Semantic Adversarial Attacks via Diffusion Models

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

Abstract

Traditional adversarial attacks concentrate on manipulating clean examples in the pixel space by adding adversarial perturbations. By contrast, semantic adversarial attacks focus on changing semantic attributes of clean examples, such as color, context, and features, which are more feasible in the real world. In this paper, we propose a framework to quickly generate a semantic adversarial attack by leveraging recent diffusion models since semantic information is included in the latent space of well-trained diffusion models. Then there are two variants of this framework: 1) the **S**emantic **T**ransformation (ST) approach fine-tunes the latent space of the generated image and/or the diffusion model itself; 2) the **L**atent **M**asking (LM) approach masks the latent space with another target image and local backpropagation-based interpretation methods. Additionally, the ST approach can be applied in either white-box or black-box settings. Extensive experiments are conducted on CelebA-HQ and AFHQ datasets, and our framework demonstrates great fidelity, generalizability, and transferability compared to other baselines. Our approaches achieve $\sim$100% attack success rate in multiple settings with the best FID as 36.61. Code is available at https://github.com/steven202/semantic_adv_via_dm.

1 Introduction

Deep neural networks have achieved breakthroughs in many domains [1, 13, 40, 41], however, their intrinsic vulnerabilities to adversarial examples raise security concerns [5, 38, 43, 45]. Most of the literature on adversarial machine learning has been generalized to adversarial perturbations within a ℓ_p norm ball with a small radius ϵ around the clean input example.

Vanilla-trained models achieve high accuracy in classifying benign examples, while misclassifying inputs with such imperceptible perturbations. Instead of globally attacking input images on pixel space, [16, 30] proposed semantic adversarial attacks gaining insight into real-world robustness by manipulating semantically meaningful visual attributes. Semantic attacks may be perceptible; however, such attacks are semantically meaningful and thus hard to detect. Following the concept of semantic adversarial attacks, there is a growing literature on this topic [2, 17, 24, 27, 35, 39]. In the real world, adversarial attacks in ℓ_p-norm based constraint rarely happen due to fragile perturbations. Compared with ℓ_p-norm adversarial attacks in the pixel space, semantic adversarial images are more feasible since they are unrestricted in the magnitude of perturbation while preserving perceptual similarity and realism. Such attacks include changes in texture or any semantic attribute that lead to misclassification.

Currently, there are two existing approaches to implement semantic attacks on clean images: transformations in color or texture [2, 16], or by performing manipulation in the latent space of a generative model [17, 24], such as Generative Adversarial Networks (GANs) [12]. The former leverages various techniques to gather color and texture information, while the latter relies upon attribute annotations to generate semantic adversarial images with generative models such as GANs. Most previous works on semantic adversarial attacks utilize generative models to change attributes, relying on attribute annotations [17, 24], color or texture information [2, 16, 27].

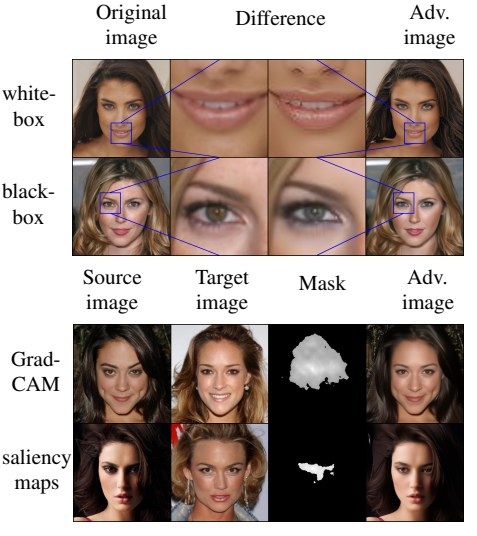

Figure 1: Top two rows: our framework with ST approach under white-box and black-box settings with small semantic changes. Bottom two rows: our framework with LM approach masking by Grad-CAM and saliency maps to transplant features from the target image.

While previous approaches have demonstrated the feasibility of semantic attacks, images that have been attacked using these algorithms can often be easily spotted by the human eye. In order for a semantic attack to be successful, the attacked image should not only fool the classifier but also appear convincingly realistic. Furthermore, existing approaches take a significant time to generate a single attacked image. As a result, it is not feasible to launch these attacks at scale. To address these problems, we develop our framework by leveraging diffusion models (DMs) without any other annotations. Recently, DMs have drawn significant attention in the image generation area with higher fidelity [15, 31] over GANs on image synthesis [8]. Also, the latent space in a DM intrinsically contains semantic information. Similar to other generative models, DMs also provide a latent space, but because it is in the same dimensions as input and output, attack methods can easily map the features from the latent space to the generated image.

