, CIFAR10 generative models: https://github.com/NVlabs/stylegan2-ada-pytorch, ImageNet recognition models: https://pytorch.org/vision/stable/models.html, ImageNet generative models: https://github.com/ajbrock/BigGAN-PyTorch



Attacker TAP-TIDIM GenAP-DI

Figure 1. Visualization of adversarial eyeglass frames generated by the TAP-TIDIM and the GenAP-DI methods for dodging attack. The first three rows are the demonstrations on CelebA-HQ dataset and the others are from LFW dataset. And the three columns denotes the pictures of attackers and attackers with the adversarial eyeglass frames by generated by TAP-TIDIM and GenAP-DI methods separately. In TAP-TIDIM, we use $\epsilon = 255$.

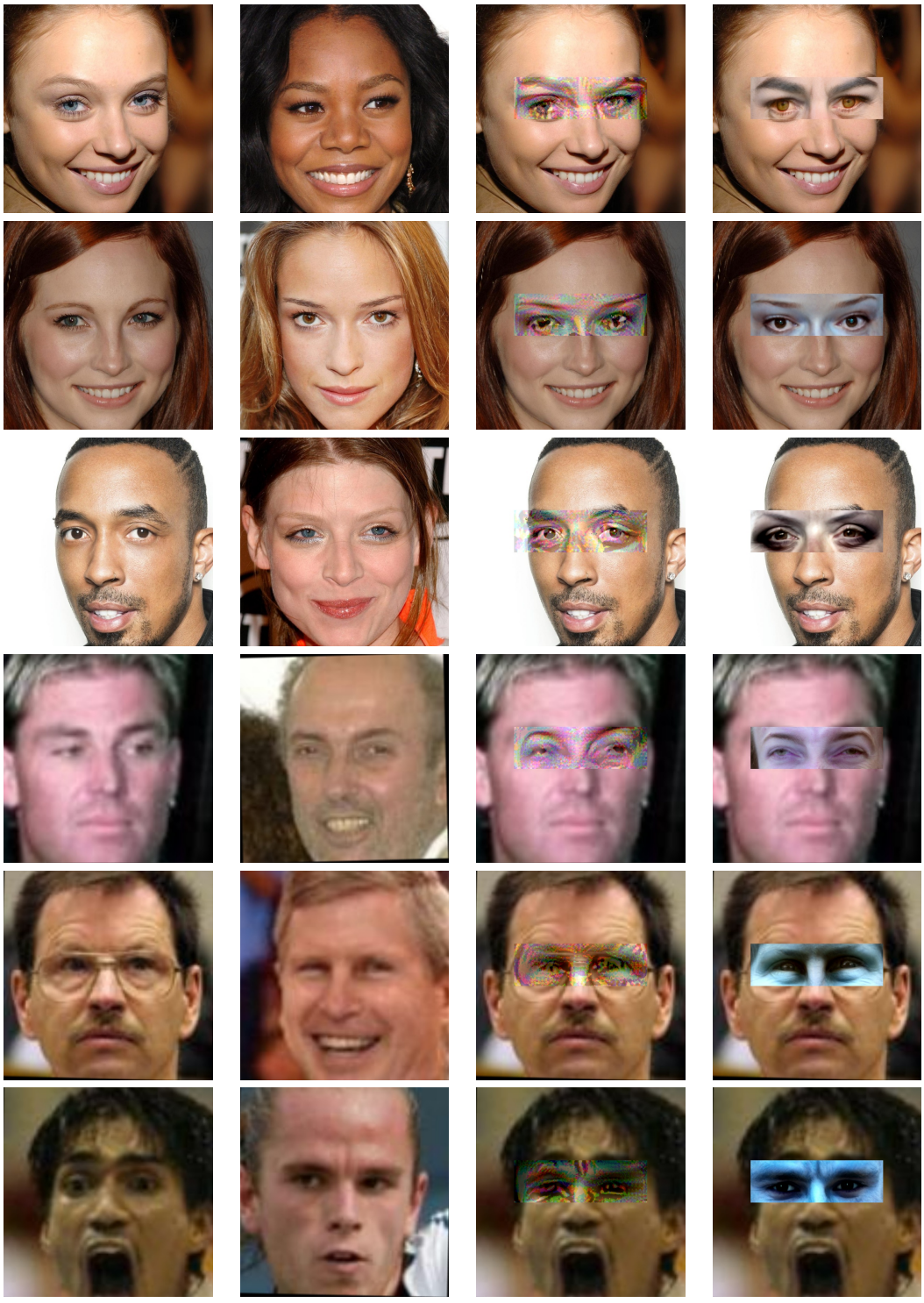


Figure 2. Visualization of adversarial eyeglass frames generated by the TAP-TIDIM and the GenAP-DI methods for impersonation attack. The first three rows are the demonstrations on CelebA-HQ dataset and the others are from LFW dataset. The first two columns are the photos of attackers and their target identities, and the following two columns show the attackers with the adversarial eyeglass frames generated by the proposed TAP-TIDIM and GenAP-DI methods. In TAP-TIDIM, we use $\epsilon = 40$.

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	**1.0000***	0.0800	0.3050	0.9950*	0.0600	0.1175
	TAP-TIDIM	**1.0000***	0.1175	0.3300	**1.0000***	0.0875	0.1275
	GenAP	**1.0000***	**0.3875**	**0.8875**	**1.0000***	**0.2950**	**0.7300**
	GenAP-DI	**1.0000***	0.3150	0.8150	**1.0000***	0.2675	0.6325
CosFace	TAP-MIM	0.0400	0.9975*	0.2600	0.0250	0.9975*	0.0975
	TAP-TIDIM	0.0650	0.9975*	0.2825	0.0200	**1.0000***	0.1150
	GenAP	0.5425	**1.0000***	**0.8650**	0.3975	**1.0000***	0.7350
	GenAP-DI	**0.5725**	**1.0000***	**0.8650**	**0.4025**	**1.0000***	**0.7750**
FaceNet	TAP-MIM	0.0475	0.0525	0.9675*	0.0125	0.0575	0.9425*
	TAP-TIDIM	0.0425	0.0425	0.9925*	0.0100	0.0375	0.9800*
	GenAP	**0.2200**	**0.1375**	0.9950*	0.1425	**0.1400**	0.9850*
	GenAP-DI	0.1925	**0.1375**	**0.9975***	**0.1500**	**0.1400**	**0.9900***

