Supplementary Material: Robustness and Generalization via Generative Adversarial Training

1. Comparison with Song et al. [2]

We show that adversarial training with examples generated by [33] hurts the classifier's performance on clean images. Table 1 demonstrates the results. We use the same classifier architectures as [2] and consider their basic attack. We observe that the test accuracy on clean images drops by 1.3%, 1.4% and 1.1% on MNIST, SVHN and CelebA respectively. As we show in Table 1 of the main paper training with our examples improves the accuracy, demonstrating difference of our approach with that of [2].

To further illustrate and compare distributions of real and adversarial images, we use a pre-trained VGG network to extract features of each image from CelebA-HQ, our adversarial examples, and those of [2], and then plot them with t-SNE embeddings as shown in Figure 1. We can see that the embeddings of CelebA-HQ real and our adversarial images are blended while the those of CelebA-HQ and Song et al.'s adversarial examples are more segregated. This again provides evidence that our adversarial images stay closer to the original manifold and hence could be more useful as adversarial training data.

2. Number of Iterations

To make sure the iterative process always converges in a reasonable number of steps, we measure the number of updates required to fool the classifier on 1000 randomly-selected images. Results are shown in Table 2. Note that for targeted attacks we first randomly sample a target class different from the ground-truth label for each image.

3. Evaluation on Certified Defenses

Adversaries can circumvent defenses tailored for a specific type of attack by using new threat models. To demonstrate this, we evaluate our attack on a certified defense against norm-bounded perturbations. Cohen et al. [1] propose a certified defense using randomized smoothing with Gaussian noise, which guarantees a certain top-1 accuracy for perturbations with L_2 norm less than a specific threshold. We use 400 noise-based and 400 style-based adversarial images from the object categories of LSUN. Our adversarial examples are evaluated against a randomized smoothing classifier based on ResNet-50 using Gaussian noise with standard deviation of 0.5. Table 3 shows accuracy of the model on clean and adversarial images. As we observe, the accuracy drops on adversarial inputs, and the certified defense is not effective against our attack. Note that we stop updating adversarial images as soon as the model is fooled. If we keep updating for more iterations afterwards, we can achieve even stronger attacks.

4. Object Detection Results

Figure 2 illustrates results on the object detection task using the RetinaNet target model [28]. We observe that small changes in the images lead to incorrect bounding boxes and predictions by the model.

5. Impact of γ and β on Semantic Segmentation

In segmentation results shown in Figure 5 we simultaneously modify both γ and β parameters of the SPADE module. We can also consider the impact of modifying each parameter separately. Figure 3 illustrates the results. As we observe, changing γ and β modifies fine details of the images which are barely perceptible yet they lead to large changes in predictions of the segmentation model.

6. Adversarial Changes to Single Images

Figure 4 illustrates how images vary as we manipulate specific layers of the network. We observe that each set of layers creates different adversarial changes. For instance, layers 12 to 18 mainly change low-level color details.

7. Adversarial Training with Norm-bounded Perturbations

We consider adversarial training with norm-bounded perturbations and limit the number of iterations to make the setup comparable with our unrestricted adversarial training. Specifically, we use Iterative-FGSM with $\epsilon = 4$ and a bounded number of steps. Results are shown in Table 4. Note that accuracy of the models drop on clean images although we use a weak attack. This is in contrast to training

	MNIST		SVHN		CelebA	
	Clean	Adversarial	Clean	Adversarial	Clean	Adversarial
Adv. Trained	98.2%	84.5%	96.4%	86.4%	96.9%	85.9%
Original	99.5%	12.8%	97.8%	14.9%	98.0%	16.2%

Table 1: Accuracy of adversarially trained and original models on clean and adversarial test images from [2].