With exploiting DMs, in this paper, we propose the **S**emantic **T**ransformation (ST) approach, which requires gradient information from the target classifier in the white box setting or leverages a surrogate model in the black box setting to generate minimal semantic changes

on the original inputs. Also, we provide another variant, the **L**atent **M**asking (LM) approach, that can transplant auto-selected features by different masking methods from a target image. Some visual examples are shown in Figure 1. We summarize our main contributions as follows:

- We propose the **S**emantic **T**ransformation (ST) approach, either white-box or black-box, a generalized way to generate semantic adversarial attacks via fine-tuning a latent space and/or a diffusion model.
- We propose the **L**atent **M**asking (LM) approach to fast generate semantic adversarial attacks. Two interpretation methods (Grad-CAM or saliency map) are used to mask a latent space by transplanting another target image.
- Our framework with ST and LM approaches is the first systematic way to generate semantic adversarial attacks leveraging diffusion models.

Related Work Unlike ℓ_p-norm based attacks with pixel-wise perturbations, [4, 9, 37] proposed unrestricted adversarial attacks with other techniques such as spatial transformations. At the same time, [16] generate adversarial examples by manipulating the colors of a clean image in the HSV color space, such as randomly shifting the hue and saturation components. Concurrently, [30] proposed a non ℓ_p-norm based attack, generated with conditional generative models. Even the concept of semantic attack was not mentioned in this work, since the attacks are generated with semantic information, it proposed the prototype of semantic attack: a kind of unrestricted perturbations with ℓ_p-norm by manipulating semantic information while keeping perceptual similarity realism. As aforementioned, changing color or texture as in [2, 16], and manipulating attributes by generative models are both considered semantic attacks, and many studies [2, 17, 24, 27, 35, 39] built on top of them. Most of such attacks are visible; such as [24], where it proposed a human face-based semantic attack algorithm by slightly changing the attributes (e.g., with the additional annotation. Though some [22, 35, 39] of them are not visible, it is still a semantic attack, as long as it uses semantic information to modify a clean image. For example, [35] crafted invisible semantic adversarial perturbations by manipulating semantic information with Perceptual Similarity (PS), and [22] generate targeted unrestricted adversarial attacks with a decision-based attacking algorithm in a latent space of an adversarial generative model (GAN). However, when generating semantic adversarial attacks with generative models, previous studies either rely on a dataset with attribute annotations [24] or require thousands of queries [22]. Traditional image generation techniques usually require adversarial generative networks (GANs); recently diffusion models (DMs) [10, 15, 31] achieved superior image quality to GANs on image synthesis [8].

2 Methodology

2.1 Preliminary: Diffusion Models

Diffusion models [15, 31] include a diffusion process (forward process) and a sampling process (reverse process). The diffusion process transforms data to a simple noise distribution while the sampling process reverses this process. Either of the two steps is a Markov chain and consists of a sequence of steps, where every step can be approximated to a Gaussian

distribution. The diffusion and sampling processes can be defined as follows:

$$q_\theta(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q_\theta(\mathbf{x}_t|\mathbf{x}_{t-1}),\ q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}), \tag{1}$$

$$p_\theta(\mathbf{x}_{0:T}) = p(\mathbf{x}_T)\prod_{t=1}^{T} p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t),\ p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t,t), \Sigma_\theta(\mathbf{x}_t,t)), \tag{2}$$

where $\mathbf{x}_t$ is the latent space for $t = 1, \cdots, T$. The latent space $\mathbf{x}_t$ in diffusion process can be expressed as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\mathbf{w},\ \mathbf{w} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{3}$$

where $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{s=1}^{t} \alpha_s$. In the forward process, the parameter $\{\beta_t\}_{t=0}^{T}$ can be either a learnable parameter by reparameterization [15] or fixed constant. In the reverse process, $\mu_\theta(\mathbf{x}_t,t)$ can be expressed as $\mu_\theta(\mathbf{x}_t,t) = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{x}_t,t))$, where $\varepsilon_\theta(\mathbf{x}_t,t)$ is a noise approximation model, which predicts ε from a latent space $\mathbf{x}_t$ and a time step t. This model can be trained by minimizing the following loss function over model parameters θ as $\mathcal{L}_{simple}(\theta) = \mathbb{E}_{x_0\sim q(\mathbf{x}_0), \mathbf{w}\sim\mathcal{N}(\mathbf{0},\mathbf{I}), t}\|\mathbf{w} - \varepsilon_\theta(\mathbf{x}_t,t)\|_2^2$. After the model is trained, data can be sampled with the sampling process as $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}}(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\varepsilon_\theta(\mathbf{x}_t,t)) + \sigma_t\mathbf{z}$, where $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$. Meantime, [29] proposed a non-Markovian diffusion process that leverages the same DDPM in the forward process as they share the same forward marginals, and it has a distinct sampling process:

$$\mathbf{x}_{t-1} = \sqrt{\bar{\alpha}_{t-1}}\mathbf{f}_\theta(\mathbf{x}_t,t) + \sqrt{1-\bar{a}_{t-1}-\sigma_t^2}\varepsilon_\theta(\mathbf{x}_t,t) + \sigma_t^2\mathbf{z}, \tag{4}$$

where $\mathbf{z} \sim \mathcal{N}(\mathbf{0},\mathbf{I})$ and $\mathbf{f}_\theta(\mathbf{x}_t,t)$ is a estimation of $\mathbf{x}_0$ at t given $\mathbf{x}_t$ and $\varepsilon_\theta(\mathbf{x}_t,t)$, as $\mathbf{f}_\theta(\mathbf{x}_t,t) = \frac{\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\varepsilon_\theta(\mathbf{x}_t,t)}{\sqrt{\bar{\alpha}_t}}$.