Table 3. The success rates of black-box dodging attack on FaceNet, CosFace, ArcFace in the digital world under the face verification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to a respirator region. * indicates white-box attacks.

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	**1.0000***	0.2625	0.1225	0.9850*	0.2450	0.1025
	TAP-TIDIM	**1.0000***	0.3125	0.1450	**1.0000***	0.2800	0.1125
	TAP-TIDIMv2	**1.0000***	0.4250	0.1975	**1.0000***	0.3725	0.1475
	GenAP	**1.0000***	**0.4850**	**0.3025**	0.9975*	**0.4675**	**0.2750**
	GenAP-DI	**1.0000***	0.4450	0.2800	**1.0000***	0.4525	0.2275
CosFace	TAP-MIM	0.4100	**1.0000***	0.1475	0.2500	0.9925*	0.1200
	TAP-TIDIM	0.4025	**1.0000***	0.1750	0.2225	**1.0000***	0.1300
	TAP-TIDIMv2	0.5425	**1.0000***	0.2250	0.3400	**1.0000***	0.1775
	GenAP	0.6250	**1.0000***	**0.2975**	0.5025	**1.0000***	0.2900
	GenAP-DI	**0.6325**	**1.0000***	**0.2975**	**0.5050**	0.9975*	**0.3100**
FaceNet	TAP-MIM	0.1925	0.1625	0.6525*	0.0900	0.1450	0.6325*
	TAP-TIDIM	0.1900	0.1950	**0.8700***	0.1000	0.1725	0.8550*
	TAP-TIDIMv2	**0.3700**	**0.2700**	0.8600*	**0.2275**	**0.2400**	**0.8575***
	GenAP	0.1950	0.1525	0.7075*	0.1475	0.1800	0.7550*
	GenAP-DI	0.2025	0.1550	0.7325*	0.1325	0.1700	0.7450*

Table 4. The success rates of black-box impersonation attack on FaceNet, CosFace, ArcFace in the digital world under the face verification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to a respirator region. * indicates white-box attacks.

cond algorithm uses the image directly generated by the generative model conditional on the target class. Second, the GenAP-DI-cond-opt algorithm is the GenAP-DI algorithm optimized in the $\mathbf{Z}$ space, with the conditional variable set to the target class. Third, the GenAP-DI-uncond-opt algorithm is a GenAP-DI algorithm optimized in the $\mathbf{W}^+$ plus the noise space. Similarly, GenAP algorithms are introduced on ImageNet. Since BigGAN is also a conditional generative model, we can also define the corresponding GenAP-cond and GenAP-DI-cond-opt algorithm with it.

