SemanticAdv: Generating Adversarial Examples via Attribute-conditioned Image Editing Supplementary Materials

Haonan Qiu*[1] Chaowei Xiao*[2] Lei Yang*[3] Xinchen Yan[2,4]
Honglak Lee[2] Bo Li[5]

[1]The Chinese University of Hong Kong, Shenzhen
[2]University of Michigan, Ann Arbor
[3]The Chinese University of Hong Kong
[4] Uber ATG, [5] UIUC

A Implementation details

In this section, we provide implementation details used in our experiments. We implement our *SemanticAdv* using PyTorch [15].

A.1 Face identity verification

We use Adam optimizer [9] to generate adversarial examples for both our *SemanticAdv* and the pixel-wise attack method CW [2]. More specifically, we run optimization for up to 200 steps with a fixed updating rate 0.05 under G-FPR $< 10^{-4}$. Under cases with a slightly higher G-FPR, we run the optimization for up to 500 steps with a fixed updating rate 0.01. For the pixel-wise attack method CW, we use additional pixel reconstruction objective with the weight set to 5. Specifically, we run optimization for up to $1,000$ steps with a fixed updating rate 10^{-3}.

Evaluation metrics. To evaluate the performance of *SemanticAdv* under different attributes, we consider three metrics as follows:

- *Best*: the attack is successful as long as one single attribute among 17 can be successfully attacked;
- *Average*: we calculate the average attack success rate among 17 attributes for the same face identity;
- *Worst*: the attack is successful only when all of 17 attributes can be successfully attacked;

Please note that we use the *Best* metric as a fair comparison to the attack success rate reported by existing pixel-wise attack methods, while *SemanticAdv* can be generated with different attributes as one of our advantages. In practice, both our *SemanticAdv* (*Best*) and CW achieve 100% attack success rate. In addition, we report the performance using the *average* and *worst* metric, which enables us to analyze the adversarial robustness towards certain semantic attributes.

* The first three authors contributed equally.

Pixel-wise defense methods. `Feature squeezing` [19] is a simple but effective method by reducing color bit depth to remove the adversarial effects. We compress the image represented by 8 bits for each channel to 4 bits for each channel to evaluate the effectiveness. For `Blurring` [11], we use a 3×3 Gaussian kernel with standard deviation 1 to smooth the adversarial perturbations. For `JPEG` [4], it leverages the compression and decompression to remove the adversarial perturbation. We set the compression ratio as 0.75 in our experiment.

A.2 Face landmark detection

We use Adam optimizer [9] to generate *SemanticAdv* against the face landmark detection model. Specifically, we run optimization for up to $2,000$ steps with a fixed updating rate 0.05 with the balancing factor λ set to 0.01 (see Eq. 3 in the main paper).

Evaluation Metrics. We apply different metrics for two adversarial attack tasks, respectively. For "Rotating Eyes" task, we use a widely adopted metric *Normalized Mean Error (NME)* [1] for experimental evaluation.

$$r_{\text{NME}} = \frac{1}{N} \sum_{k=1}^{N} \frac{||\mathbf{p}_k - \hat{\mathbf{p}}_k||_2}{\sqrt{W_B * H_B}}, \tag{1}$$

where $\mathbf{p}_k$ denotes the k-th ground-truth landmark, $\hat{\mathbf{p}}_\mathbf{k}$ denotes the k-th predicted landmark and $\sqrt{W_B * H_B}$ is the square-root area of ground-truth bounding box, where W_B and H_B represents the width and height of the box.

For "Out of Region" task, we consider the attack is successful if the landmark predictions fall outside a pre-defined centering region on the portrait image. We introduce a metric that reflects the portion of landmarks outside of the pre-defined centering region: $r_{\text{OR}} = \frac{N_{\text{out}}}{N_{\text{total}}}$, where N_{out} denotes the number of predicted landmarks outside the pre-defined bounding box and N_{total} denotes the total number of landmarks.

A.3 Ablation study: feature-space interpolation

We include an ablation study on feature-space interpolation by analyzing attack success rates using different feature-maps in the main paper. We illustrate the choices of StarGAN feature-maps used in Figure A. Table 1 in the main paper shows the attack success rate on R-101-S. As shown in Figure A, we use $\mathbf{f}_i$ to represent the feature-map after i-th up-sampling operation. $\mathbf{f}_0$ denotes the feature-map before applying up-sampling operation. The result demonstrates that samples generated by interpolating on $\mathbf{f}_0$ achieve the highest success rate. Since $\mathbf{f}_0$ is the feature-map before decoder, it still well embeds semantic information in the feature space. We adopt $\mathbf{f}_0$ for interpolation in our experiments.