2.2 The Semantic Transformation (ST) approach

One crude way of using a diffusion model to perform a semantic adversarial attack is to directly manipulate the latent space through an attack loss as shown in Figure 2. Given a clean image $\mathbf{x}_0$, we use Eq.(3) to obtain its latent space $\mathbf{x}_T$ with a diffusion process as in Eq. (1). The generated semantic adversarial image is denoted as $\hat{\mathbf{x}}_0(\hat{\theta}, \hat{\mathbf{x}}_T)$ with the fine-tuned diffusion model parameter $\hat{\theta}$ and the fine-tuned latent space $\hat{\mathbf{x}}_T$. We fine-tune the latent space and/or the diffusion model to transform semantic information during the fine-tuning process until the generated image $\hat{\mathbf{x}}_0(\hat{\theta}, \hat{\mathbf{x}}_T)$ mislead the classifier. After the fine-tuning process, we evaluate our attacks with a sampling process as in Eq. (2). For diffusion and sampling processes, we use DDIM [29] as it is a deterministic process. The fine-tuning process is performed until the target classifier can be fooled by generated image $\hat{\mathbf{x}}_0(\hat{\theta}, \hat{\mathbf{x}}_T)$. As shown in [20], we believe manipulating a latent space $\mathbf{x}_T$ at step T affects the generated image after sampling, since $\mathbf{x}_T$ contains semantic information of the original image $\mathbf{x}_0$. The algorithm is in the supplementary material.

Loss Function in the Finetuning Process. There are many perceptual metrics for assessing the perceptual similarity between two images, such as Peak signal to noise ratio (PSNR) and structural index similarity (SSIM); however, these metrics fail to capture the nuances of human perceptions [44]. Evaluated on BAPPS dataset, [44] proposed the Learned Perceptual Image Patch Similarity (LPIPS) metric, which recognizes similarities well even with various distortions for a pair of images. Hence, we minimize the LPIPS metric in our loss function, to

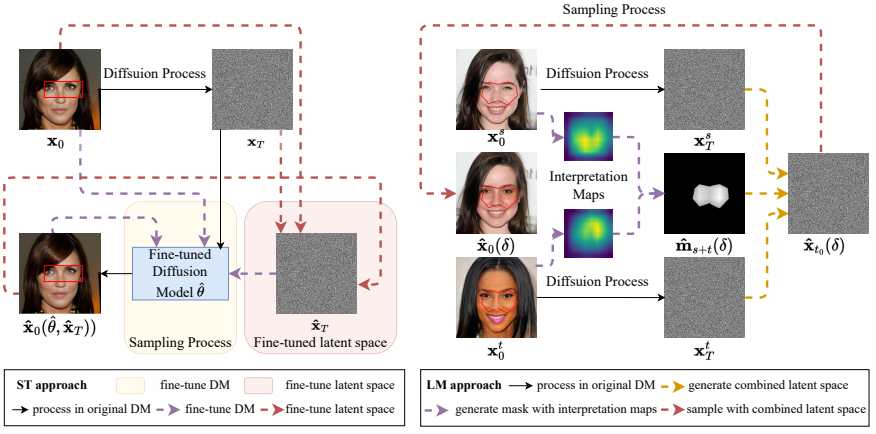

Figure 2: Pipelines of generating semantic adversarial images via a fine-tuning process by ST and LM approach.

maximize the perceptual similarity between the original image and the generated semantic adversarial image. Inspired by the TRADES loss [42], we maximize the KL divergence between the prediction logits on the original image and the prediction logits on the semantic adversarial image. Our loss function in the fine-tuning process is defined as:

$$\mathcal{L}_{ST} = \min_{\hat{\theta},\hat{\mathbf{x}}_T} \lambda D_{LPIPS}(\mathbf{x}_0, \hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T)) - D_{KL}(f(\mathbf{x}_0), f(\hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T))), \tag{5}$$

where D_{LPIPS} captures the perceptual similarities of the original image $\mathbf{x}_0$ and the generated adversarial image $\hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T))$, and minimizing this loss term keeps perceptual features of the pair of images remain same during fine-tuning. By contrast, maximizing D_{KL} encourages the generated image $\hat{\mathbf{x}}_0(\hat{\theta},\hat{\mathbf{x}}_T))$ to enlarge the logits distance with $\mathbf{x}_0$ with respect to a classifier f, either a known classifier or a random pre-trained classifier. Thus, there is a trade-off between these two terms: while maintaining the global perceptual similarities, we expect to change the local attributes misleading the classifier f. The relative strengths of two loss terms D_{LPIPS} and D_{KL} can be adjusted by the scalar λ.

