Supplementary Material for Generative Adversarial Perturbations

Omid Poursaeed[1,2] Isay Katsman[1] Bicheng Gao[3,1] Serge Belongie[1,2]
[1]Cornell University [2]Cornell Tech [3]Shanghai Jiao Tong University
{op63,isk22,bg455,sjb344}@cornell.edu

1. Runtime Analysis

Note that inference time is not an issue for universal perturbations as we just need to add the perturbation to the input image during inference. Therefore, we provide running time only for image-dependent perturbations. In this case, we need to forward the input image to the generator and get the resulting perturbation. Table 1 demonstrates the inference time for image-dependent perturbations. It also shows the generator's architecture for each task including the number of filters in the first layer. We perform model-level parallelization across two GPUs, and batch size is set to be one. Notice that inference time is in the order of milliseconds, allowing us to generate perturbations in real-time. Table 2 shows inference time for the segmentation task. Two architectures with similar performance are given. Here we deal with 1024×512 images in the Cityscapes dataset, and we need models with more capacity; hence, the inference time is larger compared with the classification task.

Task	Architecture	Titan Xp	Tesla K40
Non-targeted	ResNet Gen. 6 blocks, 50 filters	0.27 ms	4.7 ms
Targeted	ResNet Gen. 6 blocks, 57 filters	0.28 ms	4.8 ms

Table 1: Average inference time per image and generator's architecture for image-dependent classification tasks. Target model is Inception-v3.

Architecture	Titan Xp	Tesla K40m
U-Net Generator: 8 layers, 200 filters	132.8 ms	511.7 ms
ResNet Generator: 9 blocks, 145 filters	335.7 ms	2396.9 ms

Table 2: Average inference time per image and generator's architecture for the semantic segmentation task. Targeted image-dependent perturbations are considered with FCN-8s as the pre-trained model.

2. Resistance to Gaussian Blur

We examine the effect of applying Gaussian filters to perturbed images. Results for the classification task are shown in Table 3. In order to be comparable with [26], we consider non-targeted image-dependent perturbations with Destruction Rate (fraction of images that are no longer misclassified after blur) as the metric. For most σ values, our method is more resistant to Gaussian blur than I-FGSM.

We also evaluate the effect of Gaussian filters for the segmentation task. Results are given in Table 4. As we can observe, the perturbations are reasonably robust to Gaussian blur.

	$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1$	$\sigma = 1.25$
GAP	0.0%	0.8%	3.2%	8.0%
I-FGSM	0.0%	0.5%	8.0%	23.0%

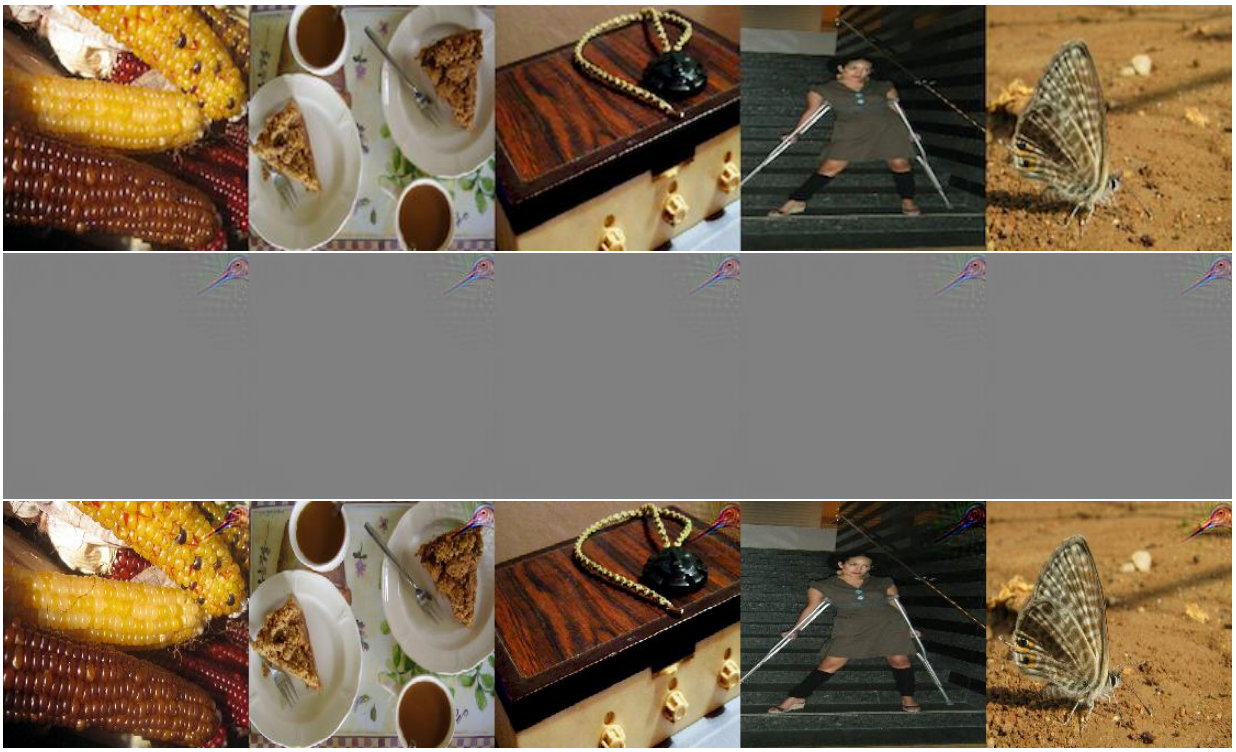
Table 3: Destruction Rate of non-targeted image-dependent perturbations for the classification task. Perturbation norm is set to $L_\infty = 16$.

	$\sigma = 0.5$	$\sigma = 0.75$	$\sigma = 1$	$\sigma = 1.25$
$L_\infty = 5$	83.2%	76.9%	66.0%	57.1 %
$L_\infty = 10$	94.8%	90.1%	80.0%	69.6%
$L_\infty = 20$	97.5%	95.7%	89.3%	78.8%

Table 4: Success rate of targeted image-dependent perturbations for the segmentation task after applying Gaussian filters.

3. Additional Examples

More examples of both classification and segmentation adversarial perturbations are given in the following figures.