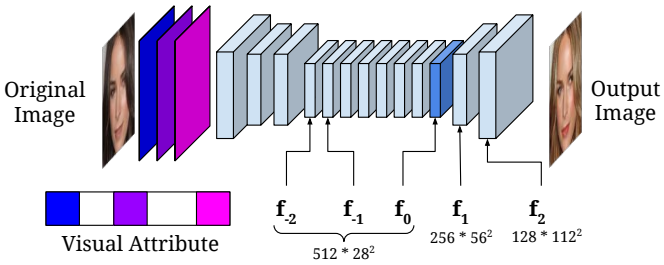

Fig. A. The illustration of the features we used in StarGAN encoder-decoder architecture.

A.4 Semantic attacks on street-view images

Given an input semantic label map at resolution 256×256, we select a target object instance (e.g., a pedestrian) to attack. Then, we create a manipulated semantic label map by inserting another object instance (e.g., a car) in the vicinity of the target object. Similar to the experiments in the face domain, for both semantic label maps, we use the image manipulation encoder to extract features (with $1,024$ channels at spatial resolution 16×16) and conduct feature-space interpolation. We synthesize the final image by feeding the interpolated features to the image manipulation decoder. By searching the interpolation coefficient that maximizes the attack rate, we are able to fool the segmentation model with the synthesized final image.

B Additional quantitative results

B.1 Face identity verification

Benchmark performance. We provide additional information about the ResNet models used in the experiments. Table A illustrates the performance on multiple face identity verification benchmarks including Labeled Face in the Wild (LFW) dataset [7], AgeDB-30 dataset [14], and Celebrities in Frontal-Profile (CFP) dataset [16]. LFW [7] is the *de facto* standard testing set for face verification under unconstrained conditions, which contains $13,233$ face images from $5,749$ identities. AgeDB [14] contains $12,240$ images from 440 identities. AgeDB-30 is the most challenging subsets for evaluating face verification models. The large variations in age makes the face model perform worse on this dataset than on LFW. CFP [16] consists of 500 identities, where each identity has 10 frontal and 4 profile images. Although good performance has been achieved on the Frontal-to-Frontal (CFP-FF) test protocol, the Frontal-to-Profile (CFP-FP) test protocol still remains challenging as most of the face training sets have very few profile face images. Table A indicates that the used face verification model achieves state-of-the-art under all benchmarks.

Table A. The verification accuracy (%) of ResNet models on multiple face recognition datasets including LFW, AgeDB-30, and CFP.

$\mathcal{M}$ / benchmarks	LFW	AgeDB-30	CFP-FF	CFP-FP
R-50-S	99.27	94.15	99.26	91.49
R-101-S	99.42	95.93	99.57	95.07
R-50-C	99.38	95.08	99.24	90.24
R-101-C	99.67	95.58	99.57	92.71

Thresholds for identity verification. To decide whether two portrait images belong to the same identity or not, we use the normalized L_2 distance between face features and set the FPR thresholds accordingly, which is a commonly used procedure when evaluating the face verification model [10, 8]. Table B illustrates the threshold values used in our experiments when determining whether two portrait images belong to the same identity or not.

Table B. The threshold values for face identity verification.

FPR/$\mathcal{M}$	R-50-S	R-101-S	R-50-C	R-101-C
10^{-3}	1.181	1.244	1.447	1.469
3×10^{-4}	1.058	1.048	1.293	1.242
10^{-4}	0.657	0.597	0.864	0.809

Quantitative analysis. Combining the results from Table C and Figure 4 in the main paper, we understand that the face verification models used in our experiments have different levels of robustness across attributes. For example, face verification models are more robust against local shape variations than color variations, e.g., pale skin has higher attack success rate than mouth open. We believe these discoveries will help the community further understand the properties of face verification models.

Table C shows the overall performance (accuracy) of face verification model and attack success rate of *SemanticAdv* and CW. As shown in Table C, although the face model trained with `cos` objective achieves higher face recognition performance, it is more vulnerable to adversarial attack compared with the model trained with `softmax` objective. Table D shows that the intermediate results of *SemanticAdv* before adversarial perturbation cannot attack successfully, which indicates the success of *SemanticAdv* comes from adding adversarial perturbations through interpolation.

Table C. Quantitative results of identity verification (%). It shows accuracy of face verification model and attack success rate of *SemanticAdv* and CW.