Nuance between White-box and Black-box Attacks. The difference between a white-box and black-box attack is whether the malicious knows the target model parameters of f. In Eq. (5), when calculating D_{KL}, the prediction logits from the target classifier are used for a white-box attack, and the outputs from a pre-trained InceptionV3 model [33] are used for a black-box attack.

2.3 The Latent Masking (LM) Approach

Fine-tuning a latent space or diffusion model requires calculating the gradients on the latent space or diffusion model parameters, resulting in huge computation expenses. In this subsection, we introduce an alternative approach to modify the generated image, that is, by masking the latent space with feature significance from a target image. The mask area is the most significant in a feature map and is intended to contain important, semantically meaningful features with respect to the target classifier. We transplant the masked area as in Figure 2. Let $\mathbf{m}$ be an interpretation map, it can be calculated as $\mathbf{m} = g(\mathbf{x}_0, \mathbf{y})$, where $\mathbf{x}_0$ is a clean image, and $\mathbf{y}$ is its label. We use Grad-CAM [11, 26] and saliency maps [28, 32] for interpretation maps in this paper.

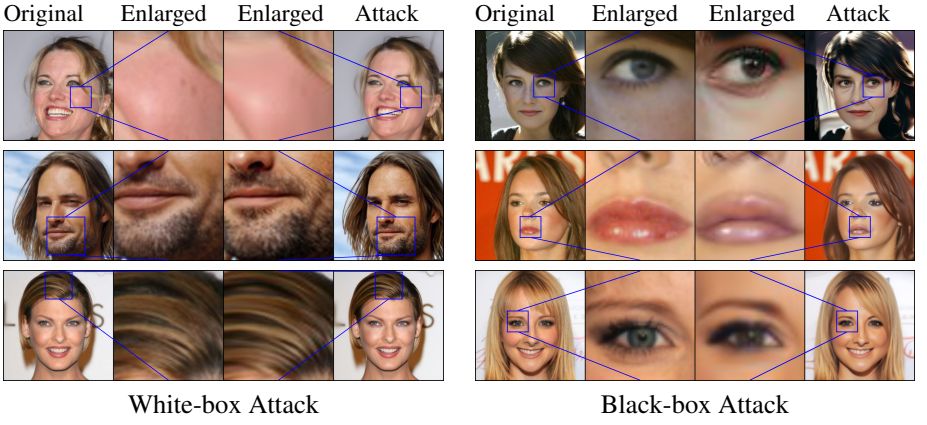

Figure 3: ST approach under different fine-tuning. First row: only fine-tuning a latent space. Second row: only fine-tuning a diffusion model. Third row: fine-tuning both a latent space and a diffusion model. First column: original image. Second column: enlarged local area from the original image. Third column: enlarged local area from the adversarial image. Fourth column: adversarial image.

Transplanting Features with Mask. We denote a pair of source and target images as $\mathbf{x}_0^s$ and $\mathbf{x}_0^t$, and their latent spaces as $\mathbf{x}_T^s$ and $\mathbf{x}_T^t$ respectively. Here, $\mathbf{x}_0^s$ is used as the victim image, and most features in $\mathbf{x}_0^s$ are kept; $\mathbf{x}_0^t$ is used as the target image, and a small portion of masked features in $\mathbf{x}_0^t$ are transplanted to $\mathbf{x}_0^s$ when generating a semantic adversarial attack. We use Grad-CAM or saliency map to generate the mask. The significance maps are calculated on $\mathbf{x}_0^s$ and $\mathbf{x}_0^t$, denoted as $\mathbf{m}_s$ and $\mathbf{m}_t$. We have three strategies to generate a mask $\hat{\mathbf{m}}$ and we denoted them as $\hat{\mathbf{m}}_s(\delta)$, $\hat{\mathbf{m}}_t(\delta)$ and $\hat{\mathbf{m}}_{s+t}(\delta)$, respectively:

$$\hat{\mathbf{m}}_s(\delta) = TopK(|\mathbf{m}_s|, \delta), \quad \hat{\mathbf{m}}_t(\delta) = TopK(|\mathbf{m}_t|, \delta), \quad \hat{\mathbf{m}}_{s+t}(\delta) = TopK(|\mathbf{m}_s| + |\mathbf{m}_t|, \delta), \tag{6}$$

where δ is a percentage threshold of $TopK$ function, ranging from 0 to 99, and $TopK$ function would only keep a given input with the $\delta\%$ largest and set other elements in the mask as zero. We design a heuristic to control the decremental speed of δ:

$$\delta = \delta - \max(\gamma \frac{z_y - \max_{i \neq y} z_i}{z_y}, 1), \tag{7}$$

where z_y is the target class confidence logit, $\max_{i \neq y} z_i$ is the second highest confidence logit, and γ is a constant. With the mask $\hat{\mathbf{m}}_{t_0}(\delta)$ created, the original latent space is modified as:

$$\hat{\mathbf{x}}_T(\delta) = (1 - \hat{\mathbf{m}}(\delta)) * \mathbf{x}_T^s + \hat{\mathbf{m}}(\delta) * \mathbf{x}_T^t \tag{8}$$

