SUPPLEMENTARY MATERIAL: FINE-GRAINED SYNTHESIS OF UNRESTRICTED ADVERSARIAL EXAMPLES

Anonymous authors
Paper under double-blind review

1 COMPARISON WITH SONG ET AL. (2018)

We show that adversarial training with examples generated by Song et al. (2018) hurts the classifier's performance on clean images. Table 1 demonstrates the results. We use the same classifier architectures as Song et al. (2018) 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 training with our examples improves the accuracy, demonstrating difference of our approach with that of Song et al. (2018).

	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 Song et al. (2018).

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 Song et al., 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.

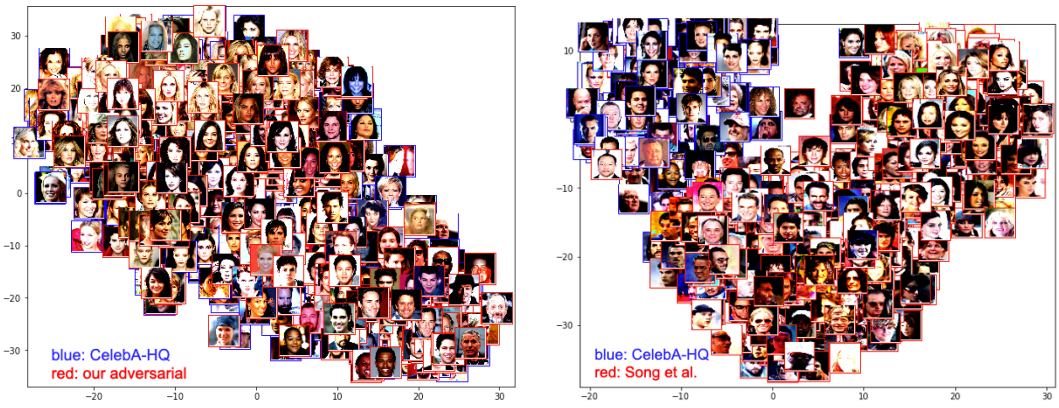


Figure 1: t-SNE plot comparing distributions of real images with adversarial examples from our approach and Song et al.

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.

	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 $\pm$ std) required to fool the classifier.

3 Object Detection Results

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

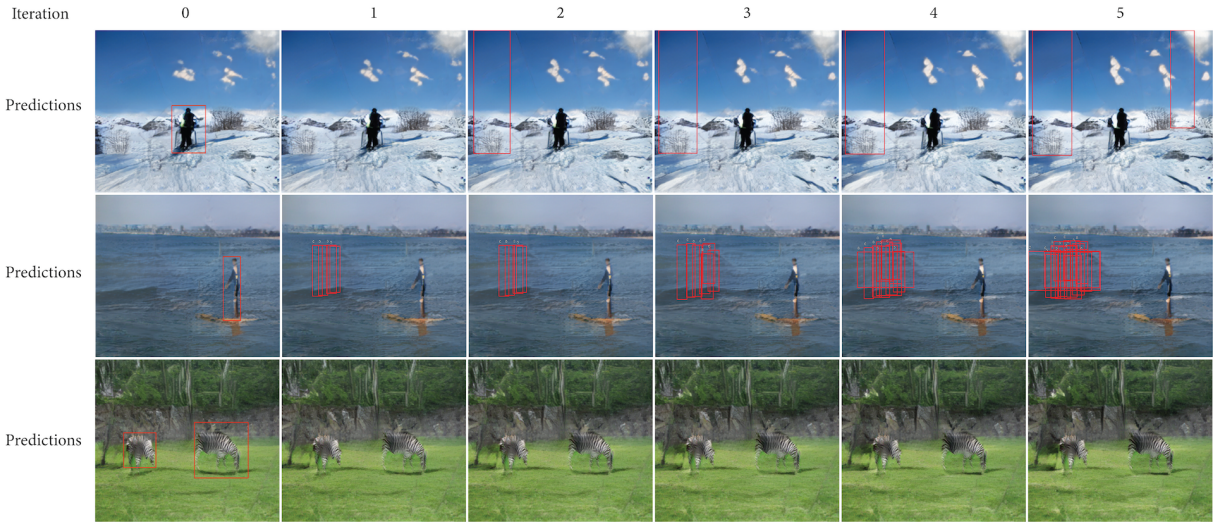


Figure 2: Unrestricted adversarial examples for object detection. Generated images and their corresponding predictions are shown for different number of iterations.

4 Additional Examples

We also provide additional examples and higher-resolution images in the following. Figure 3 depicts additional examples on the segmentation task. Figure 4 illustrates adversarial examples on CelebA-HQ gender classification, and Figure 5 shows additional examples on the LSUN dataset. Higher-resolution versions for some of the adversarial images are shown in Figure 6, which particularly helps to distinguish subtle differences between original and noise-based images.