G-FPR	Metrics / $\mathcal{M}$	R-50-S	R-101-S	R-50-C	R-101-C
10^{-3}	Verification Accuracy	98.36	98.78	98.63	98.84
	SemanticAdv (*Best*)	100.00	100.00	100.00	100.00
	SemanticAdv (*Worst*)	91.95	93.98	99.53	99.77
	SemanticAdv (*Average*)	98.98	99.29	99.97	99.99
	CW	100.00	100.00	100.00	100.00
3×10^{-4}	Verification Accuracy	97.73	97.97	97.91	97.85
	SemanticAdv (*Best*)	100.00	100.00	100.00	100.00
	SemanticAdv (*Worst*)	83.75	79.06	98.98	96.64
	SemanticAdv (*Average*)	97.72	97.35	99.92	99.72
	CW	100.00	100.00	100.00	100.00
10^{-4}	Verification Accuracy	93.25	92.80	93.43	92.98
	SemanticAdv (*Best*)	100.00	100.00	100.00	100.00
	SemanticAdv (*Worst*)	33.59	19.84	67.03	48.67
	SemanticAdv (*Average*)	83.53	76.64	95.57	91.13
	CW	100.00	100.00	100.00	100.00

Table D. Attack success rate of the intermediate output of *SemanticAdv* (%). $\mathbf{x}'$, $G(\mathbf{x}', \mathbf{c})$ and $G(\mathbf{x}', \mathbf{c}^{\text{new}})$ are the intermediate results of our method before adversarial perturbation.

G-FPR	Metrics / $\mathcal{M}$	R-50-S	R-101-S	R-50-C	R-101-C
10^{-3}	$\mathbf{x}'$	0.00	0.00	0.08	0.00
	$G(\mathbf{x}', \mathbf{c})$	0.00	0.00	0.00	0.23
	$G(\mathbf{x}', \mathbf{c}^{\text{new}})$(*Best*)	0.16	0.08	0.16	0.31
3×10^{-4}	$\mathbf{x}'$	0.00	0.00	0.00	0.00
	$G(\mathbf{x}', \mathbf{c})$	0.00	0.00	0.00	0.00
	$G(\mathbf{x}', \mathbf{c}^{\text{new}})$(*Best*)	0.00	0.00	0.00	0.00
10^{-4}	$\mathbf{x}'$	0.00	0.00	0.00	0.00
	$G(\mathbf{x}', \mathbf{c})$	0.00	0.00	0.00	0.00
	$G(\mathbf{x}', \mathbf{c}^{\text{new}})$(*Best*)	0.00	0.00	0.00	0.00

B.2 Face landmark detection

We present the quantitative results of *SemanticAdv* on face landmark detection model in Table E including two adversarial tasks, namely, "Rotating Eyes" and "Out of Region". We observe that our method is efficient to perform attacking on landmark detection models. For certain attributes such as "Eyeglasses" and "Pale Skin", *SemanticAdv* achieves reasonably-good performance.

Table E. Quantitative results on face landmark detection (%) The two row shows the measured ratios (lower is better) for "Rotating Eyes" and "Out of Region" task, respectively.

Tasks (Metrics)	Pristine	Augmented Attributes							
		Blond Hair	Young	Eyeglasses	Rosy Cheeks	Smiling	Arched Eyebrows	Bangs	Pale Skin
$r_{\text{NME}} \downarrow$	28.04	14.03	17.28	8.58	13.24	19.21	23.42	15.99	10.72
$r_{\text{OR}} \downarrow$	45.98	17.42	23.04	7.51	16.65	25.44	33.85	20.03	13.51

B.3 User study

We conduct a user study on the adversarial images of *SemanticAdv* and CW used in the experiment of API-attack and the original images. The adversarial images are generated with G-FPR$< 10^{-4}$ for both methods. We present a pair of original image and adversarial image to participants and ask them to rank the two options. The order of these two images is randomized and the images are displayed for 2 seconds in the screen during each trial. After the images disappear, the participants have unlimited time to select the more reasonably-looking image according to their perception. To maintain the high quality of the collected responses, each participant can only conduct at most 50 trials, while and each adversarial image was shown to 5 different participants. We present the images we used for user study in Figure B. In total, we collect $2,620$ annotations from 77 participants. In $39.14 \pm 1.96\%$ of trials the adversarial images generated by *SemanticAdv* are selected as reasonably-looking images and in $30.27 \pm 1.96\%$ of trails, the adversarial images generated by CW are selected as reasonably-looking images. It indicates that our semantic adversarial examples are more perceptual reasonably-looking than CW. Additionally, we also conduct the user study with larger G-FPR$= 10^{-3}$. In $45.42 \pm 1.96\%$ of trials, the adversarial images generated by *SemanticAdv* are selected as reasonably-looking images, which is very close to the random guess (50%).

B.4 Semantic attack transferability

In Table F, we present the quantitative results of the attack transferability under the setting with G-FPR $= 10^{-4}$ and T-FPR $= 10^{-4}$. We observe that with more strict testing criterion (lower T-FPR) of the verification model, the transferability becomes lower across different models.