(a) Target model: VGG-19, Fooling ratio: 94.9%

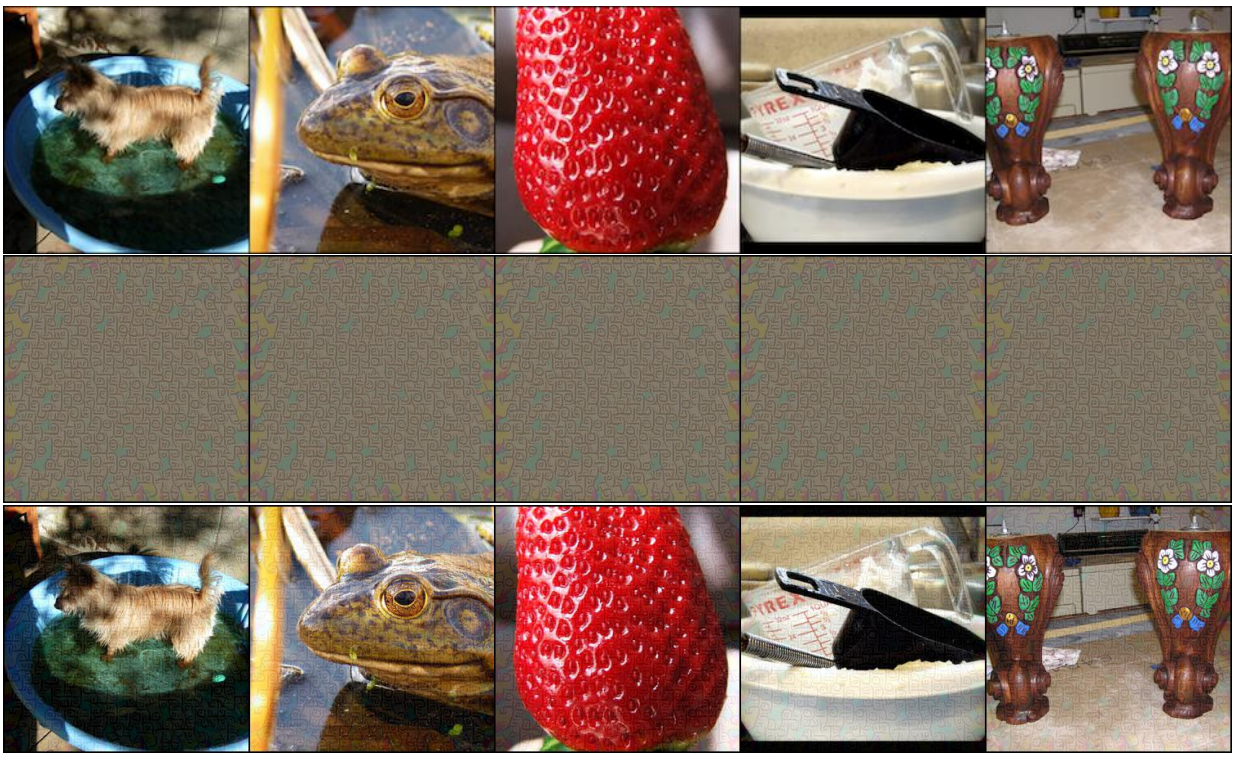
(b) Target model: VGG-16, Fooling ratio: 93.9%

Figure 1: Non-targeted universal perturbations. From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_2 = 2000$ for (a) and (b) and to $L_\infty = 10$ for (c) and (d).

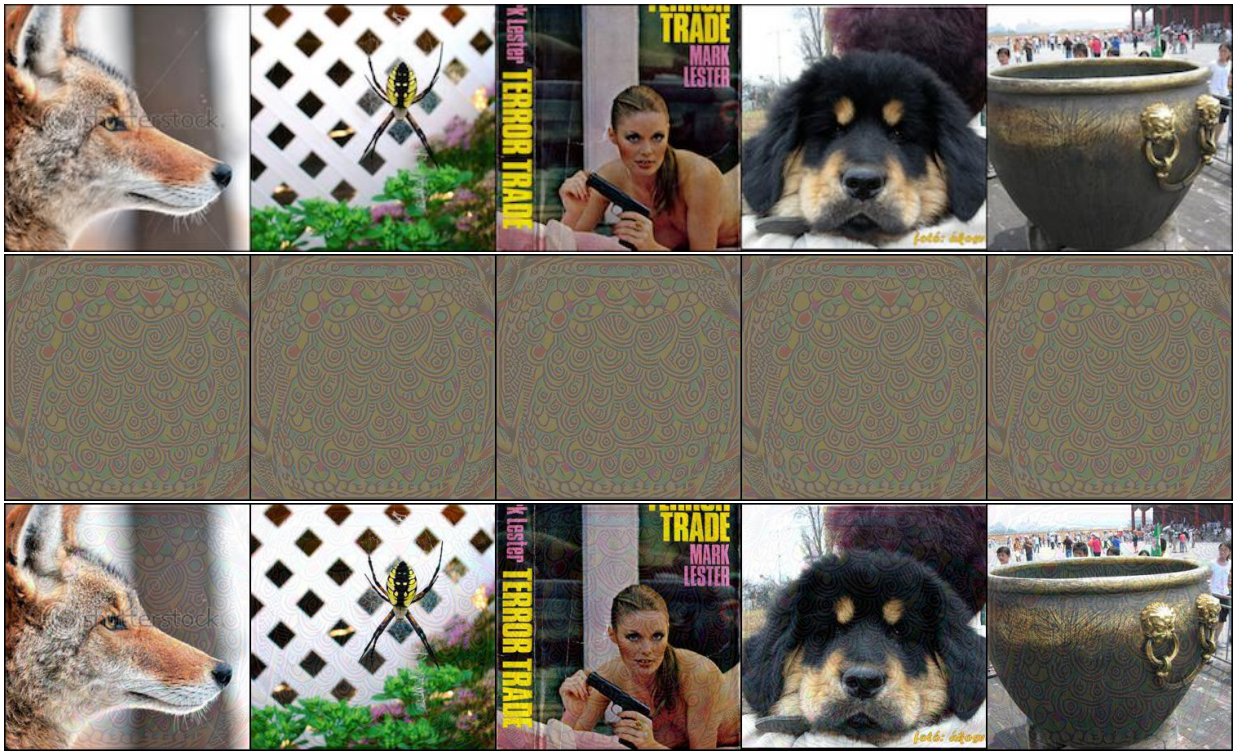
(c) Target model: Inception-v3, Fooling ratio: 79.2%

(d) Target model: VGG-19, Fooling ratio: 80.1%

Figure 1: Non-targeted universal perturbations (continued). From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_2 = 2000$ for (a) and (b) and to $L_\infty = 10$ for (c) and (d).