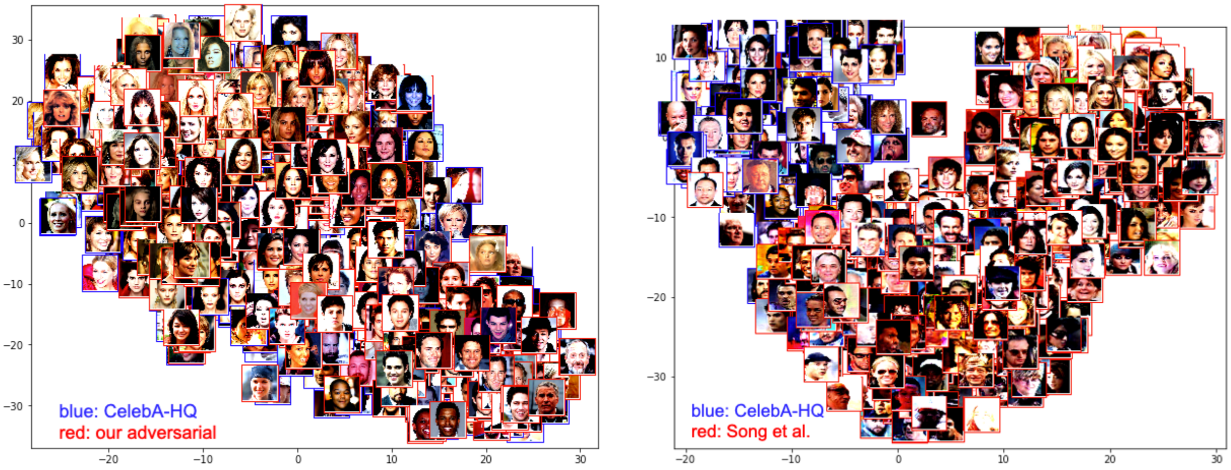


Figure 1: t-SNE plot comparing distributions of real images with adversarial examples from our approach and Song et al.

	LSUN		CelebA-HQ
	Targeted	Non-targeted	
Style-based	9.1 ± 4.2	6.8 ± 3.6	7.3 ± 3.0
Noise-based	4.5 ± 1.7	3.7 ± 1.8	6.2 ± 4.1

Table 2: Average number of iterations (mean ± std) required to fool the classifier.

	Accuracy
Clean	63.1%
Adversarial (style)	21.7%
Adversarial (noise)	37.8%

Table 3: Accuracy of a certified classifier equipped with randomized smoothing on our adversarial images.

with our unrestricted adversarial examples that improves the accuracy.

8. Additional Examples

We also provide additional examples and higher-resolution images in the following. Figure 5 depicts additional examples on the segmentation task. Figure 6 illustrates adversarial examples on CelebA-HQ gender classification, and Figure 7 shows additional examples on the LSUN dataset. Higher-resolution versions for some of the adversarial images are shown in Figure 8, which particularly helps to distinguish subtle differences between original and noise-based images.

		IFGSM-2	IFGSM-5
	Original	Adv. Trained	Adv. Trained
LSUN	88.9%	88.4%	87.8%
CelebA-HQ	95.7%	95.1%	94.6%

Table 4: Adversarial Training with norm-bounded perturbations. Iterative-FGSM ($\epsilon = 4$) with a maximum of 2 and 5 iterations is considered, and accuracy of the adversarially trained and original models on clean test images are shown.