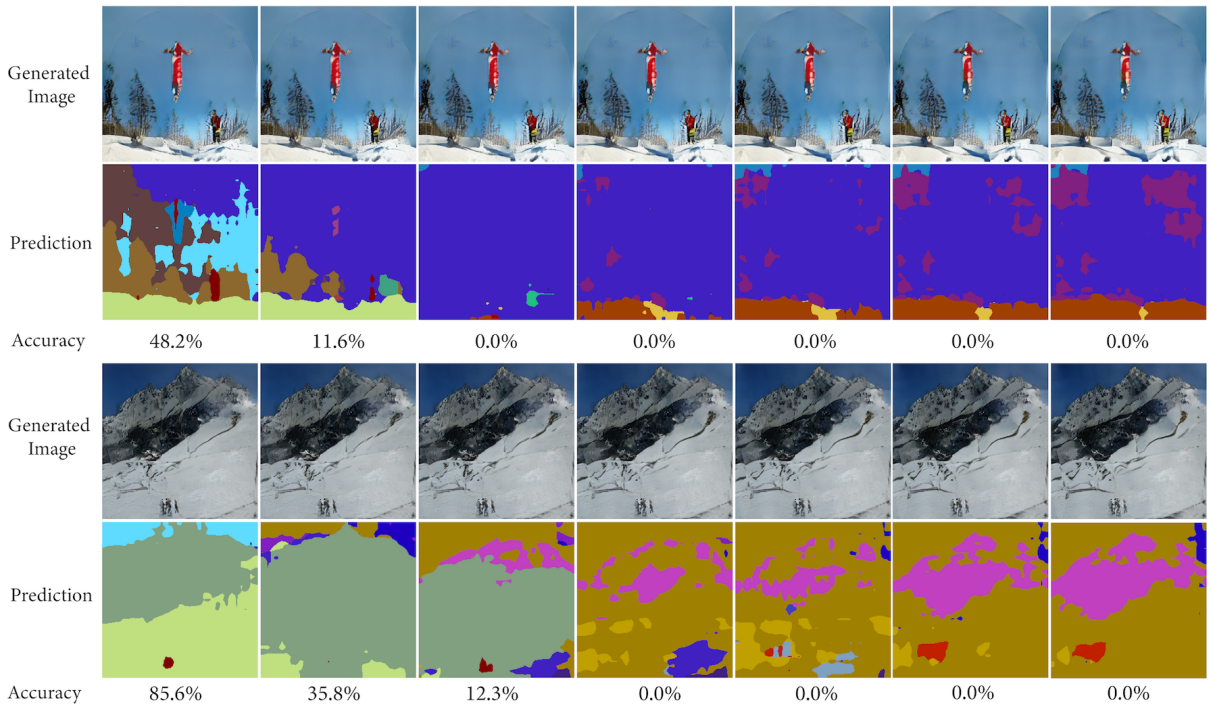
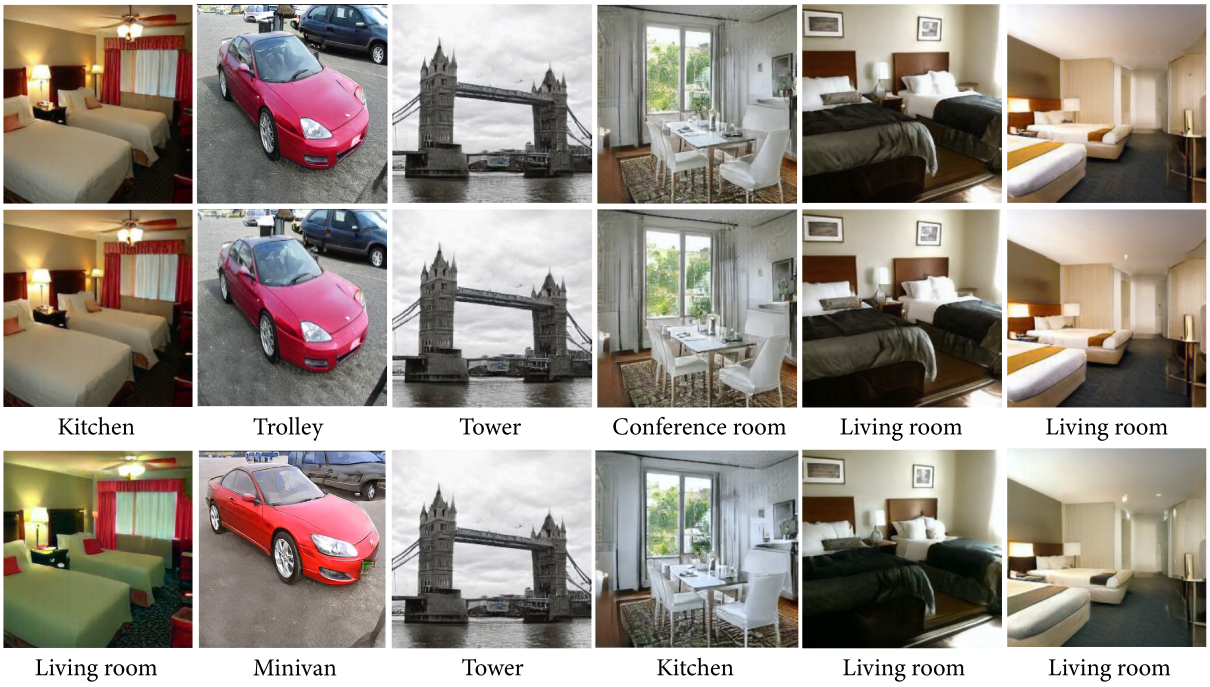