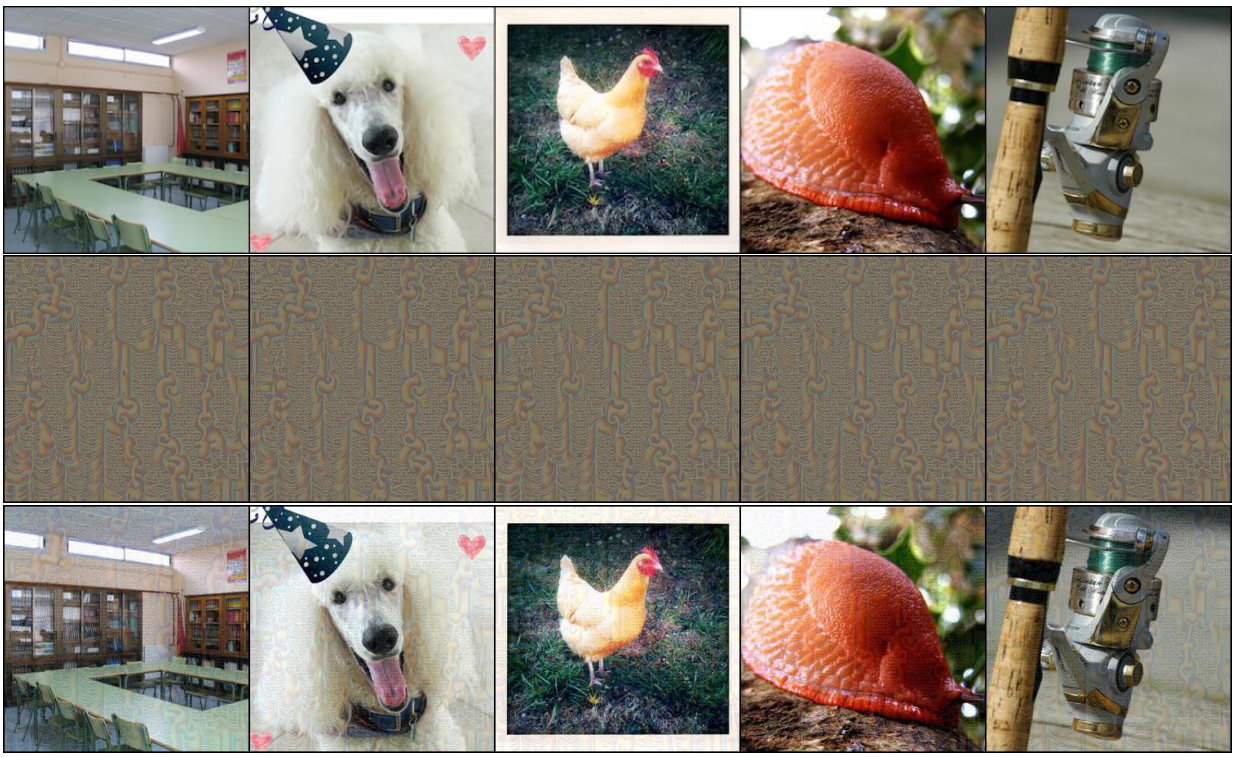


(a) Target: Jigsaw Puzzle, Top-1 target accuracy: 89.3%

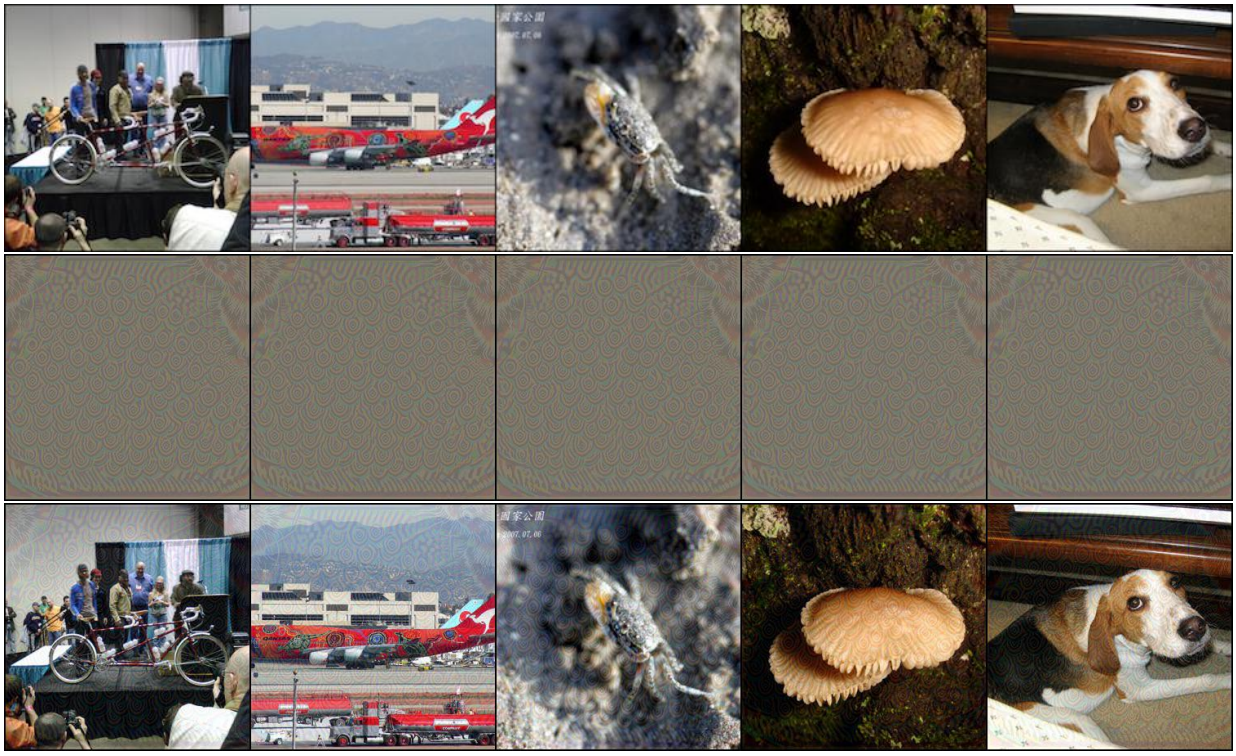


(b) Target: Teapot, Top-1 target accuracy: 62.2%

Figure 2: Targeted universal perturbations. From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_\infty = 10$, and target model is Inception-v3.



(c) Target: Chain, Top-1 target accuracy: 64.9%

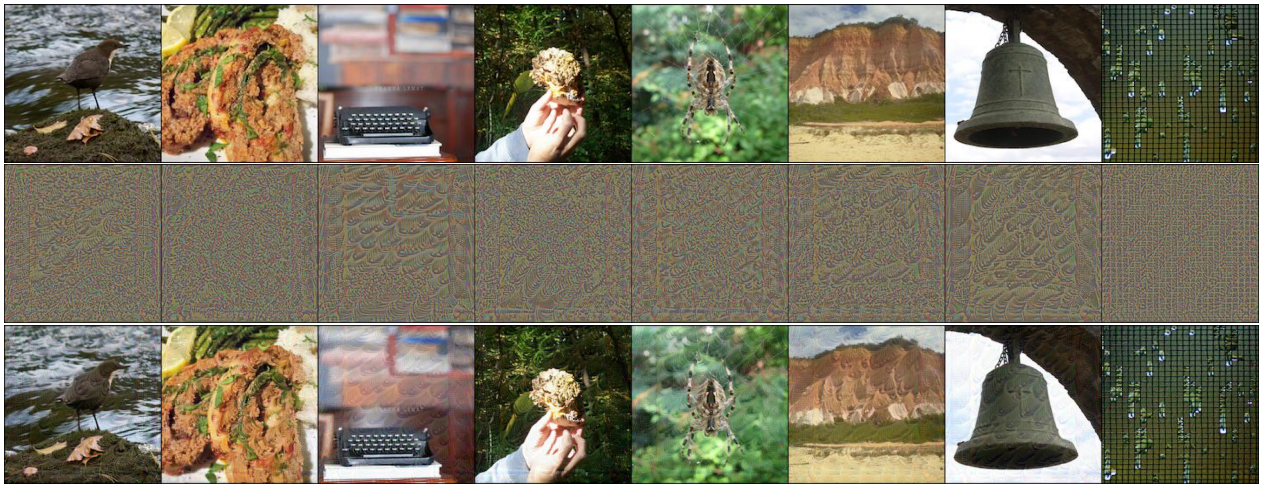


(d) Target: Hamster, Top-1 target accuracy: 60.0%

Figure 2: Targeted universal perturbations (continued). From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_\infty = 10$, and target model is Inception-v3.