References

[1] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019. 1

[2] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018. 1, 2

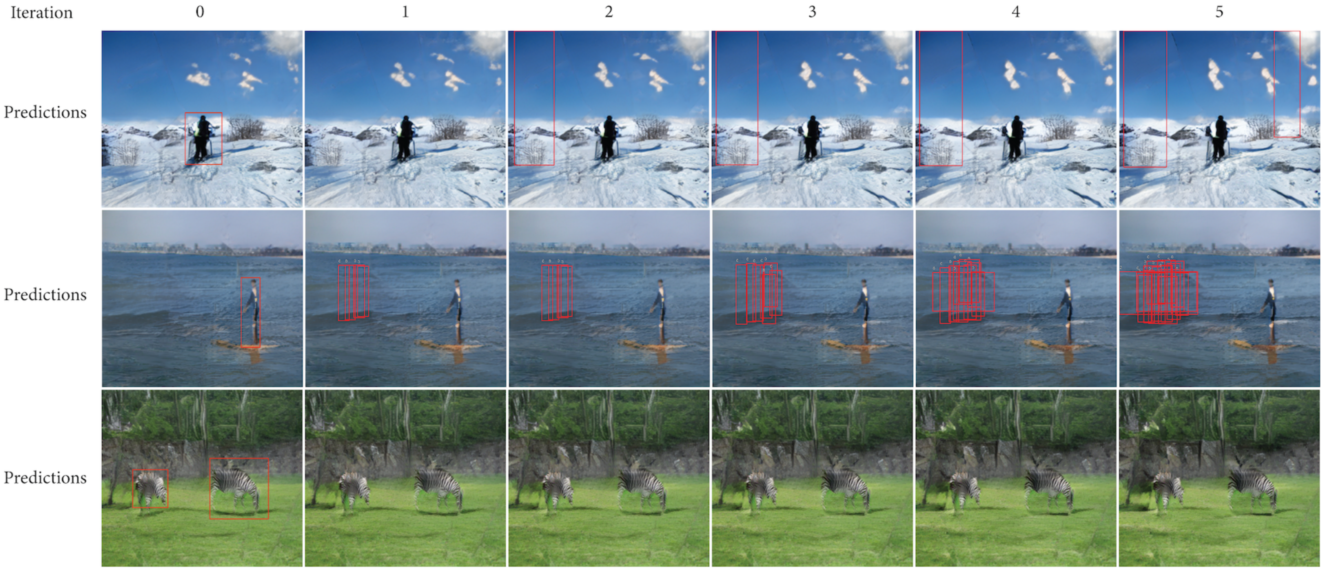


Figure 2: Unrestricted adversarial examples for object detection. Generated images and their corresponding predictions are shown for different number of iterations.

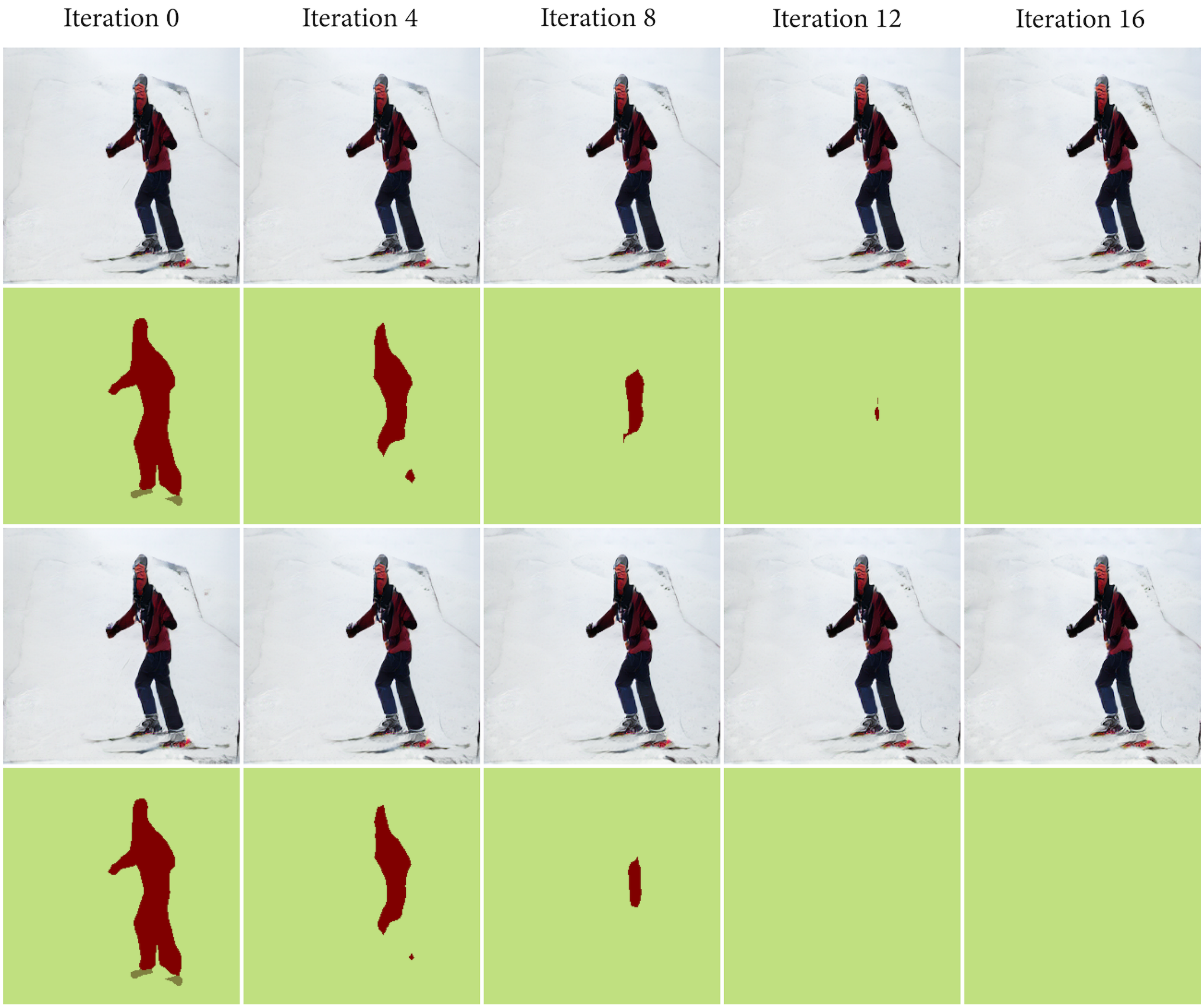


Figure 3: Impact of separately modifying γ and β parameters on segmentation results. Modified images at different iterations and corresponding predictions are shown. In the first two rows only the γ values are changed and in the last two rows only the β values are modified.