With modified latent space $\hat{\mathbf{x}}_T(\delta)$, we generated a semantic adversarial image without fine-tuning a latent space or diffusion model following Eq.(4). The whole process of the LM approach is shown in Figure 2. The algorithm is in the supplementary material.

3 Experiments

3.1 Datasets

We evaluate our white-box and black-box attacks on two tasks: human facial identity recognition and animal category recognition. For all experiments, we use 500 images with the size of 256×256.

Human Facial Identity Recognition and Gender Classification. We use Celeb-HQ Facial Identity Recognition Dataset, which is a subset of the CelebAMask-HQ dataset [21], adopted from [22]. The original CelebAMask-HQ dataset contains 30,000 face images at 512 x 512 resolution. It has 6,217 unique identities. For the target classifier, we use a subset of it as in [22], which contains 307 unique identities, 4,263 images for training, and 1,215 images for testing. A ResNet18 [13] classifier is trained for 30 epochs with 89.05% accuracy on this dataset. For the diffusion model, we used the model pretrained on CelebA-HQ [18] dataset, which contains 30,000 images at 1024 x 1024 resolution. The evaluation is in Section 3.4 and the supplementary material. Besides facial identity recognition, we also adopt the Celeb-HQ Face Gender Recognition Dataset from [22], which contains 11,057 male and 18,943 female images, and the evaluation is in the supplementary material.

Animal Category Recognition. We use the AFHQ [6] dataset, a dataset of animal faces consisting of 15,000 images at 512 x 512 resolution. The dataset contains three categories, cat, dog, and wildlife, and the evaluation is in the supplementary material.

3.2 Attack Details

Fine-tuning Process and Evaluation. In Eq. (5), we set λ to 1. We denote the number of iteration steps in a diffusion process, a fine-tuning process, and a sampling process as s_{df}, s_{ft}, and s_{sp}, which are set to $40, 15$ and 40 respectively. For s_{ft}, according to [19], even 6 steps would satisfy the fine-tuning purpose. In our experiments, we set it to 15 due to the VRAM limitation on our GPUs. However, the semantic adversarial images would demonstrate a smoother modification with higher image quality by increasing the fine-tuning step. To ensure the adversary of our generated image, we initially verify the target classifier output throughout the fine-tuning process, and if the label remains unchanged, we will conduct additional 15 steps of the fine-tuning process based on the last run iteratively until the attack is successful. The procedure of sampling will then be carried out to improve the image quality.

Constructing Mask and Evaluation. Every pair of source and target images, $\mathbf{x}_0^s$ and $\mathbf{x}_0^t$, is randomly sampled as long as they have different class labels and can be classified correctly by the target classifier. When applying Grad-CAM or saliency map to an image, we combine $\mathbf{m}_0$ and $\mathbf{m}_1$ from the RGB channels into one channel and then filter by *TopK* in order to better observe how features are transplanted from the target image to the original clean image. For the saliency map, we adopt SimpleFullGrad from [32]. For the Grad-CAM [25], we directly use the original implementation.

After integrating the generated mask from Eq. (6) and Eq. (8), the diffusion process does not always generate adversarial examples, and we need to decrement hyper-parameter δ in every itera-

Setting	strategy	ASR (%)↑	FID↓	KID↓	average query↓	average time (s)↓
clean images	-	-	30.67	0.000	-	-
LatentHSJA	-	100.0	83.52	0.046	1000†	45.87
AttAttack	-	71.80	48.92	0.018	146.82	49.71
ST approach						
fine-tune latent space	white-box	**100.0**	37.93	0.014	7.72	37.10
	black-box	59.18	114.99	0.098	43.15	206.13
fine-tune diffusion model	white-box	99.2	**36.61**	**0.006**	4.98	**30.78**
	black-box	**100.0**	96.88	0.068	11.73	66.57
fine-tune both	white-box	99.4	36.66	**0.006**	**4.96**	**30.78**
	black-box	**100.0**	94.36	0.066	11.672	64.97
LM approach						
GradCAM	$\hat{\mathbf{m}}_s(\delta)$	98.8	65.84	0.015	15.33	20.96
	$\hat{\mathbf{m}}_t(\delta)$	99.2	**64.38**	**0.014**	15.21	**18.89**
	$\hat{\mathbf{m}}_{s+t}(\delta)$	99.0	65.47	**0.014**	**14.65**	20.81
SimpleFullGrad	$\hat{\mathbf{m}}_s(\delta)$	99.6	67.10	0.016	16.17	24.03
	$\hat{\mathbf{m}}_t(\delta)$	99.6	65.21	0.016	15.32	27.48
	$\hat{\mathbf{m}}_{s+t}(\delta)$	**99.8**	65.67	0.015	14.73	23.77

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

Table 1: Performance of our framework with the ST and the LM approach on CelebA-HQ dataset compared with other two baselines.

tion with Eq. (7).

