Semantic Adversarial Attacks via Diffusion Models (Supplementary Material)

Chenan Wang[1]
cw3344@drexel.edu
Jinhao Duan[1]
jd3734@drexel.edu
Chaowei Xiao[2]
xiaocw@umich.edu
Edward Kim[1]
ek826@drexel.edu
Matthew Stamm[3]
mcs382@drexel.edu
Kaidi Xu[1]
kx46@drexel.edu

[1] Department of Computer Science
College of Computing & Informatics
Drexel University
Philadelphia, USA

[2] The Information School
University of Wisconsin
Madison, USA

[3] Electrical and Computer Engineering
Drexel University
Philadelphia, USA

1 Algorithms of our framework

The ST and LM approaches of our framework are shown in Algorithm 1 and Algorithm 2, respectively. The equations mentioned in these algorithms are listed below.
Diffusion process:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{w},\ \mathbf{w} \sim \mathcal{N}(\mathbf{0},\mathbf{I}). \tag{1}$$

Sampling process:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{f}_\theta(\mathbf{x}_t,t) + \sqrt{1-\bar{a}_{t-1}-\sigma_t^2}\varepsilon_\theta(\mathbf{x}_t,t) + \sigma_t^2\mathbf{z}. \tag{2}$$

Loss function in the fine-tuning process:

$$\mathcal{L}_{ST} = \min_{\hat{\theta},\hat{\mathbf{x}}_T} \lambda D_{LPIPS}(\mathbf{x}_0, \hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T)) - D_{KL}(f(\mathbf{x}_0), f(\hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T))). \tag{3}$$

Three strategies to generate a mask $\hat{\mathbf{m}}$:

$$\hat{\mathbf{m}}_s(\delta) = TopK(|\mathbf{m}_s|,\delta), \quad \hat{\mathbf{m}}_t(\delta) = TopK(|\mathbf{m}_t|,\delta), \quad \hat{\mathbf{m}}_{s+t}(\delta) = TopK(|\mathbf{m}_s|+|\mathbf{m}_t|,\delta). \tag{4}$$

Modify the original latent space:

$$\hat{\mathbf{x}}_T(\delta) = (1-\hat{\mathbf{m}}(\delta)) * \mathbf{x}_T^s + \hat{\mathbf{m}}(\delta) * \mathbf{x}_T^t \tag{5}$$

Heuristic to control the decremental speed of δ:

$$\delta = \delta - \max(\gamma \frac{z_y - \max_{i \neq y} z_i}{z_y}, 1), \tag{6}$$

Algorithm 1 ST approach

Input: Input image $\mathbf{x}_0$, diffusion model with weight parameters θ, number of iterations N, s_{ft} and s_{sp};
Output: Semantic adversarial image $\hat{\mathbf{x}}_0$;

$\mathbf{x}_T \leftarrow$ update by Eq. (1) with $\mathbf{x}_0$;
Initialize $\hat{\mathbf{x}}_T \leftarrow \mathbf{x}_T$, $\hat{\theta} \leftarrow \theta$;
for $i = 0$ **to** $N-1$ **do**
  $\hat{\mathbf{x}}_0 \leftarrow$ update by Eq. (2) with $\hat{\theta}$ and $\hat{\mathbf{x}}_T$ in s_{ft} steps;
  Optimize $\hat{\mathbf{x}}_T$ and/or $\hat{\theta}$ by Eq. (3);
  if $f(\hat{\mathbf{x}}_0) \neq f(\mathbf{x}_0)$ **then**
    $\hat{\mathbf{x}}_0 \leftarrow$ update by Eq. (2) with $\hat{\theta}$ and $\hat{\mathbf{x}}_T$ in s_{sp} steps;
    if $f(\hat{\mathbf{x}}_0) \neq f(\mathbf{x}_0)$ **then**
      Early stop and return $\hat{\mathbf{x}}_0$;
    end if
  end if
end for
$\hat{\mathbf{x}}_0 \leftarrow$ update by Eq. (2) with $\hat{\theta}$ and $\hat{\mathbf{x}}_T$ in s_{sp} steps;
return $\hat{\mathbf{x}}_0$

Algorithm 2 LM approach

Input: A pair of source and target image as $\mathbf{x}_0^s$ and $\mathbf{x}_0^t$, an interpretation map function g, threshold δ;
Output: Semantic adversarial image $\hat{\mathbf{x}}_0$;

$\mathbf{x}_T^s, \mathbf{x}_T^t \leftarrow$ Eq. (1) with $\mathbf{x}_0^s$ and $\mathbf{x}_0^t$ respectively;
Obtain original interpretation maps $\mathbf{m}_s$ and $\mathbf{m}_t$ with function $\mathbf{m} = g(\mathbf{x}_0, \mathbf{y})$;
Initialize $\delta \leftarrow 100$;
while $\delta \neq 0$ **do**
  $\hat{\mathbf{m}} \leftarrow$ update by Eq. (4) with $\mathbf{m}_s$ and $\mathbf{m}_t$;
  $\hat{\mathbf{x}}_T \leftarrow$ update by Eq. (5) with $\hat{\mathbf{m}}$;
  $\hat{\mathbf{x}}_T \leftarrow$ update by Eq. (2) with $\hat{\mathbf{x}}_0$;
  if $f(\hat{\mathbf{x}}_0) \neq f(\mathbf{x}_0)$ **then**
    Early stop and return $\hat{\mathbf{x}}_0$;
  end if
  $\delta \leftarrow$ update by Eq. (6);
end while
$\hat{\mathbf{x}}_0 \leftarrow$ update by Eq. (2) with $\hat{\mathbf{x}}_T$;
return $\hat{\mathbf{x}}_0$

2 Additional Evaluations

2.1 Evaluation on Facial Identity Recognition

All experiments are executed on a single NVIDIA A40 GPU. The classifier is ResNet-18 with 89.63% accuracy. For AttAttack [1] on Celeb-HQ facial identity recognition dataset [3] in this paper, the perturbed attribute is Age to balance quality and ASR. For LatentHSJA [3], we set the number of queries to 1,000 for the same reason. Visual examples for comparison between our methods and other methods are shown in Figure 2. Our semantic attack accuracy before and after natural perturbations are shown in Table 1, where our semantic perturbations are still preserved against natural perturbations: even with the clean accuracy changes 47.8% under Gaussian blur, our semantic attack accuracy only changes 14.6%.