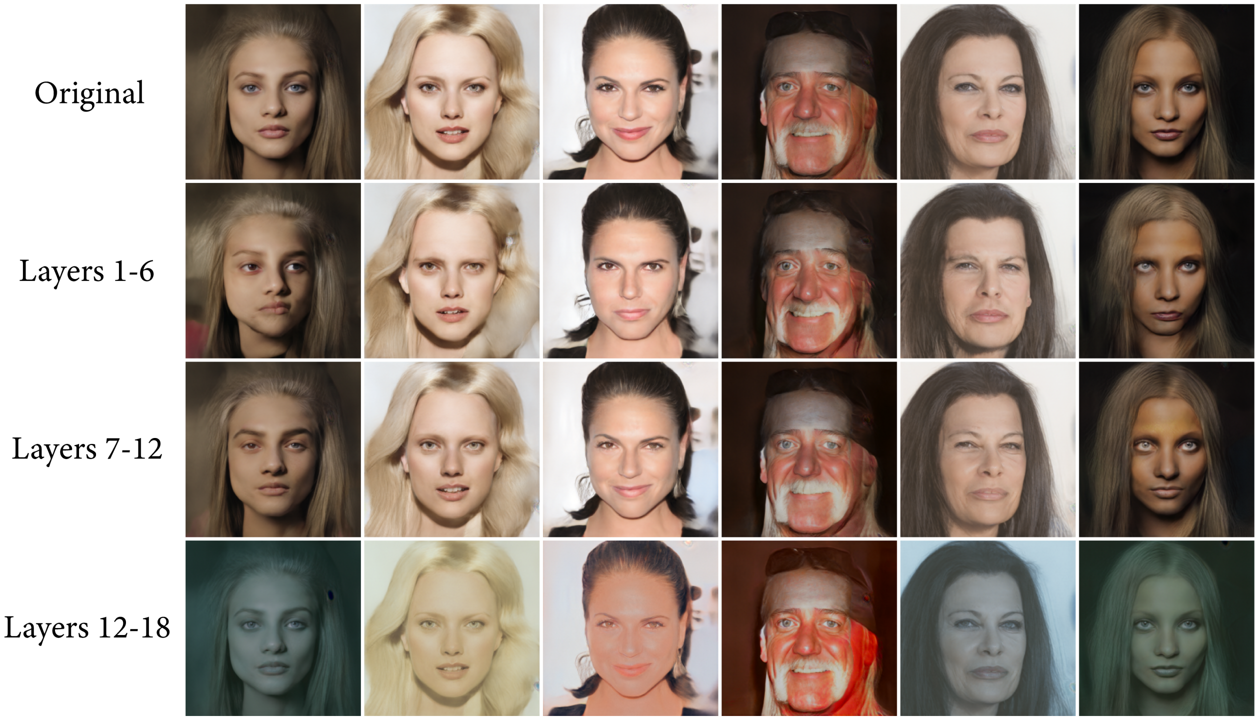


Figure 4: Impact of manipulating different layers of the network on generated adversarial images.