In this approach, semantic adversarial images are generated using only the mask and a target image without the fine-tuning process. We set both s_{df} and s_{sp} to 40. During each iteration, with decremental δ, we check if the generated image from the sampling process is an adversarial example against the targeted classifier f, and we stop the attack when an adversarial example is generated.

3.3 Evaluation Metrics and Benchmarks

We quantify the attack success rate (ASR), Fréchet Inception Distance (FID) [12] and Kernel Inception Distance (KID) [3] in the fidelity of semantic adversarial attacks. FID measures the Fréchet distance between two data distributions, and KID measures the dissimilarity between two distributions. For both measurements, lower is better. In addition, for the ST method, we measure the average number of fine-tuning iterations, denoted as δ_{avg}; for the LM method, we measure the average threshold for successful semantic adversarial attacks, denoted as η_{avg}. We also measure the average elapsed time for generating a semantic adversarial attack to evaluate efficiency.

3.4 Results

The results and analysis focus on CelebA-HQ identity dataset, for which the results with the ST and the LM approach of our framework are shown in Table 1. For baselines, we use two groups of clean images to calculate FID and KID, each group with 500 images. For comparison, we use recent semantic adversarial attacks LatentHSJA [22] and AttAttack [17] as benchmarks. For LatentHSJA, we run experiments with default fixed 1,000 queries. The number of queries is preset by the user, and the default is 20,000. We use 1,000 as a trade-off between quality and efficiency. For AttAttack, we run several benchmarks of it and choose AttGAN to perturb the Age attribute, since this setting balances quality and ASR. In addition, we use the same target model in AttAttack as in our attacks. Both baselines are evaluated on Celeb-HQ Facial Identity Recognition Dataset as mentioned in Section 3.1. For all experiments, we focus on untargeted attacks, that is, as long as the predicted label is different from the original one, the attack will be counted as a success.

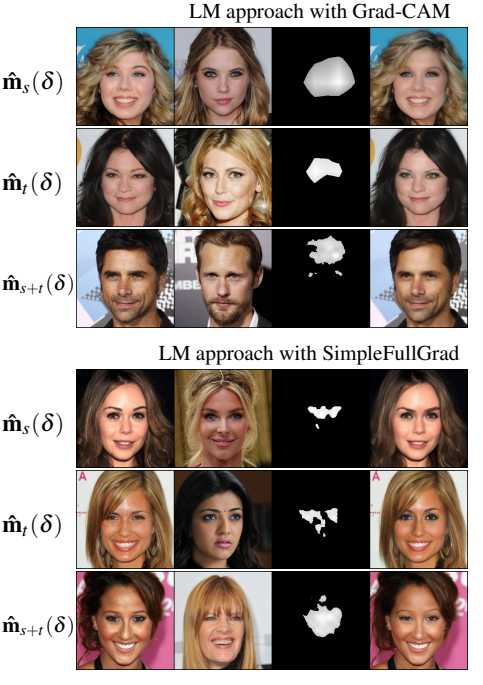

Figure 4: LM approach with three strategies $\hat{\mathbf{m}}_s(\delta)$, $\hat{\mathbf{m}}_t(\delta)$ and $\hat{\mathbf{m}}_{s+t}(\delta)$. From left to right: source images, target images, constructed masks with given strategies and generated semantic adversarial images.

Analysis for the ST Approach. Table 1 represents the performance of our framework on CelebA-HQ dataset. The top half shows the statistics of our framework with the ST approach under white-box and black-box settings. Our attack achieves almost 100% ASR in all cases. For the ST approach, from FID and KID, we can clearly observe that our framework under white-box settings obtains higher-quality images than black-box settings. Under black-box settings, our framework creates more deformation than white-box settings and makes the generated images unrealistic and dissimilar to the original image. Visual examples generated with the ST approach are in Figure 3.

From these generated images with the ST approach, we find that our framework under white-box settings tends to achieve a minimal amount of modification concentrated in a small area of the original image. Under black-box settings, our framework tends to randomly modify details on a relatively large scale. This can be explained by the classifier f in the KL divergence loss term D_{KL} of Eq. (5). Under white-box settings, f is our targeted classifier in ResNet18, whereas under black-box settings, f is a pretrained InceptionV3 network on ImageNet [7]. Thus, f could not efficiently capture the most important area with respect to a face identity under black-box settings. In addition, we find that fine-tuning both a latent space and a diffusion model achieves the best balance between quality and efficiency.