The experimental results on CIFAR10 are shown in Tab. 9. We have the following observations. First, when the patch size is very small, the TAP algorithms achieve higher success rates on white-box attack by leveraging the larger search space (*i.e.*, without regularization). Second, optimizing the latent spaces of the generative models yield better results than naively using the inference results from the conditional generative model (GenAP-DI-uncond-opt and GenAP-DI-cond-opt v.s. GenAP-cond). Third, the GenAP-DI-cond-opt outperforms the TAP-TIDIM in black-box attacks when the patch size is as small as 8×8, which occupies 6% of the whole image.

The experiments results on ImageNet are shown in Tab. 10. The observations are consistent with those on CIFAR-10. The GenAP-DI-cond-opt outperforms the TAP-TIDIM in black-box attacks when the patch size is as small as 60×60, which occupies 7% of the whole image.

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	**1.0000***	0.1875	0.4800	**1.0000***	0.1075	0.2375
	TAP-TIDIM	**1.0000***	0.2350	0.4925	**1.0000***	0.1550	0.2550
	GenAP	**1.0000***	**0.5450**	**0.9400**	**1.0000***	**0.5025**	**0.8525**
	GenAP-DI	**1.0000***	0.4850	0.9000	**1.0000***	0.4400	0.7950
CosFace	TAP-MIM	0.1250	0.9975*	0.4500	0.0325	**1.0000***	0.2075
	TAP-TIDIM	0.1650	0.9975*	0.4500	0.0300	**1.0000***	0.2375
	GenAP	**0.7325**	0.9975*	**0.9250**	**0.5750**	**1.0000***	0.8850
	GenAP-DI	**0.7325**	**1.0000***	0.9150	0.5500	**1.0000***	**0.9075**
FaceNet	TAP-MIM	0.1300	0.1475	0.9800*	0.0425	0.1200	0.9650*
	TAP-TIDIM	0.0900	0.1325	0.9950*	0.0200	0.0750	0.9850*
	GenAP	**0.3500**	0.2725	**0.9975***	0.2425	0.3275	0.9875*
	GenAP-DI	0.3350	**0.2800**	**0.9975***	**0.2600**	**0.3425**	**0.9950***

Table 5. The success rates of black-box dodging attack on FaceNet, CosFace, ArcFace in the digital world under the face identification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to a respirator frame region. * indicates white-box attacks.

	Attack	CelebA-HQ			LFW		
		ArcFace	CosFace	FaceNet	ArcFace	CosFace	FaceNet
ArcFace	TAP-MIM	0.6650*	0.0575	0.0200	0.6555*	0.0600	0.0200
	TAP-TIDIM	0.9350*	0.0750	0.0275	0.9425*	0.0875	0.0300
	TAP-TIDIMv2	**0.9625***	0.1875	0.0675	**0.9650***	0.1500	0.0600
	GenAP	0.8800*	**0.2425**	**0.1325**	0.8800*	0.2025	**0.1275**
	GenAP-DI	0.9525*	0.1850	0.0975	0.9625*	**0.2075**	0.0975
CosFace	TAP-MIM	0.1100	0.6900*	0.0375	0.0800	0.7125*	0.0450
	TAP-TIDIM	0.1150	0.8875*	0.0375	0.0950	0.9425*	0.0450
	TAP-TIDIMv2	0.2125	**0.9375***	0.0825	0.1425	**0.9875***	0.0800
	GenAP	**0.2725**	0.8725*	0.1275	0.2200	0.9325*	0.1375
	GenAP-DI	0.2550	0.8825*	**0.1300**	**0.2275**	0.9225*	**0.1400**
FaceNet	TAP-MIM	0.0425	0.0325	0.3250*	0.0350	0.0275	0.3000*
	TAP-TIDIM	0.0475	0.0550	0.6375*	0.0300	0.0425	0.5825*
	TAP-TIDIMv2	**0.1275**	**0.0975**	**0.6800***	**0.0900**	**0.0850**	**0.6550***
	GenAP	0.0450	0.0550	0.4400*	0.0625	0.0725	0.5100*
	GenAP-DI	0.0525	0.0400	0.4700*	0.0575	0.0575	0.4950*