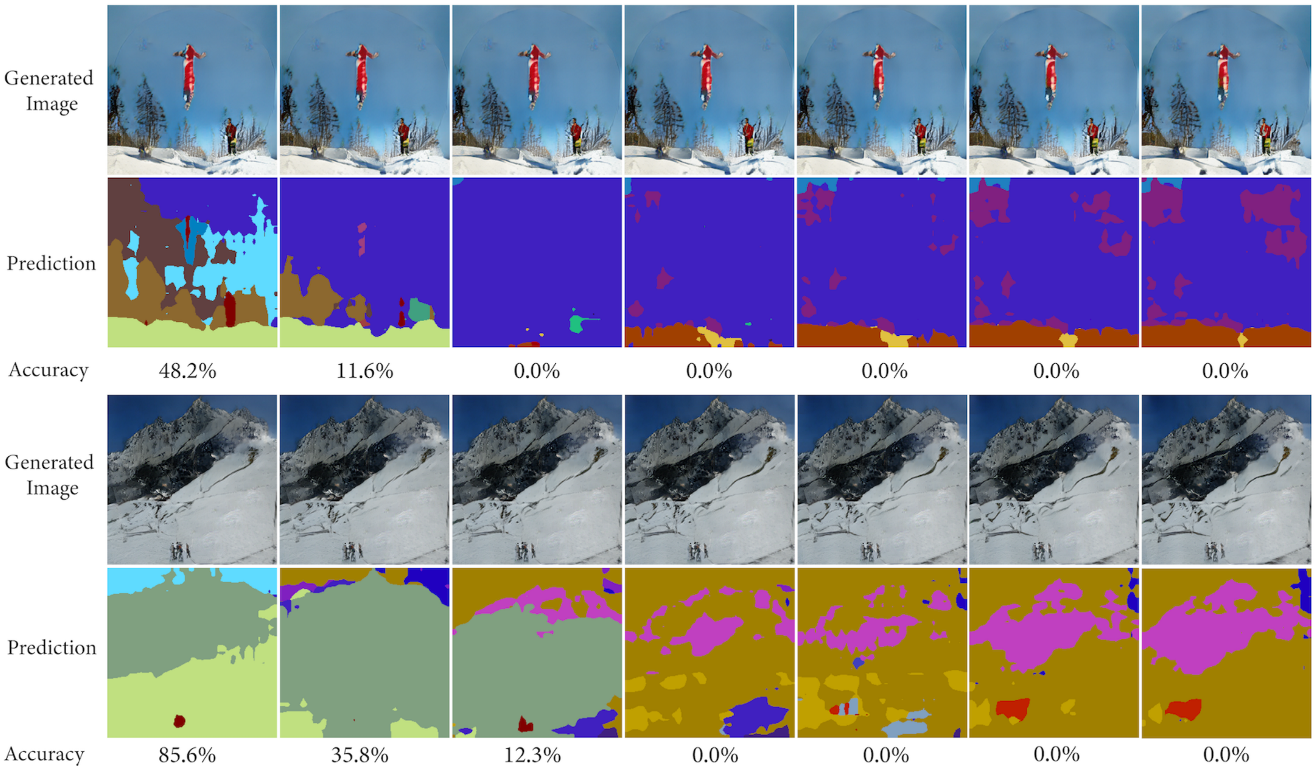


Figure 5: Unrestricted adversarial examples for semantic segmentation. Generated images, corresponding predictions and their accuracy (ratio of correctly predicted pixels) are shown for different number of iterations.