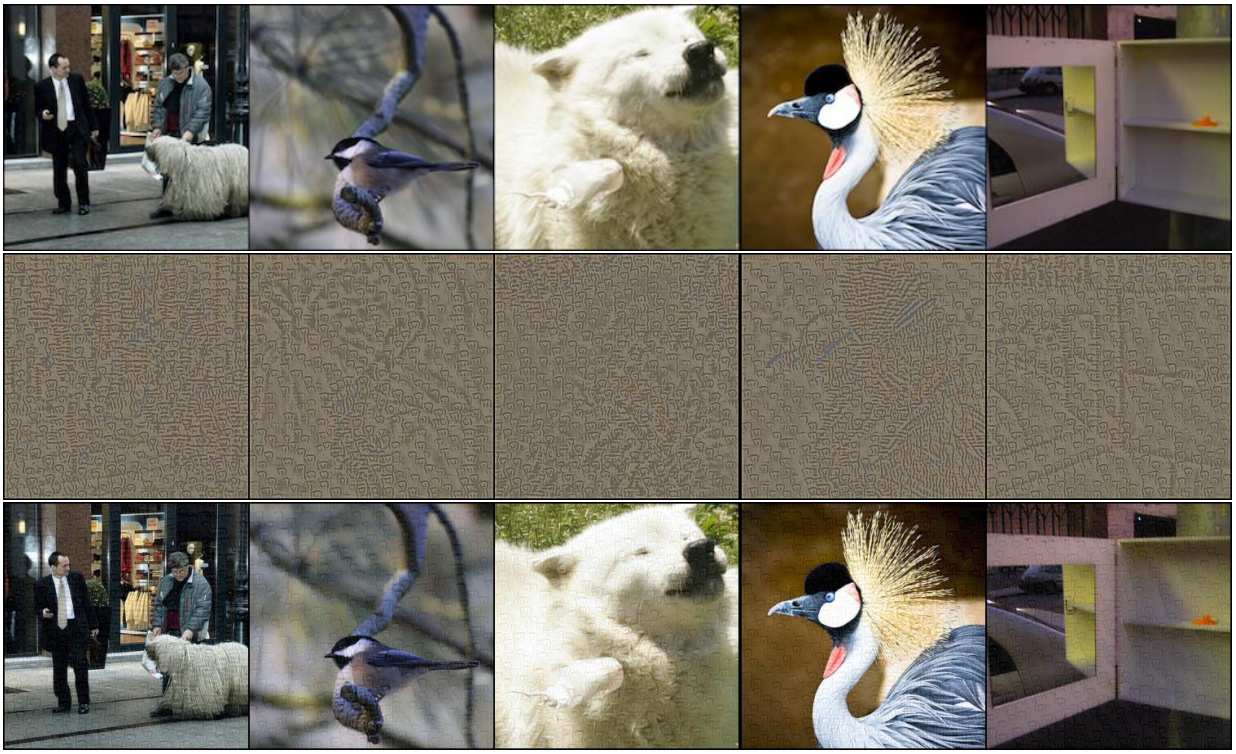
(a) $L_\infty = 7$

(b) $L_\infty = 10$



(c) $L_\infty = 13$

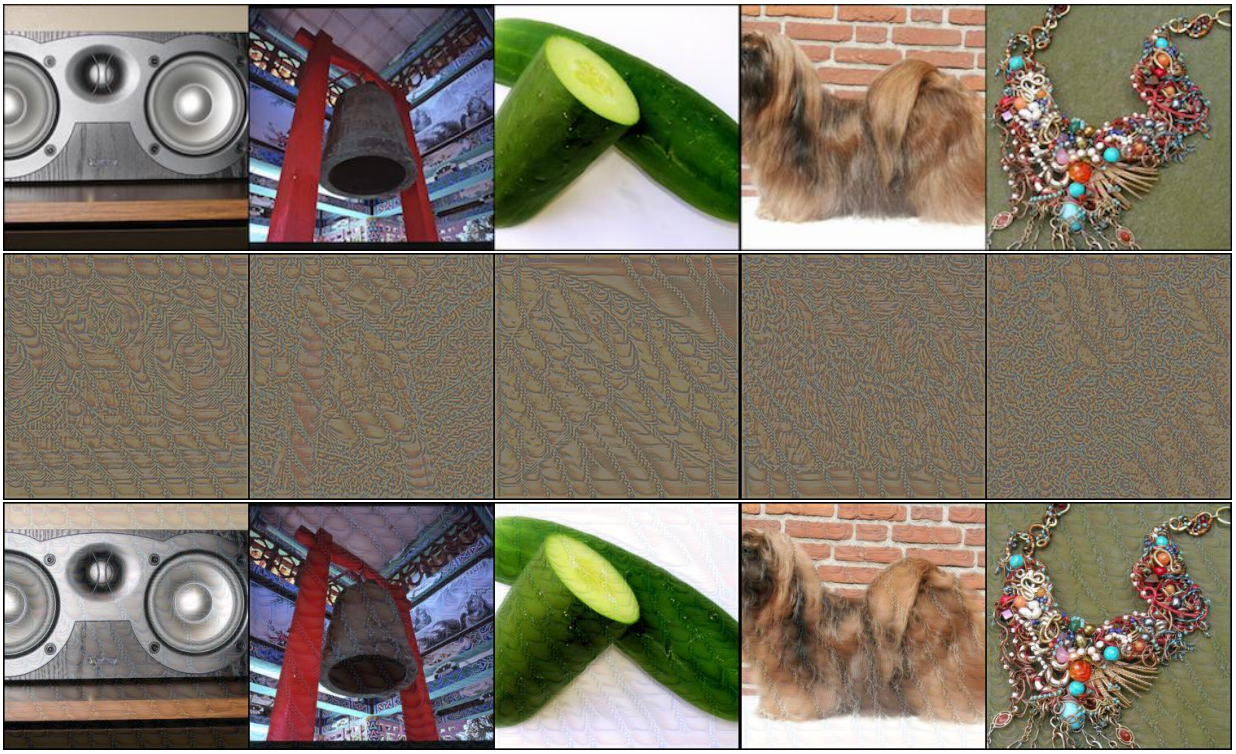
Figure 3: Non-targeted image-dependent perturbations. From top to bottom: original image, enhanced perturbation and perturbed image. Three different thresholds are considered with Inception-v3 as the target model.