Setting	None	JPEG	Gaussian Blur	Defocus Blur	Brightness
Clean	100.0	100.0	52.2	95.4	100.0
ST black-box fine-tune both	0.0	32.8	14.6	21.8	5.8

Table 1: The accuracy for clean images and the semantic adversarial images (the ST approach, black-box, fine-tuning both the latent space and diffusion model) before and after natural perturbations. Most semantic adversarial perturbations are preserved.

2.2 Evaluation on Face Gender Classification

The classifier is ResNet-18 with 98.38% accuracy. For AttAttack [1] on Celeb-HQ facial gender recognition dataset [3], the perturbed attribute is Smiling to balance quality and ASR. For LatentHSJA [3], we set the number of queries to 1,000 for the same reason. Results on Celeb-HQ facial gender recognition dataset with the ST approach are shown in Table 2, and visual examples for comparison between our framework and other methods are shown in Figure 3.

2.3 Evaluation on AFHQ dataset

The task is animal category recognition into three domains of cat, dog, and wildlife, and the classifier is ResNet-18 with 99.73% accuracy. Since we obtained a pretrained diffusion model on AFHQ-Dog dataset from [2], the diffusion model in our experiments could only generate dog images, and work well with the ST approach. Visual examples using the ST approach of our framework on AFHQ dataset are shown in Figure 1.

Setting	strategy	ASR (%)↑	FID↓	KID↓	average query↓	average time (s)↓
clean images	-	-	30.67	0.000	-	-
LatentHSJA	-	100.0	28.64	0.005	1000 †	45.83
AttAttack, age	-	34.2	72.99	0.029	373.20	94.43
AttAttack, 6 attrs	-	85.80	96.31	0.060	78.86	26.14
ST approach						
fine-tune latent space	white-box	76.3	115.58	0.051	30.58	168.31
fine-tune diffusion model	white-box	78.2	102.29	0.024	16.17	**90.88**
	black-box	96.8	226.81	0.116	26.01	137.49
fine-tune both	white-box	78.2	**98.19**	**0.023**	**16.13**	90.96
	black-box	**97.4**	235.45	0.119	26.02	137.59

† Elapsed time varies, depending on the query steps, which is preset by the user.

Table 2: Our framework with the ST approach on CelebA-HQ gender classification dataset. For AttAttack attack, we run the AttAttack either on age attribute, or 6 attributes: Eyeglasses, Age, Pale Skin, Mustache, Eyebrows, and Hair Color.

Figure 1: Examples of generated images using the ST approach of our framework on AFHQ dataset.

Figure 2: Comparison of generated images between different methods on Celeb-HQ facial identity recognition dataset. For AttAttack, the perturbed attribute is Age.

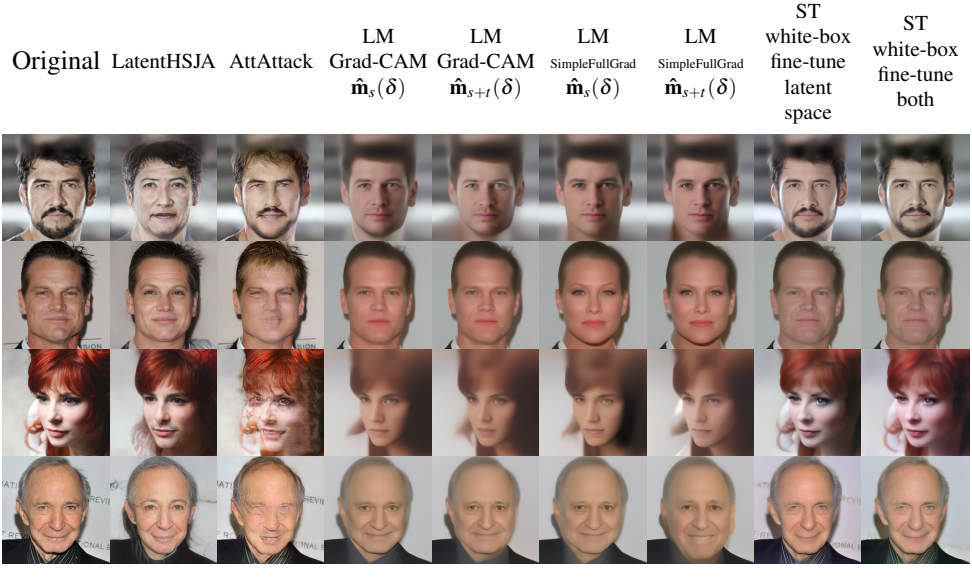

Figure 3: Comparison of generated images between different methods on Celeb-HQ face gender classification dataset. For AttAttack attack, the perturbed attribute is Smiling.

References

[1] Ameya Joshi, Amitangshu Mukherjee, Soumik Sarkar, and Chinmay Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4773–4783, 2019.

[2] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2426–2435, 2022.

[3] Dongbin Na, Sangwoo Ji, and Jong Kim. Unrestricted black-box adversarial attack using gan with limited queries. *arXiv preprint arXiv:2208.11613*, 2022.