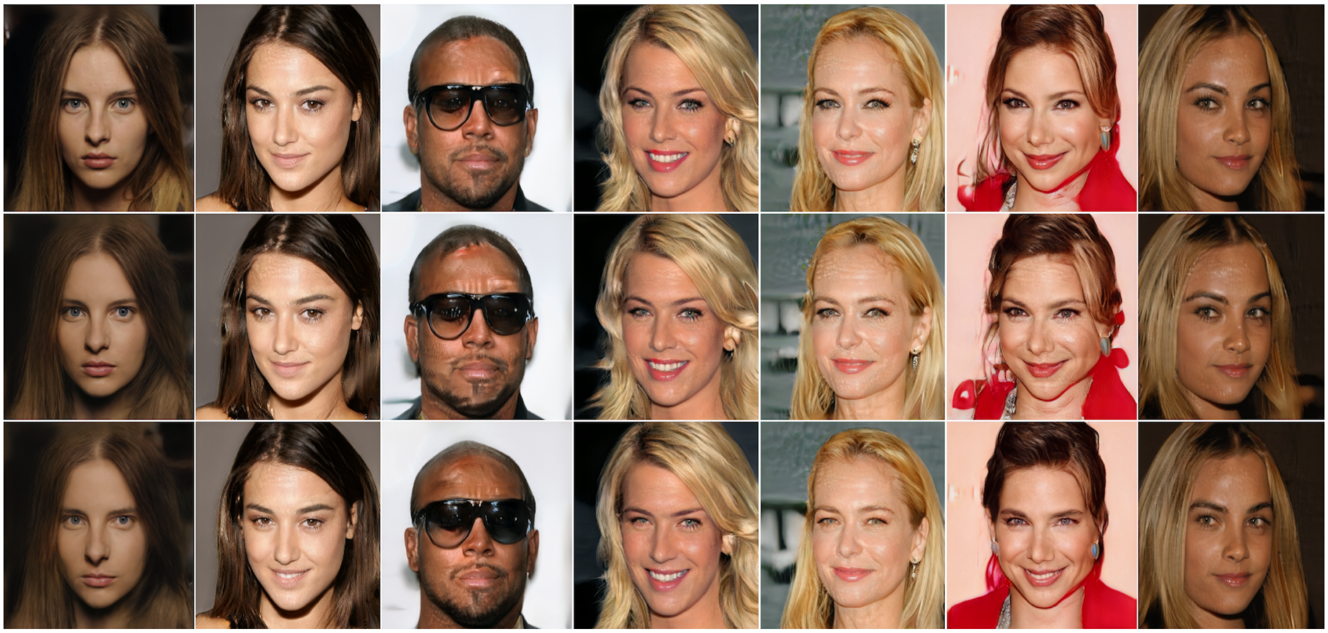
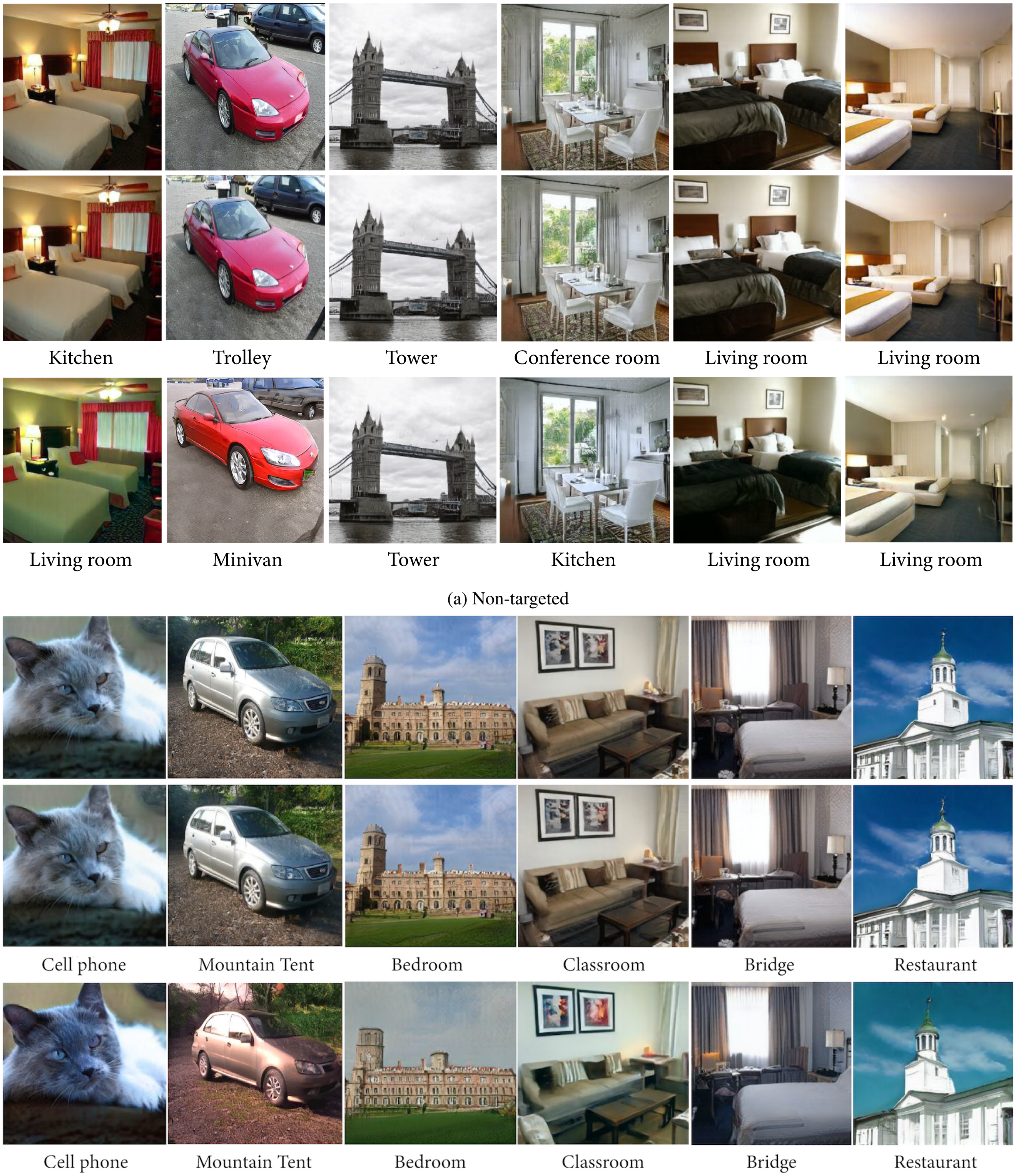


Figure 6: Unrestricted adversarial examples on CelebA-HQ gender classification. From top to bottom: Original, noise-based and style-based adversarial images. Males are classified as females and vice versa.