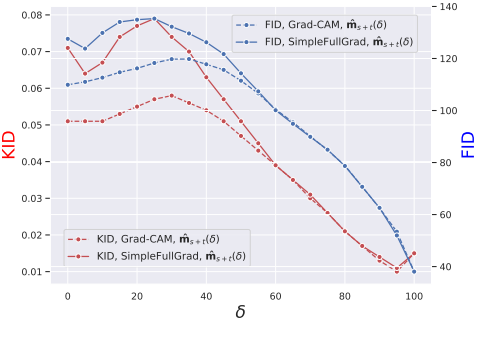

Figure 5: The relationship between FID/KID and δ in the LM approach of our framework.

Analysis for the LM Approach. We found that constructing masks using GradCAM or SimpleFullGrad saliency maps yields similar results like nearly 100% ASR and similar FIDs and KIDs. In most human faces, the features related to a face's identity usually come to be in a similar area on an image (e.g. nose, eyes, chin, and forehead). Hence, we can transplant features by directly applying masks to source and target images.

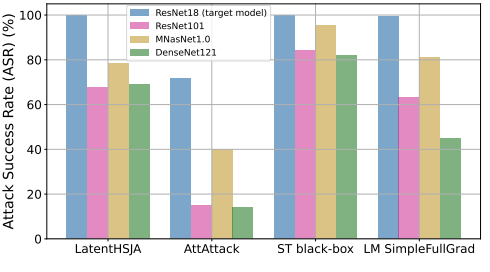

Figure 6: Transfer attack results on LatentHSJA, AttAttack, our ST and LM approach.

Visual examples are shown in Figure 4. Note that the mask may not exactly correspond to the area being modified since the generative process is performed via a diffusion model, the Gaussian noise added during the sampling process cannot precisely only modify a specified area without special techniques such as image inpainting. However, the most significant modifications occur in the mask areas overall. We also investigate the sensitivity of masking threshold δ, as shown in Figure 5, and we find that the LM approach with Grad-CAM achieves better fidelity and quality than SimpleFullGrad in the same δ, This implies that Grad-CAM has a better ability to expose attackable areas than SimpleFullGrad.

Comparison with Benchmarks. Visual examples for comparison are shown in Figure 7, and more of them can be found in the supplementary material. In terms of FID and KID, our framework with the ST approach under white-box settings achieves comparable perfor-

mance. For average elapsed time per generated image, our framework with the LM approach shows the best performance compared to other methods. LatentHSJA starts from the target image and iteratively tries to transplant pixels in latent space, with a fixed number of queries, but with a low similarity compared with the original image. AttAttack greatly improves performance by utilizing manual attribute annotations, yet it is challenging to generalize to other datasets. In addition, we evaluate the transferability in Figure 6. All methods targeted their attack on Resnet18 (except for ST Black-box) and generated 500 adversarial images. We test these on three additional classifiers, ResNet101, MNasNet1.0, and DenseNet121 to calculate the ASR. Our ST black-box approach performs the best among the others as expected, since it does not require exact information from the target model. Compared with the ST approach, the generated adversarial attacks should be harder to denoise by diffusion-based purification algorithms such as [23, 34, 36]. We further evaluate the robustness of our semantic attack under natural perturbations, such as JPEG, Gaussian Blur, Defocus Blur, and Brightness transformation in the supplementary material. The results demonstrate that our semantic adversarial perturbations are still preserved after natural perturbations.

Figure 7: Comparison of semantic adversarial attacks between ours and others. From left to right: original image, LatentHSJA, AttAttack, our ST black-box approach, and LM with Grad-CAM approach.

4 Conclusion

In this paper, we first proposed a framework for semantic adversarial attacks by leveraging Diffusion Models with the ST approach and LM approach. The ST approach manipulates the latent space of a benign image or the parameters of a diffusion model via fine-tuning, whereas the LM approach manipulates the latent space via masking of significance maps in a more direct way. In our empirical study, the proposed framework achieves excellent performance for both approaches, under different settings. In total, our framework shows great generalizability, efficiency, and transferability compared to other baselines and exposes a novel usage of Diffusion Models in the semantic adversarial attack domain. However, there are limitations to our framework. For the ST approach, the quality of adversarial images under the black-box setting is not as good as the white-box setting; for the LM approach, it cannot precisely control the masked area to be modified compared to a clean image.

References

[1] David Bau, Jun-Yan Zhu, Hendrik Strobelt, Agata Lapedriza, Bolei Zhou, and Antonio Torralba. Understanding the role of individual units in a deep neural network. *Proceedings of the National Academy of Sciences*, 117(48):30071–30078, 2020.