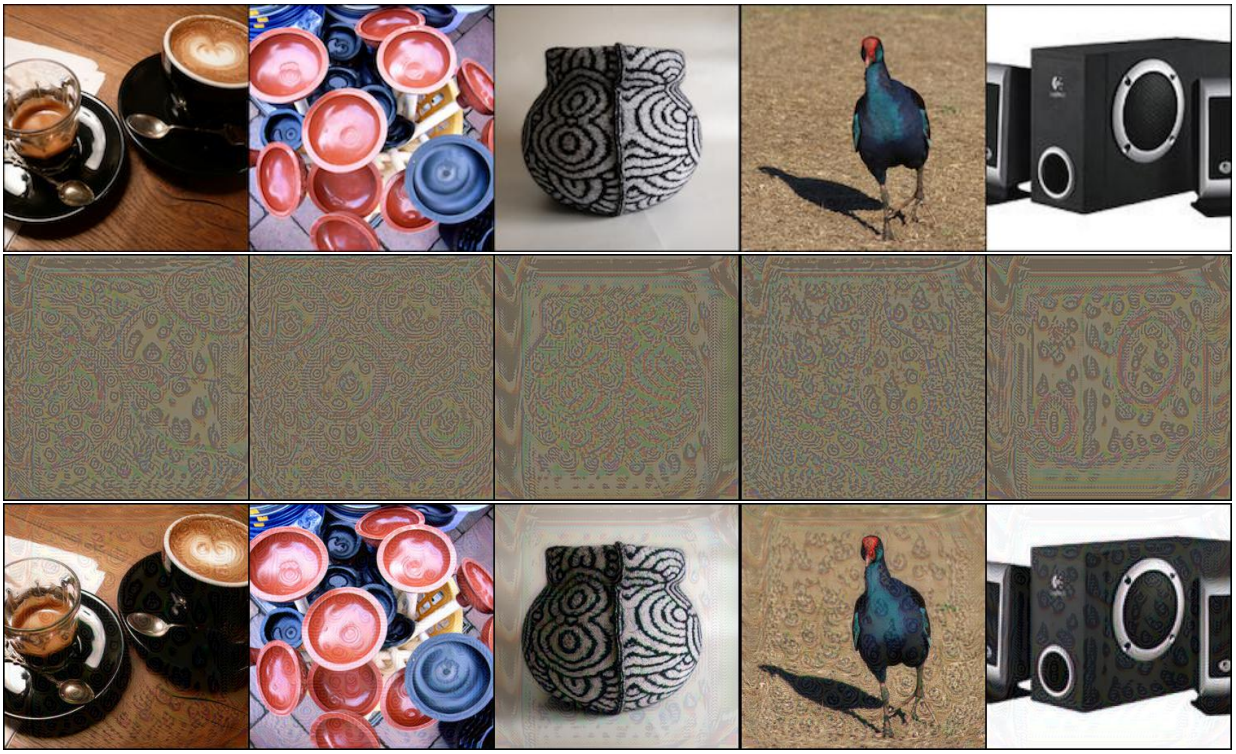
(a) Target: Jigsaw puzzle, Top-1 target accuracy: 98.1%

(b) Target: Knot, Top-1 target accuracy: 95.0%

Figure 4: Targeted image-dependent perturbations. From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_\infty = 10$, and Inception-v3 is the pre-trained model.

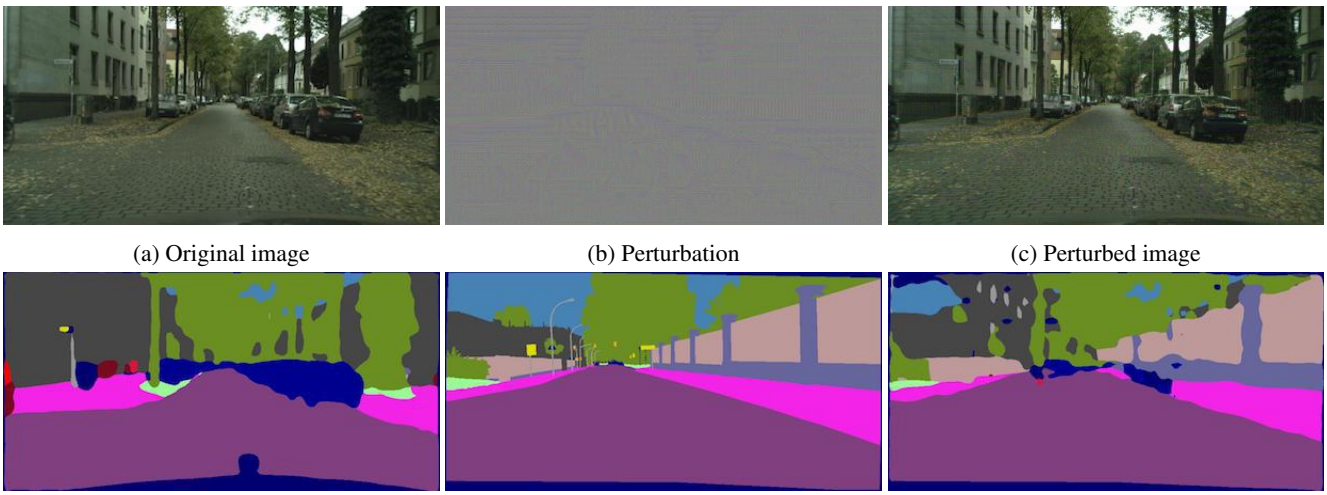


(c) Target: Chain, Top-1 target accuracy: 89.7%



(d) Target: Teapot, Top-1 target accuracy: 90.6%

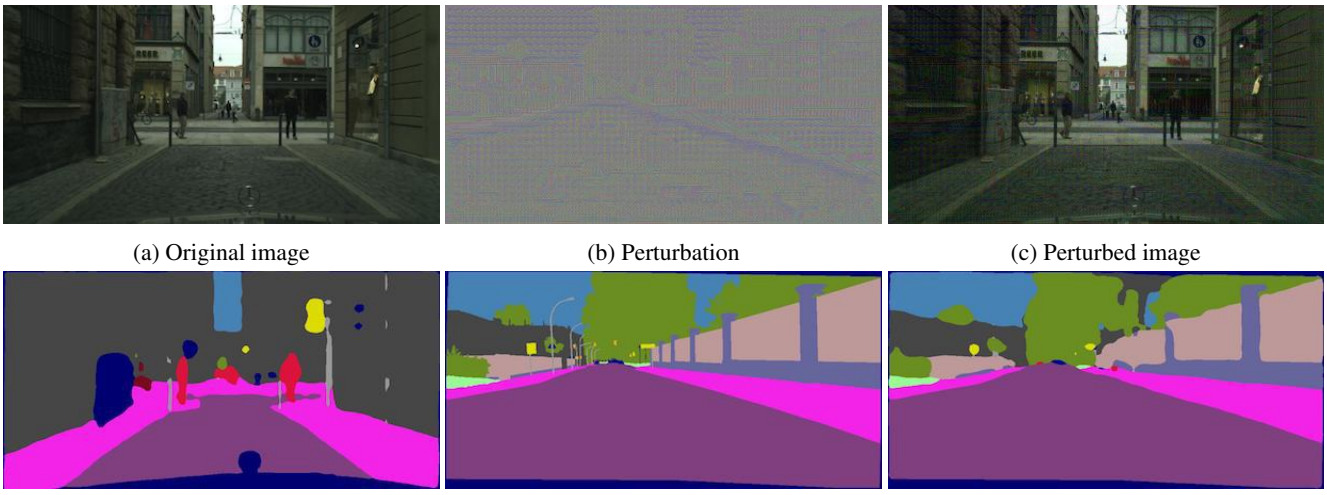
Figure 4: Targeted image-dependent perturbations (continued). From top to bottom: original image, enhanced perturbation and perturbed image. Perturbation norm is set to $L_\infty = 10$, and Inception-v3 is the pre-trained model.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