To further showcase that our *SemanticAdv* is non-trivially different from pixel-wise attack added on top of semantic image editing, we provide one additional baseline called StarGAN+CW and evaluate its attack transferability. This baseline first performs semantic image editing using the StarGAN model (non-adversarial) and then conducts the standard L_p CW attacks on the generated images. As shown in Table G, the StarGAN+CW baseline has noticeable performance gap to our proposed *SemanticAdv*. This also justifies that our *SemanticAdv* is able to

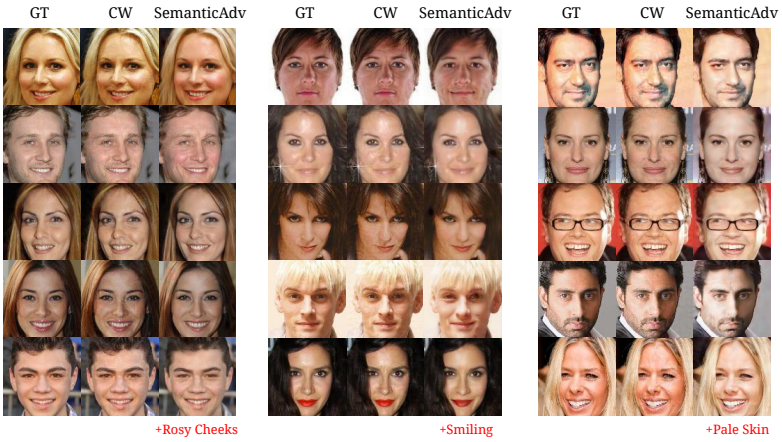

Fig. B. Qualitative comparisons among ground truth, pixel-wise adversarial examples generated by CW, and our proposed *SemanticAdv*. Here, we present the results from G-FPR $< 10^{-4}$ so that perturbations are visible.

Table F. Transferability of *SemanticAdv*: cell (i, j) shows attack success rate of adversarial examples generated against j-th model and evaluate on i-th model. Results are generated with G-FPR $= 10^{-4}$ and T-FPR $= 10^{-4}$.

$\mathcal{M}_{\text{test}}$ / $\mathcal{M}_{\text{opt}}$	R-50-S	R-101-S	R-50-C	R-101-C
R-50-S	1.000	0.005	0.000	0.000
R-101-S	0.000	1.000	0.000	0.000
R-50-C	0.000	0.000	1.000	0.000
R-101-C	0.000	0.000	0.000	1.000

produce novel adversarial examples which cannot be simply achieved by combining attribute-conditioned image editing model with L_p bounded perturbation.

B.5 Query-free black-box API attack

In Table J, we present the results of *SemanticAdv* performing query-free black-box attack on three online face verification platforms. *SemanticAdv* outperforms CW and StarGAN+CW in all APIs under all FPR thresholds. In addition, under the same T-FPR, we achieve higher attack success rate on APIs using samples generated using lower G-FPR compared to samples generated using higher G-FPR. Original $\mathbf{x}$ and generated $\mathbf{x}^{\text{new}}$ are regarded as reference point of the performance of online face verification platforms. In Figure C, we also show several examples of our API attack on Microsoft Azure face verification system, which further demonstrates the effectiveness of our approach.

Table G. Transferability of *StarGAN+CW*: cell (i, j) shows attack success rate of adversarial examples generated against j-th model and evaluate on i-th model. Results of *SemanticAdv* are listed in brackets.

$\mathcal{M}_{\text{test}}$ / $\mathcal{M}_{\text{opt}}$	R-101-S
R-50-S	0.035 (0.108)
R-101-S	1.000 (1.000)
R-50-C	0.145 (0.202)
R-101-C	0.085 (0.236)

Table H. G-FPR=10^{-3}, T-FPR=10^{-3}

$\mathcal{M}_{\text{test}}$ / $\mathcal{M}_{\text{opt}}$	R-101-S
R-50-S	0.615 (0.862)
R-101-S	1.000 (1.000)
R-50-C	0.570 (0.837)
R-101-C	0.695 (0.888)

Table I. G-FPR=10^{-4}, T-FPR=10^{-3}

B.6 *SemanticAdv* against adversarial training

We evaluate our *SemanticAdv* against the existing adversarial training based defense method [13]. In detail, we randomly sample 10 persons from CelebA [12] and then randomly split the sampled dataset into training set, validation set and testing set according to a proportion of 80%, 10% and 10%, respectively. We train a ResNet-50 [6] to identify these face images by following the standard face recognition training pipeline [17]. As CelebA [12] does not contain enough images for each person, we finetune our model from a pretrained model trained on MS-Celeb-1M [5, 20]. We train the robust model by using adversarial training based method [13]. In detail, we follow the same setting in [13]. We use 7-step PGD L_∞ attack to generate adversarial examples to solve the inner maximum problem for adversarial training. During test process, we evaluate by using adversarial examples generated by 20-step PGD attacks. The perturbation is bounded by 8 pixel (ranging from [0,255]) in terms of L_∞ distance).