Figure 3: Unrestricted adversarial examples for semantic segmentation. Generated images, corresponding predictions and their accuracy (ratio of correctly predicted pixels) are shown for different number of iterations.

Figure 4: 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 5: Unrestricted adversarial examples on LSUN for a) non-targeted and b) targeted attacks. From top to bottom: original, noise-based and style-based images.

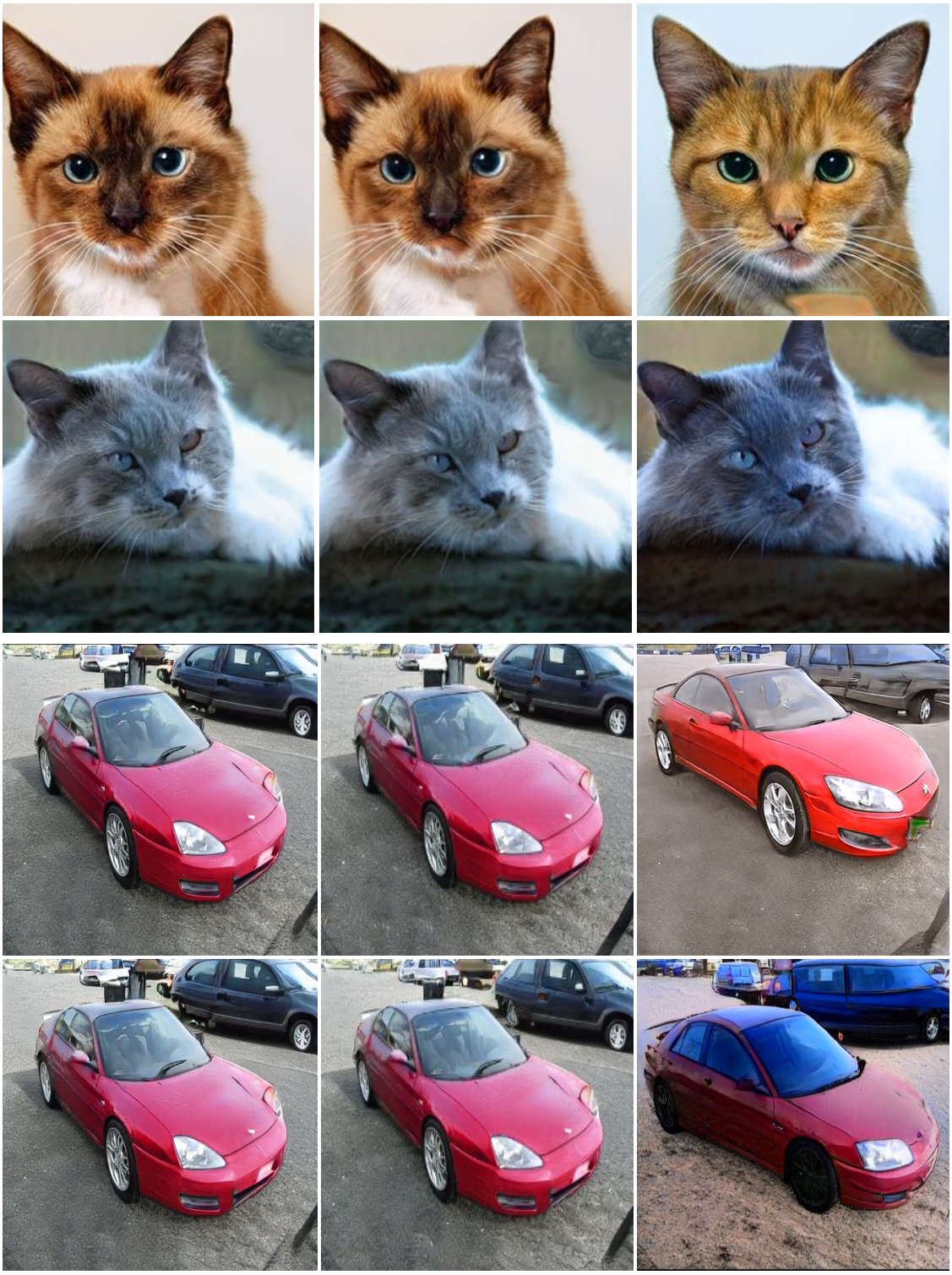


Figure 6: High resolution versions of adversarial images. From left to right: original, noise-based and style-based images.