(a) Non-targeted

(b) Targeted

Figure 7: Unrestricted adversarial examples on LSUN for a) non-targeted and b) targeted attacks. From top to bottom: original, noise-based and style-based images.

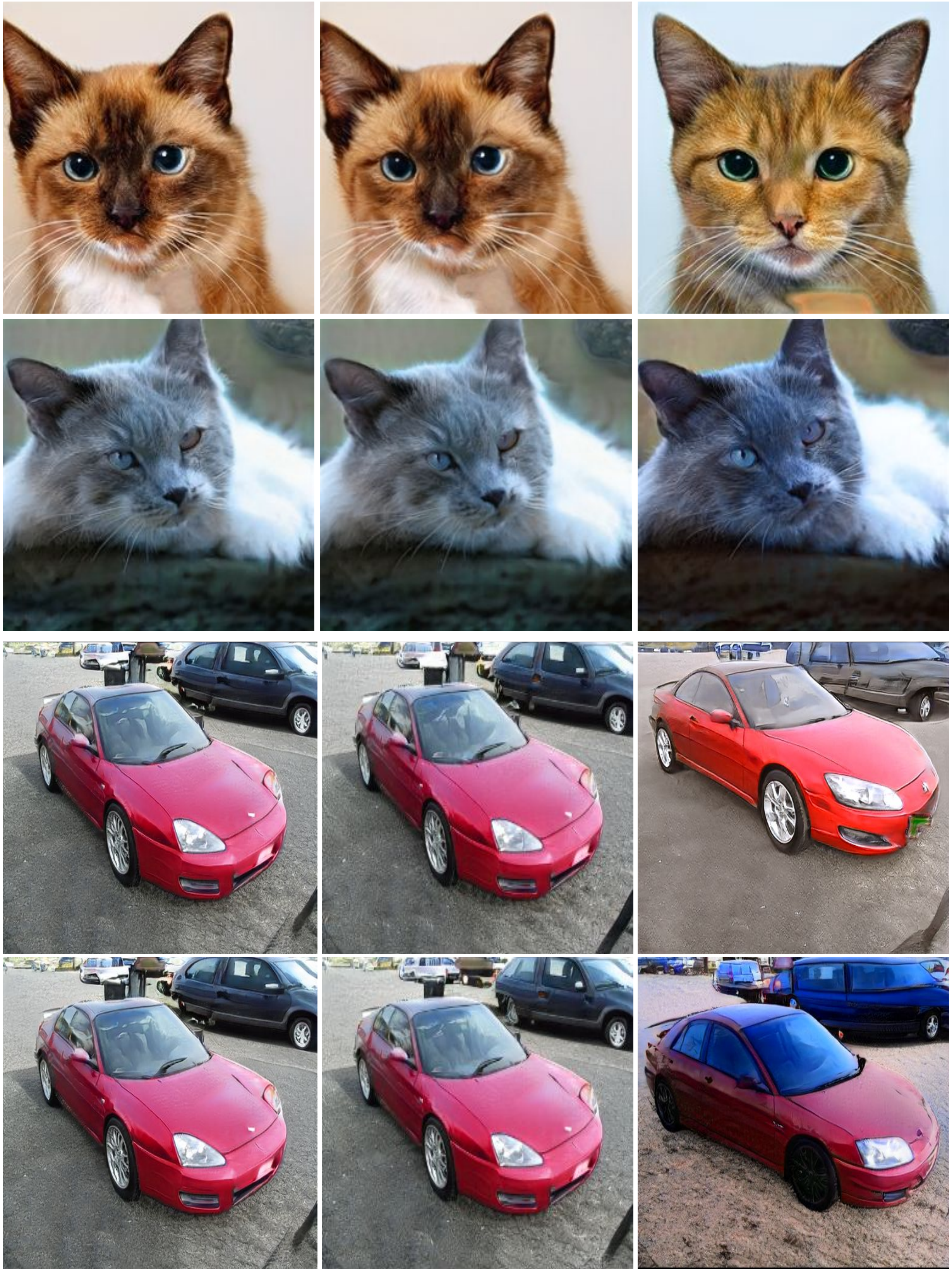


Figure 8: High resolution versions of adversarial images. From left to right: original, noise-based and style-based images.