[2] Anand Bhattad, Min Jin Chong, Kaizhao Liang, Bo Li, and David A Forsyth. Unrestricted adversarial examples via semantic manipulation. *arXiv preprint arXiv:1904.06347*, 2019.

[3] Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans. *arXiv preprint arXiv:1801.01401*, 2018.

[4] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018.

[5] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *Security and Privacy (SP), 2017 IEEE Symposium on*, pages 39–57. IEEE, 2017.

[6] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. Stargan v2: Diverse image synthesis for multiple domains. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021.

[9] Yinpeng Dong, Fangzhou Liao, Tianyu Pang, Hang Su, Jun Zhu, Xiaolin Hu, and Jianguo Li. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9185–9193, 2018.

[10] Jinhao Duan, Fei Kong, Shiqi Wang, Xiaoshuang Shi, and Kaidi Xu. Are diffusion models vulnerable to membership inference attacks? *International Conference on Machine Learning*, 2023.

[11] Jacob Gildenblat and contributors. Pytorch library for cam methods. https://github.com/jacobgil/pytorch-grad-cam, 2021.

[12] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[14] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

[15] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

[16] Hossein Hosseini and Radha Poovendran. Semantic adversarial examples. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1614–1619, 2018.

[17] 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.

[18] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196*, 2017.

[19] 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.

[20] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space. *arXiv preprint arXiv:2210.10960*, 2022.

[21] Cheng-Han Lee, Ziwei Liu, Lingyun Wu, and Ping Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020.

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

[23] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460*, 2022.

[24] Haonan Qiu, Chaowei Xiao, Lei Yang, Xinchen Yan, Honglak Lee, and Bo Li. Semanticadv: Generating adversarial examples via attribute-conditioned image editing. In *European Conference on Computer Vision*, pages 19–37. Springer, 2020.

[25] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017.

[26] RR Selvaraju, M Cogswell, A Das, R Vedantam, D Parikh, and D Batra. Grad-cam: visual explanations from deep networks via gradient-based localization. 2016. *arXiv preprint arXiv:1610.02391*, 2016.

[27] Ali Shahin Shamsabadi, Ricardo Sanchez-Matilla, and Andrea Cavallaro. Colorfool: Semantic adversarial colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1151–1160, 2020.

[28] Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Deep inside convolutional networks: Visualising image classification models and saliency maps. *arXiv preprint arXiv:1312.6034*, 2013.

[29] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2020.

[30] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. *Advances in Neural Information Processing Systems*, 31, 2018.

[31] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2020.

[32] Suraj Srinivas and François Fleuret. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32, 2019.

[33] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jonathon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. 2015. *arXiv preprint arXiv:1512.00567*, 2015.

[34] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022.

[35] Yajie Wang, Shangbo Wu, Wenyi Jiang, Shengang Hao, Yu-an Tan, and Quanxin Zhang. Demiguise attack: Crafting invisible semantic adversarial perturbations with perceptual similarity. *arXiv preprint arXiv:2107.01396*, 2021.

[36] Quanlin Wu, Hang Ye, and Yuntian Gu. Guided diffusion model for adversarial purification from random noise. *arXiv preprint arXiv:2206.10875*, 2022.

[37] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. *arXiv preprint arXiv:1801.02612*, 2018.

[38] Kaidi Xu, Gaoyuan Zhang, Sijia Liu, Quanfu Fan, Mengshu Sun, Hongge Chen, Pin-Yu Chen, Yanzhi Wang, and Xue Lin. Adversarial t-shirt! evading person detectors in a physical world. In *European Conference on Computer Vision (ECCV)*, pages 665–681. Springer, 2020.

[39] Qiuling Xu, Guanhong Tao, Siyuan Cheng, and Xiangyu Zhang. Towards feature space adversarial attack. *arXiv preprint arXiv:2004.12385*, 2020.

[40] Chenxi Yuan and Mohsen Moghaddam. Attribute-aware generative design with generative adversarial networks. *IEEE Access*, 8:190710–190721, 2020.

[41] Chenxi Yuan, Jinhao Duan, Nicholas J Tustison, Kaidi Xu, Rebecca A Hubbard, and Kristin A Linn. Remind: Recovery of missing neuroimaging using diffusion models with application to alzheimer's disease. *medRxiv*, pages 2023–08, 2023.

[42] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

[43] Huan Zhang, Shiqi Wang, Kaidi Xu, Yihan Wang, Suman Jana, Cho-Jui Hsieh, and Zico Kolter. A branch and bound framework for stronger adversarial attacks of relu networks. In *International Conference on Machine Learning*, pages 26591–26604. PMLR, 2022.

[44] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[45] Pu Zhao, Kaidi Xu, Sijia Liu, Yanzhi Wang, and Xue Lin. Admm attack: an enhanced adversarial attack for deep neural networks with undetectable distortions. In *Proceedings of the 24th Asia and South Pacific Design Automation Conference*, pages 499–505, 2019.