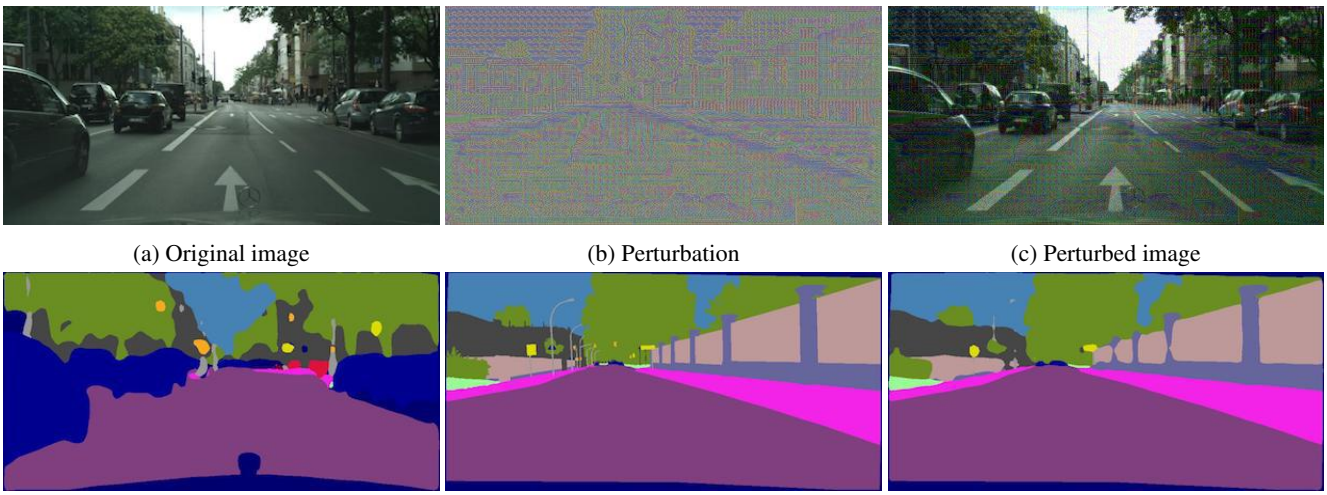
Figure 5: Targeted universal perturbations with $L_\infty = 5$. Zoom in for details.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

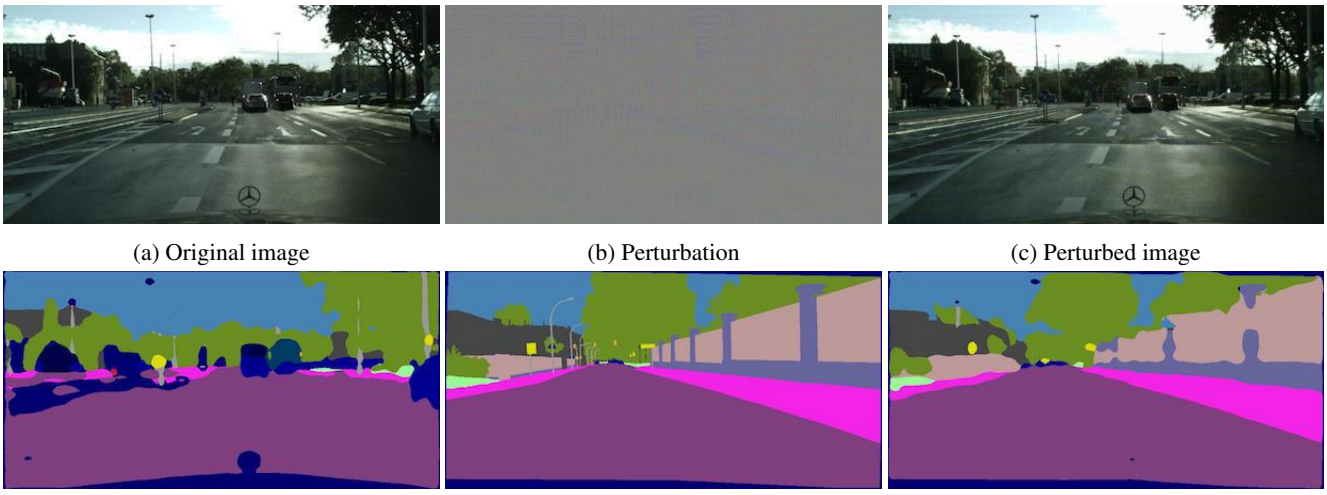
Figure 6: Targeted universal perturbations with $L_\infty = 10$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

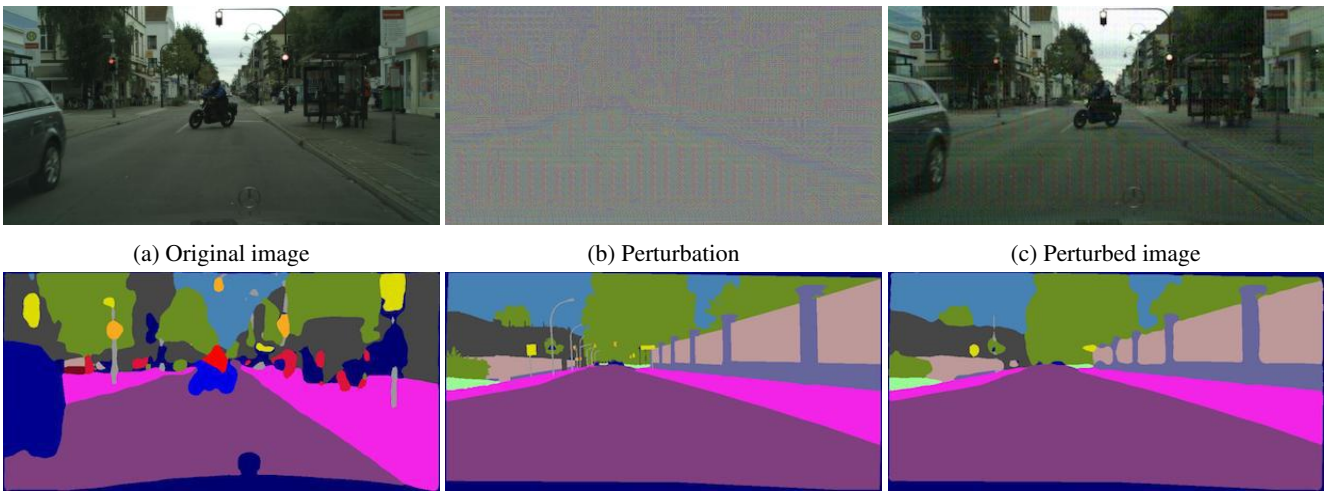
Figure 7: Targeted universal perturbations with $L_\infty = 20$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

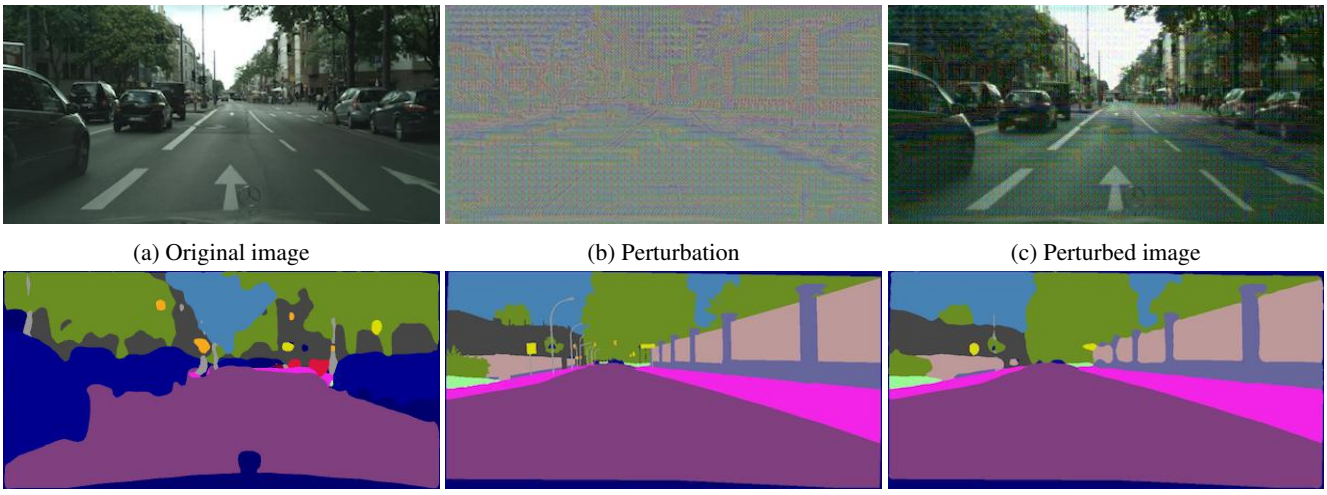
Figure 8: Targeted image-dependent perturbations with $L_\infty = 5$. Zoom in for details.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