As shown in Table K, the robust model achieves 10% accuracy against the adversarial examples generated by *SemanticAdv*, while 46.7% against the adversarial examples generated by PGD [13]. It indicates that existing adversarial training based defense method is less effective against *SemanticAdv*. It further demonstrates that our *SemanticAdv* identifies an unexplored research area beyond previous L_p-based ones.

B.7 Semantic attacks on street-view images

We conduct our experiment on CityScape dataset and use target attack success rate as our evaluation metric. The target attack success rate is measured by the pixel-wise accuracy between the predicted result and the target segmentation map. Our proposed method achieves $83.8 \pm 11.2\%$ attack success rate.

Table J. Quantitative analysis on query-free black-box attack. We use ResNet-101 optimized with `softmax` loss for evaluation and report the attack success rate(%). Note that for Micresoft Azure API, it does not provide the accept thresholds for different T-FPRs and thus we use the provided likelihood 0.5 to determine whether two faces belong to the same person.

API name	Face++		AliYun		Azure
Metric	T-FPR		T-FPR		Likelihood
Attacker / Metric value	10^{-3}	10^{-4}	10^{-3}	10^{-4}	0.5
Original $\mathbf{x}$	2.04	0.51	0.50	0.00	0.00
Generated $\mathbf{x}^{\text{new}}$	4.21	0.53	0.50	0.00	0.00
CW (G-FPR = 10^{-3})	9.18	2.04	2.00	0.50	0.00
StarGAN+CW (G-FPR = 10^{-3})	15.9	3.08	3.50	**1.00**	0.00
SemanticAdv (G-FPR = 10^{-3})	**20.00**	**4.10**	**4.00**	0.50	0.00
CW (G-FPR = 10^{-4})	28.57	10.17	10.50	2.50	1.04
StarGAN+CW (G-FPR = 10^{-4})	35.38	14.36	12.50	3.50	1.05
SemanticAdv (G-FPR = 10^{-4})	**58.25**	**31.44**	**24.00**	**10.50**	**5.73**
CW	37.24	20.41	18.00	9.50	3.09
StarGAN+CW	47.45	26.02	20.00	8.50	5.56
MI-FGSM [3]	53.89	30.57	29.50	17.50	10.82
M-DI^2-FGSM [18]	56.12	33.67	30.00	18.00	12.04
SemanticAdv (G-FPR < 10^{-4})	**67.69**	**48.21**	**36.5**	**19.5**	**15.63**

Table K. Accuracy on standard model (without adversarial training) and robust model (with adversarial training).

Training Method / Attack	Benign	PGD	*SemanticAdv*
Standard	93.3%	0%	0%
Robust [13]	86.7%	46.7%	10%

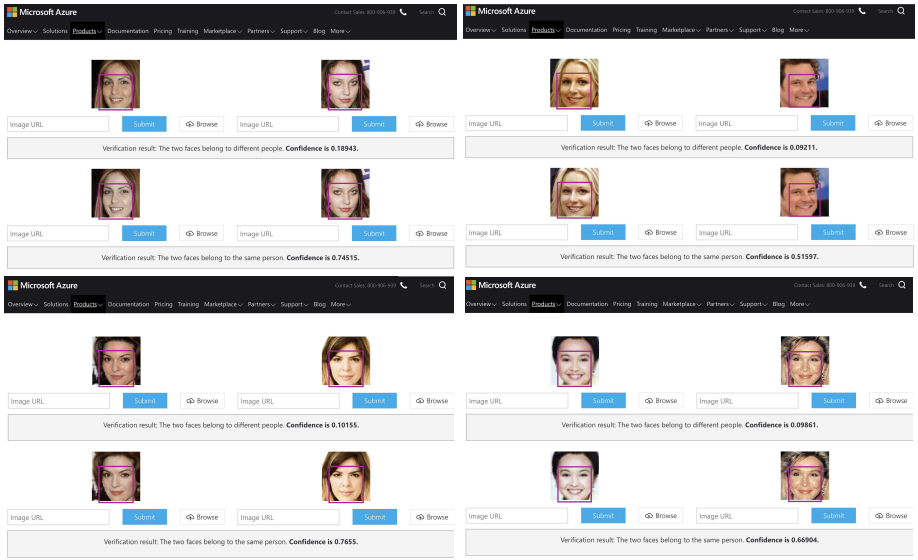

Fig. C. Illustration of our *SemanticAdv* in the real world face verification platform (editing on pale skin). Note that the confidence denotes the likelihood that two faces belong to the same person.