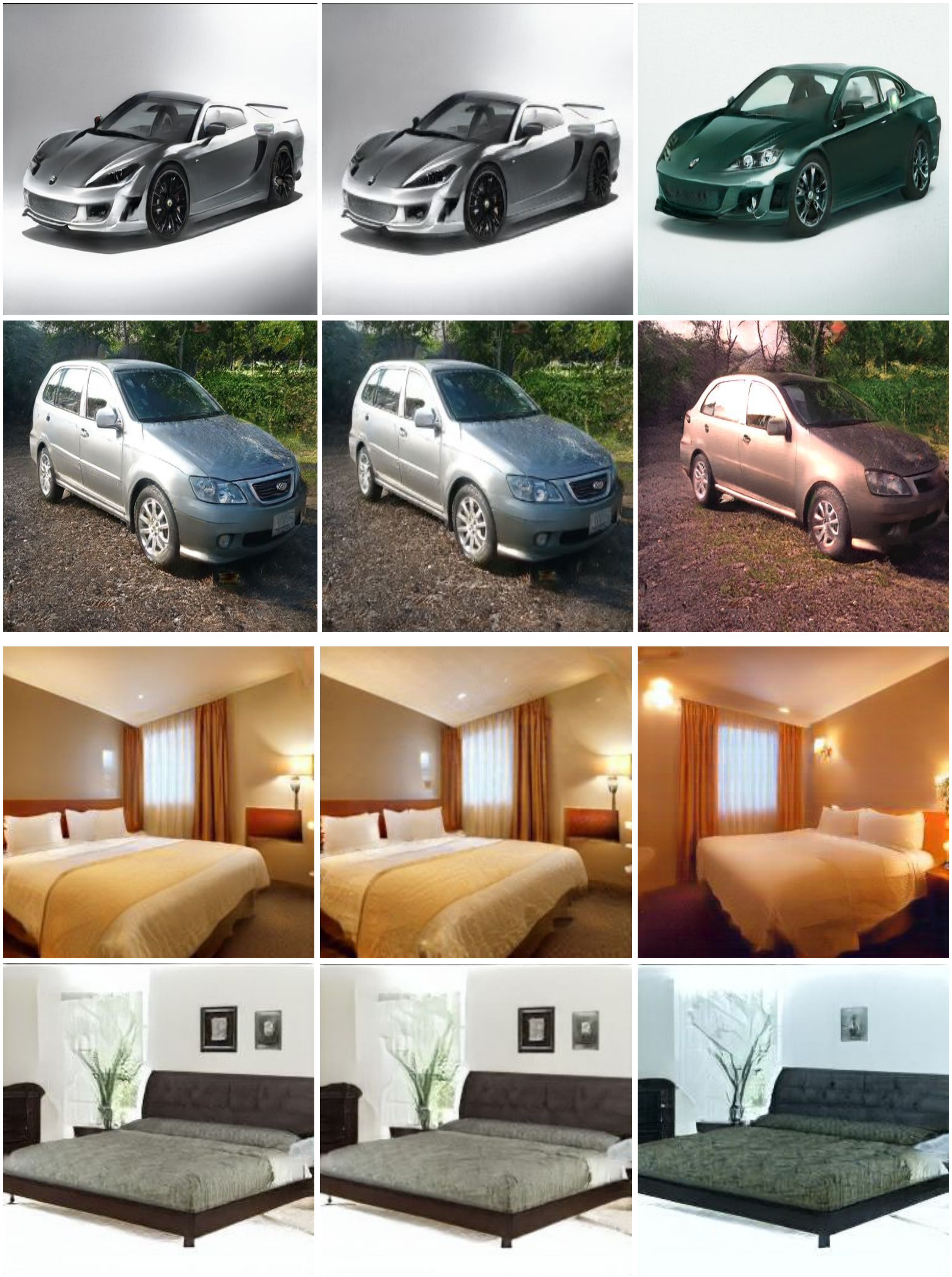


Figure 8: (cont.) High resolution versions of adversarial examples. From left to right: original, noise-based and style-based images.