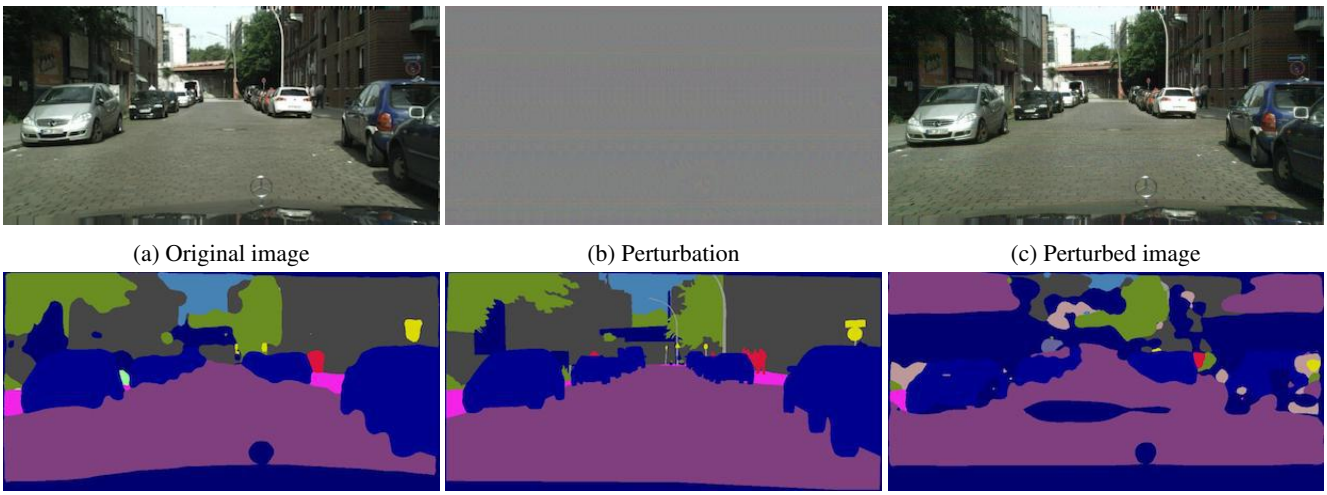
Figure 9: Targeted image-dependent perturbations with $L_\infty = 10$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Target (f) Prediction for perturbed image

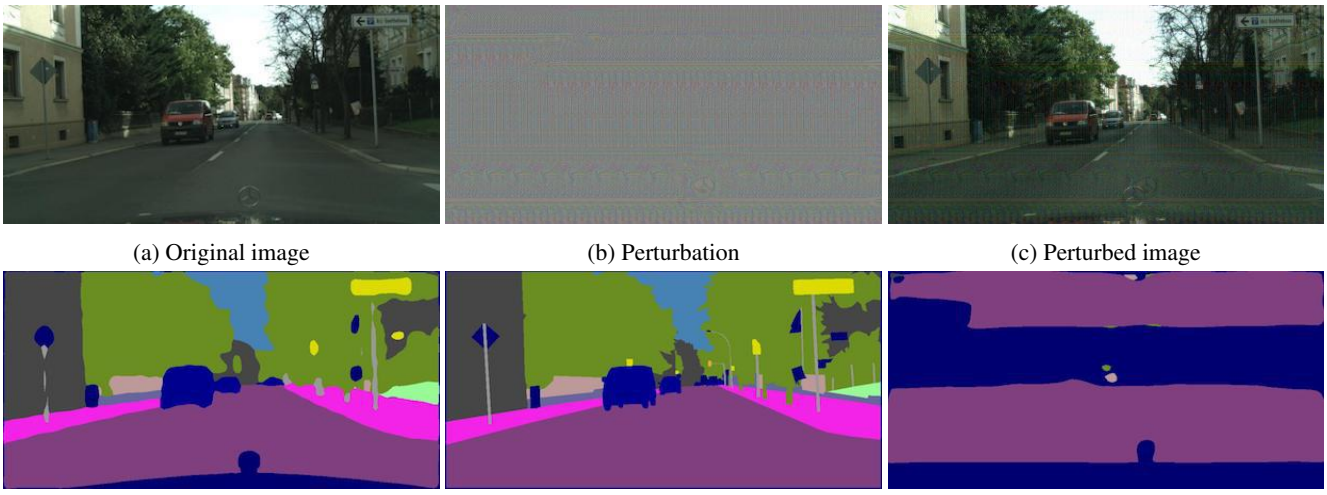
Figure 10: Targeted image-dependent perturbations with $L_\infty = 20$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

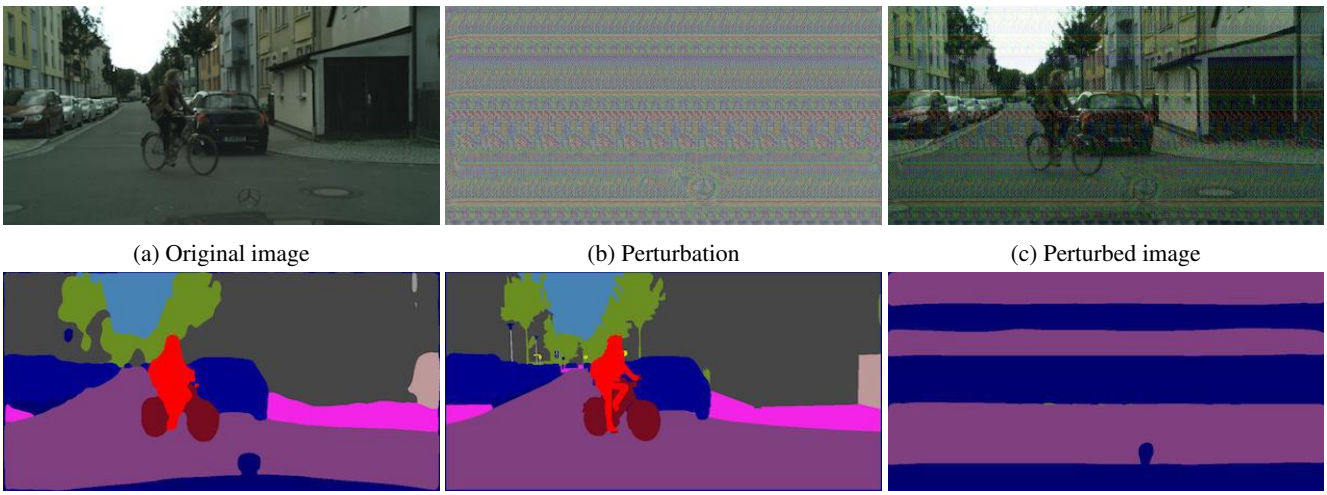
Figure 11: Non-targeted universal perturbations with $L_\infty = 5$. Zoom in for details.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