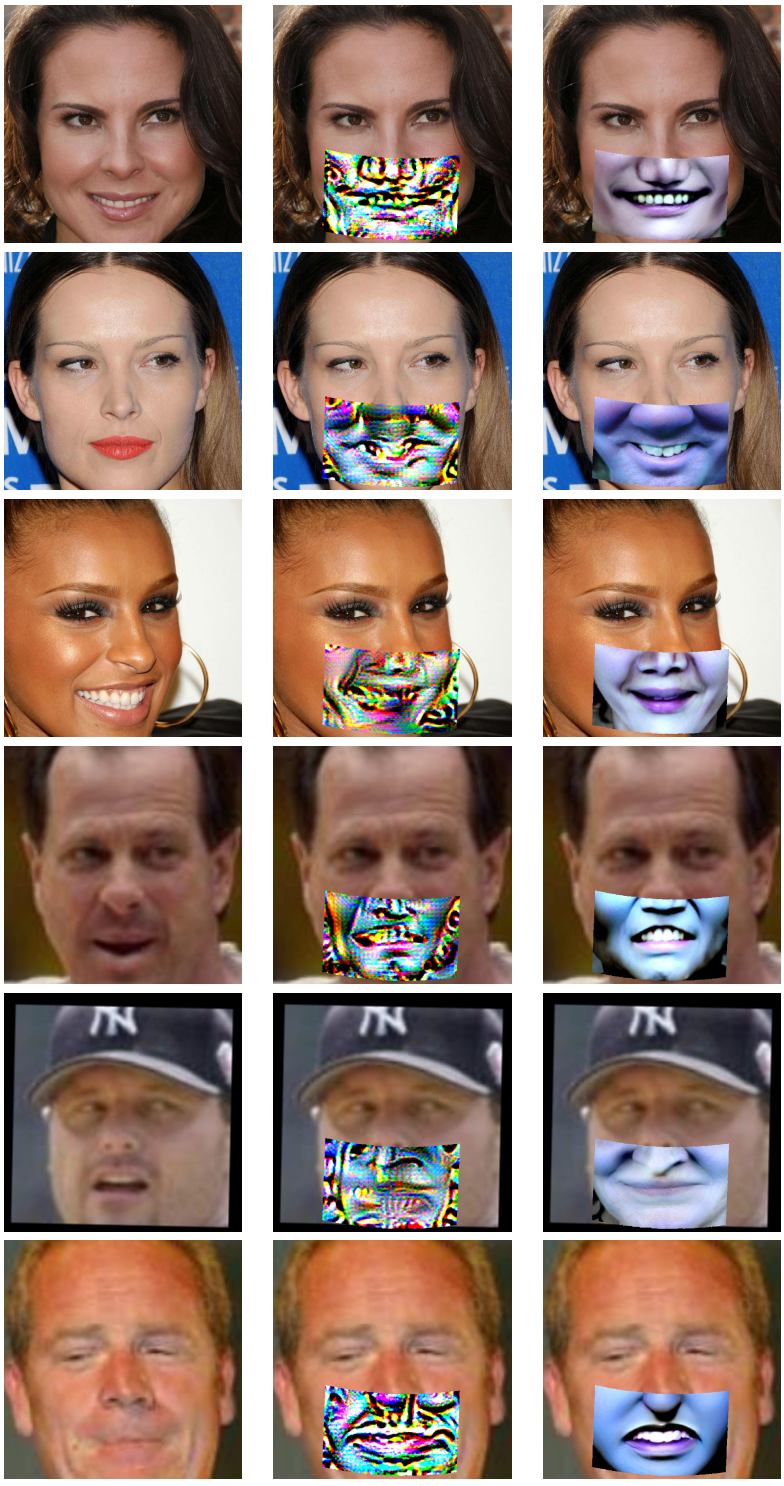
Table 6. The success rates of black-box impersonation attack on FaceNet, CosFace, ArcFace in the digital world under the face identification task. The adversarial examples are generated against FaceNet, CosFace, and ArcFace by restricting the adversarial patches to a respirator frame region. * indicates white-box attacks.