C Additional visualizations

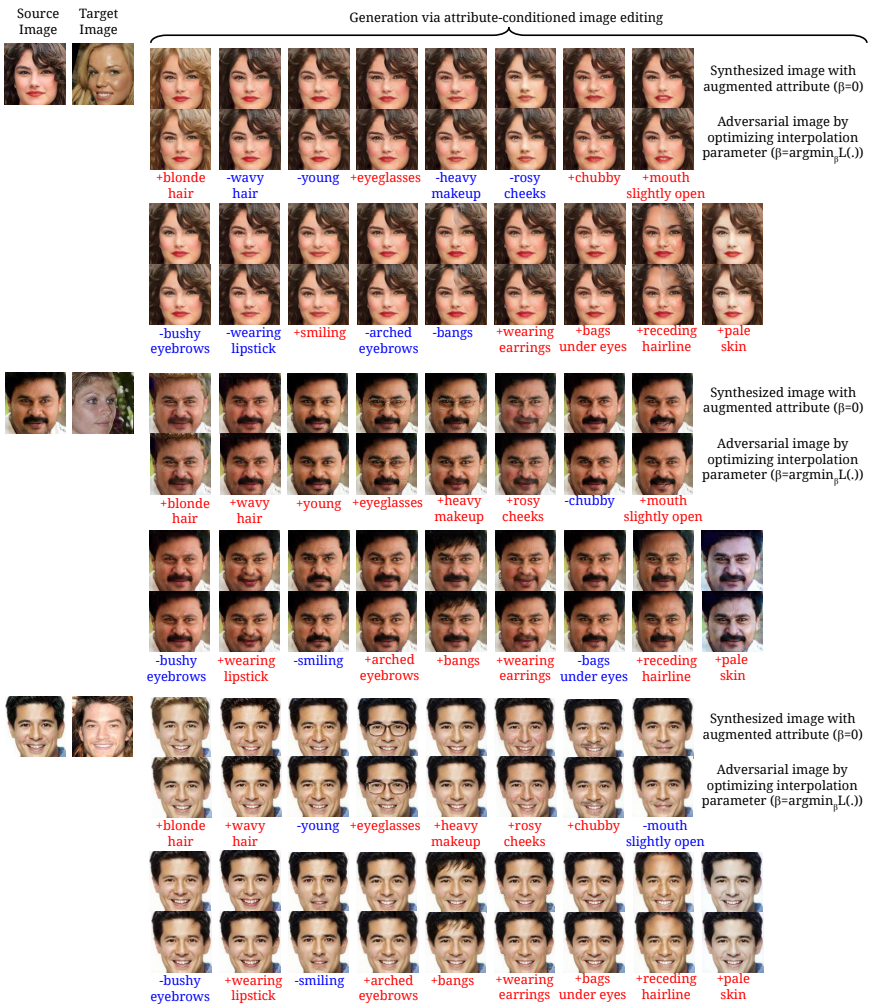

Fig. D. Qualitative analysis on single-attribute adversarial attack (G-FPR=10^{-3}).