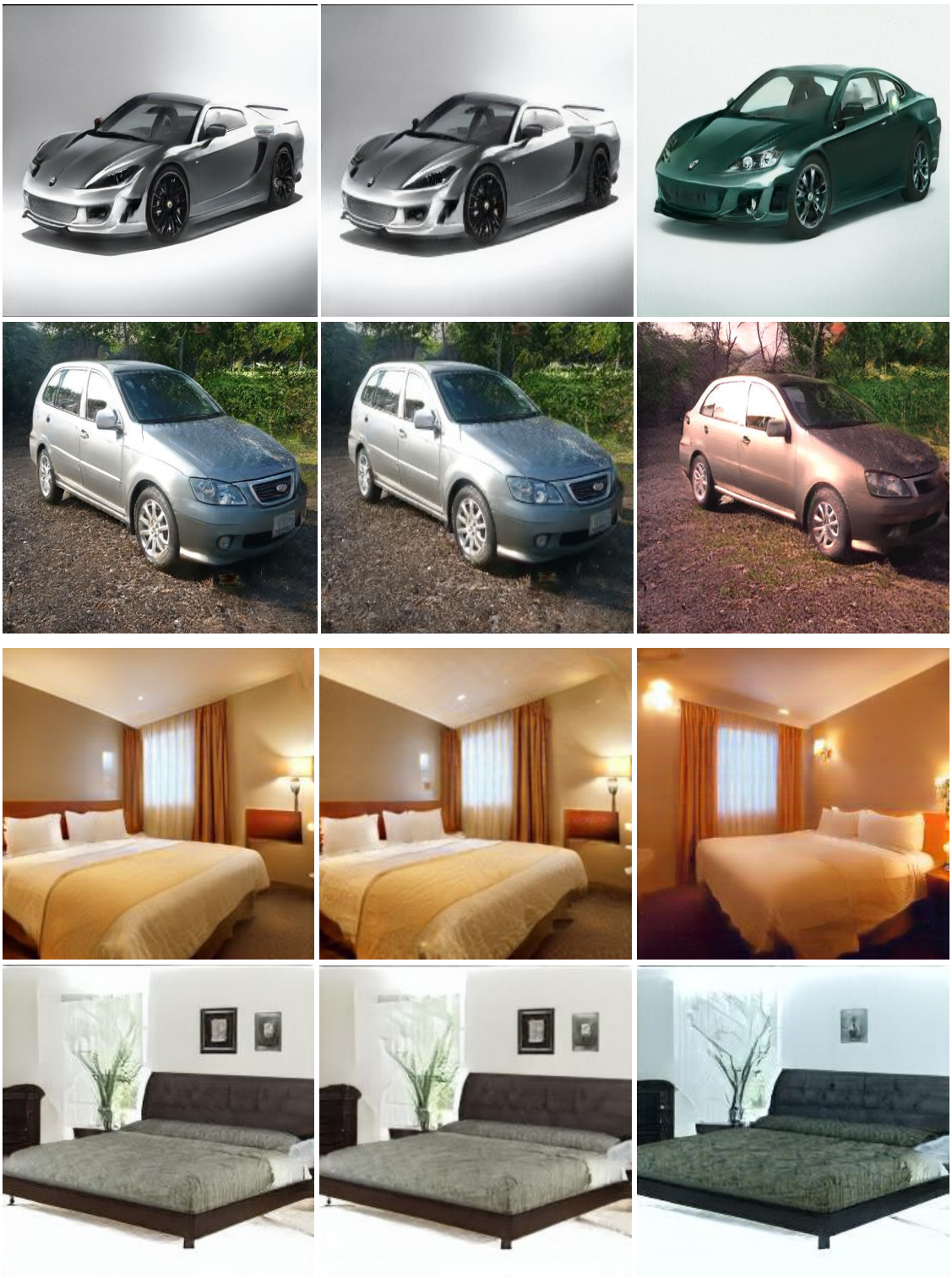


Figure 6: (cont.) High resolution versions of adversarial examples. From left to right: original, noise-based and style-based images.