References

[1] Yunjey Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8789–8797, 2018. 1

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

[3] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020. 1



Attacker TAP-TIDIM GenAP-DI

Figure 3. Visualization of adversarial respirators generated by the TAP-TIDIM and the GenAP-DI methods for dodging attack. The first three rows are the demonstrations on CelebA-HQ dataset and the others are from LFW dataset. And the three columns denotes the pictures of attackers and attackers with the adversarial respirators by generated by TAP-TIDIM and GenAP-DI methods separately. In TAP-TIDIM, we use $\epsilon = 255$.

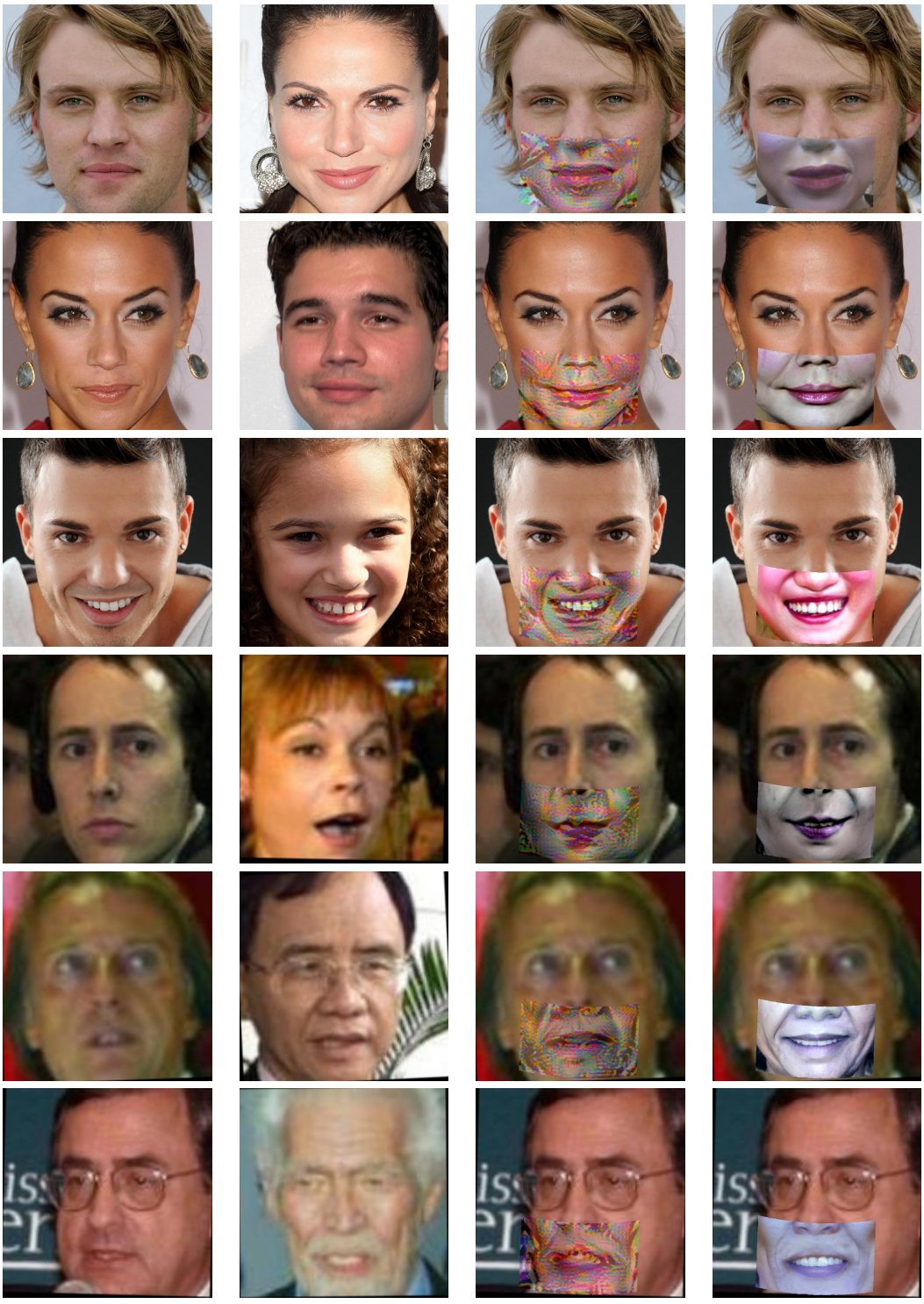


Figure 4. Visualization of adversarial respirators generated by the TAP-TIDIM and the GenAP-DI methods for impersonation attack. Table format follows the Fig.2. The first three rows are the demonstrations on CelebA-HQ dataset and the others are from LFW dataset. The first two columns are the photos of attackers and their target identities, and the following two columns show the attackers with the adversarial respirators generated by the proposed TAP-TIDIM and GenAP-DI methods. In TAP-TIDIM, we use $\epsilon = 40$.