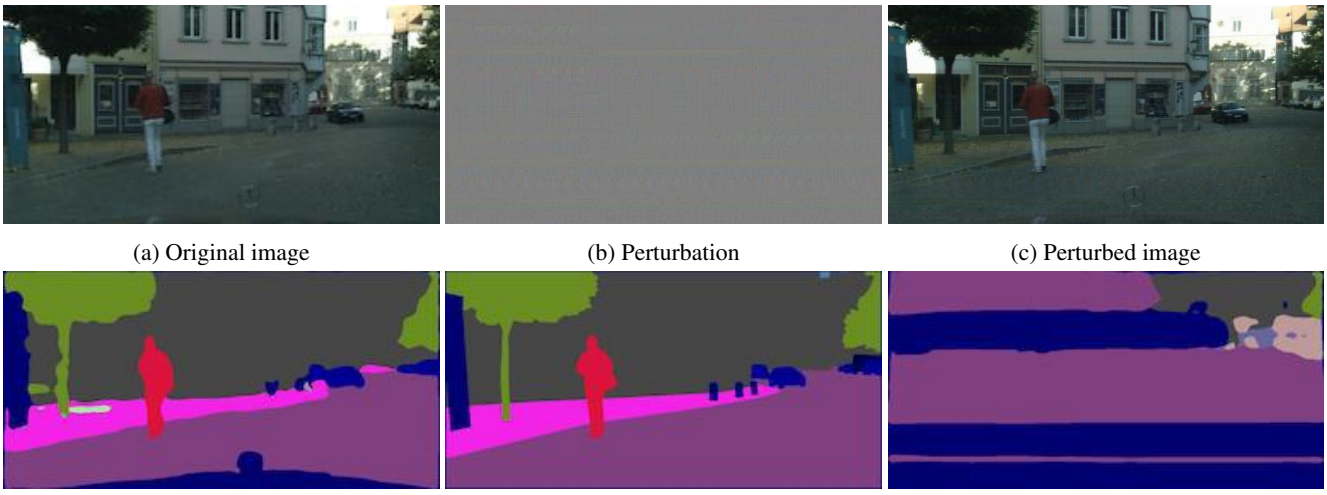
Figure 12: Non-targeted universal perturbations with $L_\infty = 10$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

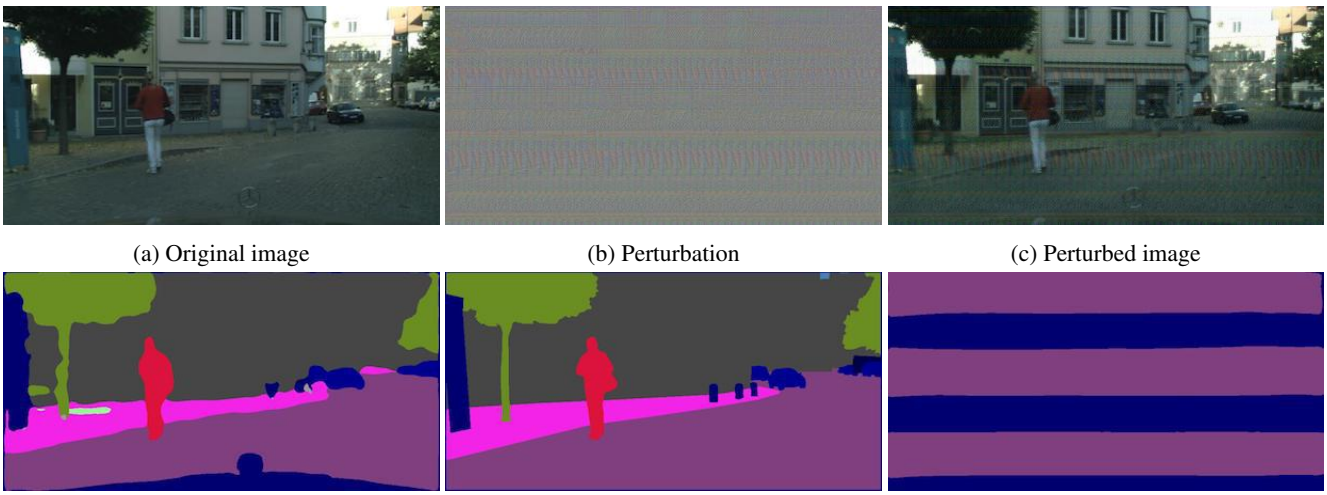
Figure 13: Non-targeted universal perturbations with $L_\infty = 20$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

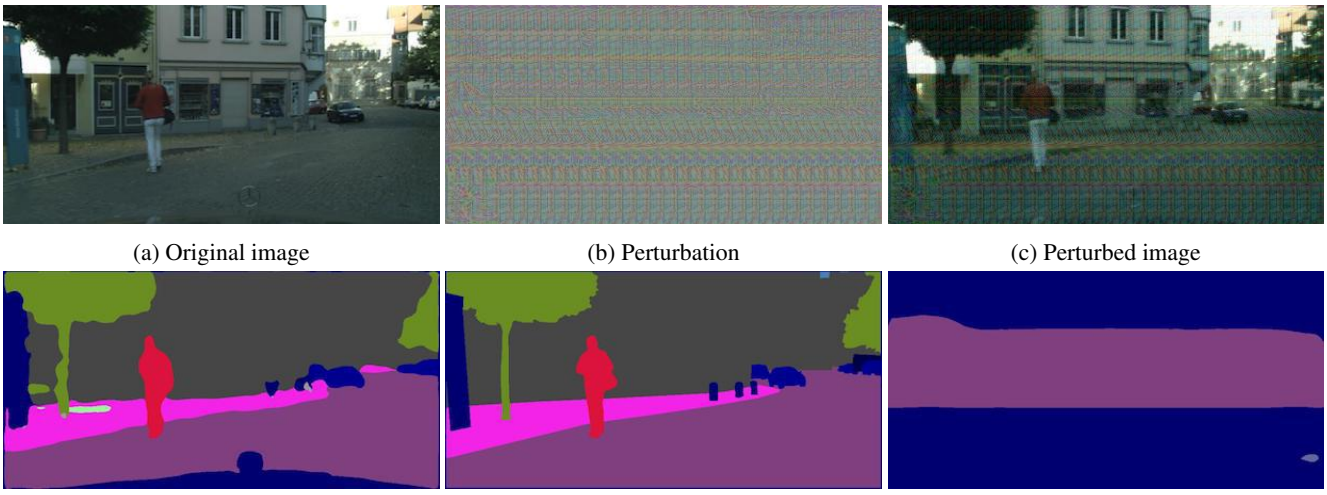
Figure 14: Non-targeted image-dependent perturbations with $L_\infty = 5$. Zoom in for details.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

Figure 15: Non-targeted image-dependent perturbations with $L_\infty = 10$.



(a) Original image (b) Perturbation (c) Perturbed image

(d) Prediction for original image (e) Groundtruth (f) Prediction for perturbed image

Figure 16: Non-targeted image-dependent perturbations with $L_\infty = 20$.