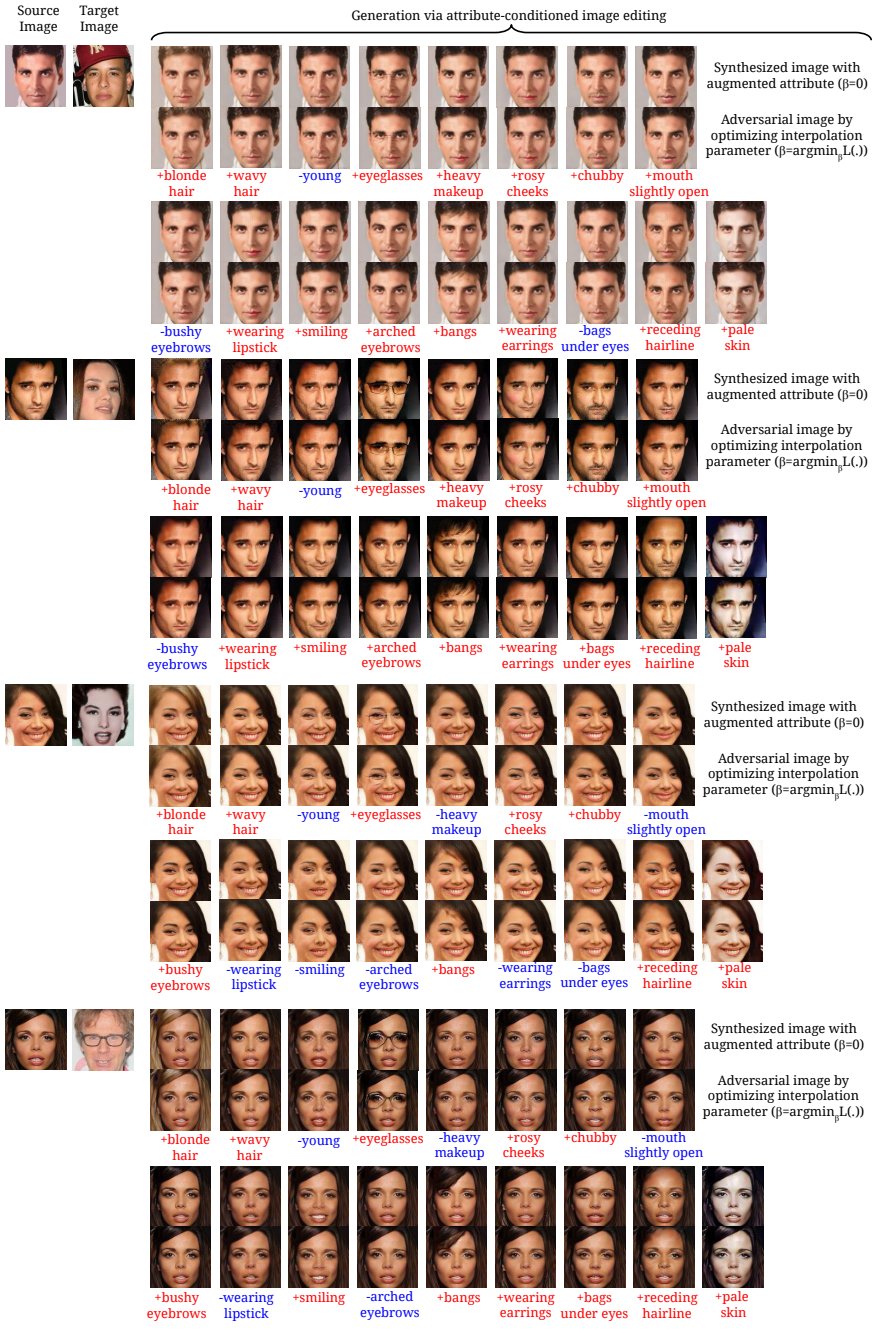

Fig. E. Qualitative analysis on single-attribute adversarial attack (G-FPR=10^{-3}).