Attack	ArcFace	CosFace	FaceNet
SemanticAdv	0.82	0.19	0.11
GenAP	**1.00**	**0.51**	**0.36**

Table 7. The attack success rate of black-box impersonation attack in face verification using SemanticAdv and GenAP on the CelebA dataset. ArcFace is the substitute model.

Dataset	Image classification models	Generative models	Lengths of the square patch
CIFAR10	ResNet50, MobileNetV2, InceptionV3, DenseNet121, VGG16	StyleGAN2-ADA	$8, 12, 16, 24$
ImageNet	ResNet101, DenseNet121, VGG16, ResNet50	BigGAN	$40, 80, 60, 80, 100, 120$

Table 8. Setting of the image classification experiments.

Patch size	Attack	ResNet50	MobileNetV2	InceptionV3	DesNet121	VGG16
8×8	TAP-TIDIM	**0.530***	0.101	0.114	0.194	0.248
	GenAP-DI-uncond-opt	0.128*	0.071	0.092	0.104	0.094
	GenAP-cond	0.037	0.036	0.040	0.043	0.034
	GenAP-DI-cond-opt	0.462*	**0.156**	**0.185**	**0.262**	**0.189**
12×12	TAP-TIDIM	0.793*	0.209	0.224	0.277	0.345
	GenAP-DI-uncond-opt	0.246*	0.170	0.198	0.218	0.193
	GenAP-cond	0.128	0.117	0.131	0.131	0.128
	GenAP-DI-cond-opt	**0.834***	**0.424**	**0.458**	**0.585**	**0.478**
16×16	TAP-TIDIM	0.865*	0.227	0.289	0.304	0.398
	GenAP-DI-uncond-opt	0.333*	0.260	0.254	0.270	0.238
	GenAP-cond	0.299	0.319	0.335	0.313	0.316
	GenAP-DI-cond-opt	**0.969***	**0.778**	**0.778**	**0.846**	**0.801**
24×24	TAP-TIDIM	0.888*	0.235	0.325	0.237	0.509
	GenAP-DI-uncond-opt	0.461*	0.400	0.398	0.398	0.360
	GenAP-cond	0.745	0.776	0.808	0.800	0.780
	GenAP-DI-cond-opt	**1.000***	**0.981**	**0.989**	**0.999**	**0.988**

Table 9. The success rates of black-box targeted attack on CIFAR-10 dataset using adversarial patches. The adversarial patches are generated against ResNet50. * indicates white-box attacks.

Patch size	Attack	ResNet101	DesNet121	VGG16	ResNet50
40×40	TAP-TIDIM	**0.501***	**0.004**	**0.006**	**0.006**
	GenAP-cond	0.000	0.001	0.001	0.000
	GenAP-DI-cond-opt	0.020*	0.000	0.001	0.001
60×60	TAP-TIDIM	**0.899***	0.004	0.001	0.002
	GenAP-cond	0.001	0.000	0.000	0.002
	GenAP-DI-cond-opt	0.110*	**0.014**	**0.009**	**0.021**
80×80	TAP-TIDIM	**0.993***	0.002	0.002	0.004
	GenAP-cond	0.012	0.005	0.007	0.005
	GenAP-DI-cond-opt	0.011*	**0.037**	**0.025**	**0.055**
100×100	TAP-TIDIM	**1.000***	0.006	0.004	0.002
	GenAP-cond	0.045	0.019	0.012	0.025
	GenAP-DI-cond-opt	0.524*	**0.090**	**0.039**	**0.129**
120×120	TAP-TIDIM	**1.000***	0.012	0.006	0.007
	GenAP-cond	0.096	0.080	0.039	0.090
	GenAP-DI-cond-opt	0.735*	**0.179**	**0.066**	**0.251**

Table 10. The success rates of black-box targeted attack on ImageNet dataset using adversarial patches. The adversarial patches are generated against ResNet101. * indicates white-box attacks.