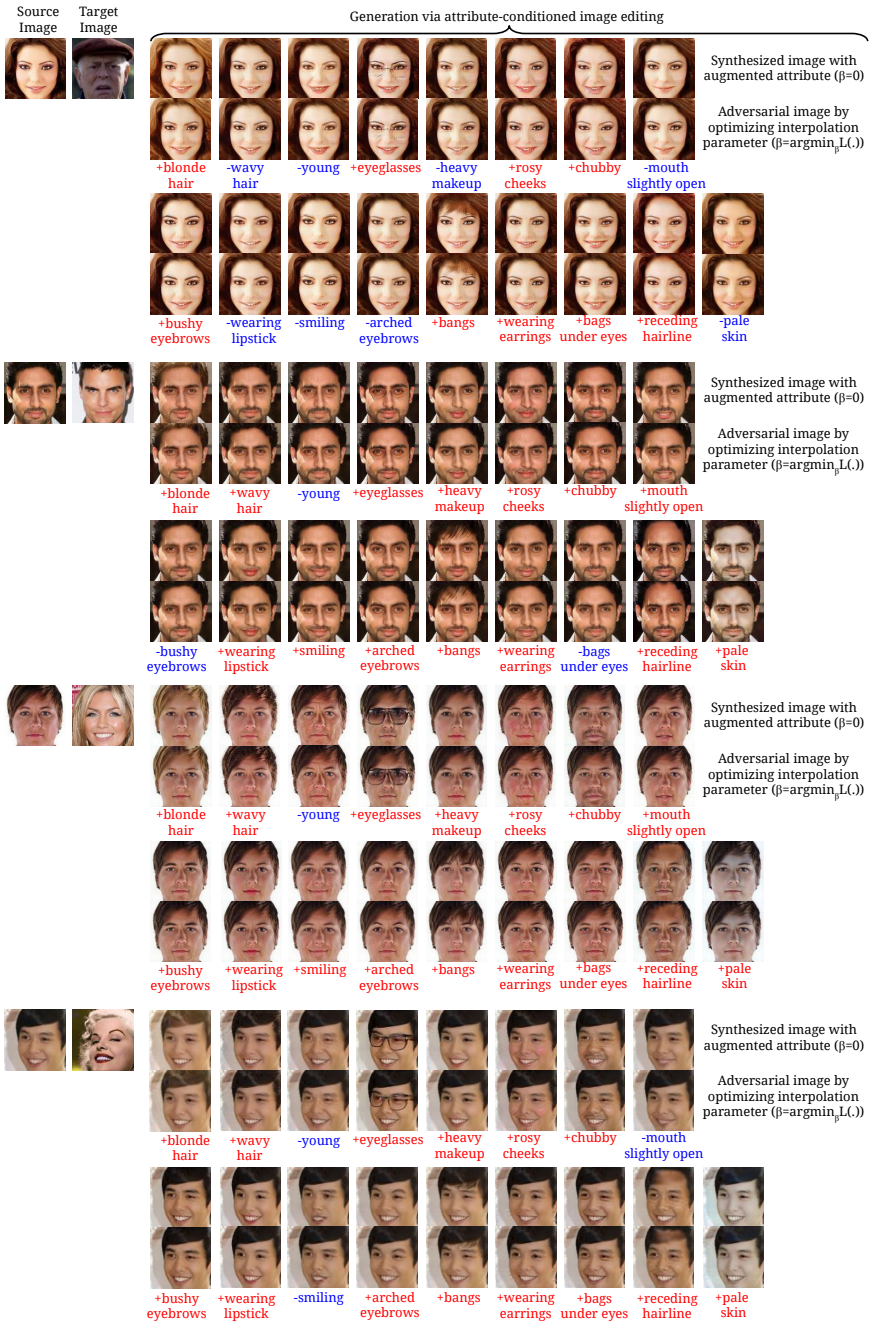

Fig. F. Qualitative analysis on single-attribute adversarial attack (G-FPR=10^{-3}).

Fig. G. Qualitative comparisons between our proposed *SemanticAdv* (G-FPR = 10^{-3}) and pixel-wise adversarial examples generated by CW. Along with the adversarial examples, we also provide the corresponding perturbations (residual) on the right.

Fig. H. Qualitative analysis on single-attribute adversarial attack (*SemanticAdv* with G-FPR = 10^{-3}) by each other. Along with the adversarial examples, we also provide the corresponding perturbations (residual) on the right.

References

1. Bulat, A., Tzimiropoulos, G.: How far are we from solving the 2d & 3d face alignment problem?(and a dataset of 230,000 3d facial landmarks). In: ICCV (2017)
2. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 IEEE Symposium on Security and Privacy (S&P). IEEE (2017)
3. 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)
4. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
5. Guo, Y., Zhang, L., Hu, Y., He, X., Gao, J.: Ms-celeb-1m: A dataset and benchmark for large-scale face recognition. In: ECCV. Springer (2016)
6. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016)
7. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
8. Kemelmacher-Shlizerman, I., Seitz, S.M., Miller, D., Brossard, E.: The megaface benchmark: 1 million faces for recognition at scale. In: CVPR. pp. 4873–4882 (2016)
9. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: ICLR (2015)
10. Klare, B.F., Klein, B., Taborsky, E., Blanton, A., Cheney, J., Allen, K., Grother, P., Mah, A., Jain, A.K.: Pushing the frontiers of unconstrained face detection and recognition: Iarpa janus benchmark a. In: CVPR (2015)
11. Li, X., Li, F.: Adversarial examples detection in deep networks with convolutional filter statistics. In: Proceedings of the IEEE International Conference on Computer Vision. pp. 5764–5772 (2017)
12. Liu, Z., Luo, P., Wang, X., Tang, X.: Deep learning face attributes in the wild. In: ICCV (2015)
13. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. In: ICLR (2018)
14. Moschoglou, S., Papaioannou, A., Sagonas, C., Deng, J., Kotsia, I., Zafeiriou, S.: Agedb: the first manually collected, in-the-wild age database. In: CVPR Workshops. pp. 51–59 (2017)
15. Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A.: Automatic differentiation in PyTorch. In: NIPS Autodiff Workshop (2017)
16. Sengupta, S., Chen, J.C., Castillo, C., Patel, V.M., Chellappa, R., Jacobs, D.W.: Frontal to profile face verification in the wild. In: WACV. pp. 1–9. IEEE (2016)
17. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: CVPR (2014)
18. Xie, C., Zhang, Z., Zhou, Y., Bai, S., Wang, J., Ren, Z., Yuille, A.L.: Improving transferability of adversarial examples with input diversity. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2730–2739 (2019)
19. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
20. Zhang, X., Yang, L., Yan, J., Lin, D.: Accelerated training for massive classification via dynamic class selection. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)