AdvDiffuser: Natural Adversarial Example Synthesis with Diffusion Models

Xinquan Chen* [a,b], Xitong Gao* [a], Juanjuan Zhao [a], Kejiang Ye [a], Cheng-Zhong Xu [c]
[a] Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences, China.
[b] University of Chinese Academy of Sciences, China.
[c] State Key Lab of IOTSC, University of Macau, Macau S.A.R., China.
{xq.chen,xt.gao,jj.zhao,kj.ye}@siat.ac.cn czxu@um.edu.mo

Abstract

Previous work on adversarial examples typically involves a fixed norm perturbation budget, which fails to capture the way humans perceive perturbations. Recent work has shifted towards natural unrestricted adversarial examples (UAEs) that breaks ℓ_p perturbation bounds but nonetheless remain semantically plausible. Current methods use GAN or VAE to generate UAEs by perturbing latent codes. However, this leads to loss of high-level information, resulting in low-quality and unnatural UAEs. In light of this, we propose AdvDiffuser, a new method for synthesizing natural UAEs using diffusion models. It can generate UAEs from scratch or conditionally based on reference images. To generate natural UAEs, we perturb predicted images to steer their latent code towards the adversarial sample space of a particular classifier. We also propose adversarial inpainting based on class activation mapping to retain the salient regions of the image while perturbing less important areas. On CIFAR-10, CelebA and ImageNet, we demonstrate that it can defeat the most robust models on the RobustBench leaderboard with near 100% success rates. Furthermore, The synthesized UAEs are not only more natural but also stronger compared to the current state-of-the-art attacks. Specifically, compared with GA-attack, the UAEs generated with AdvDiffuser exhibit 6× smaller LPIPS perturbations, $2 \sim 3\times$ smaller FID scores and 0.28 higher in SSIM metrics, making them perceptually stealthier. Finally, adversarial training with AdvDiffuser further improves the model robustness against attacks with unseen threat models.[1]

1. Introduction

Deep neural networks (DNNs) have achieved unprecedented success in various visual recognition tasks. Despite their remarkable success, DNNs are susceptible to *adversarial examples* [23], *i.e.* DNN predictions can be fooled by adding a tiny and difficult to perceive perturbation to a natural image. Furthermore, deep models face greater threats in real-world scenarios from unrestricted adversarial examples (UAEs) [2]. UAEs can make extensive changes to images without significantly affecting human perception of their meanings and faithfulness, and have thus emerged as a prominent direction in the study of adversarial attacks over the past few years.

Gradient-based unrestricted adversarial attacks perturbs original images within predefined perturbation bounds. Geometry-aware attacks [20] uses proxy models to minimize the ℓ_p budget required, and it won the 1st place in a CVPR Competition on unrestricted adversarial attacks [4]. On the other hand, perceptual attacks [19, 51] optimize perturbations using bounds on perceptual distances, such as LPIPS [49] and structural similarity [42]. Others consider image recolorization [36, 37]. However, selecting proxy models and distance metrics require subjective prior knowledge to generate adversarial examples that appear realistic.

Generative models such as generative adversarial networks (GANs) have the ability to learn and sample from the data distribution effectively. This is why [38, 50] use them to generate adversarial examples. These approaches search for perturbations in the latent space that can cause the targeted model to misclassify the images after decoding, in order to find adversarial examples. Nevertheless, perturbing latent codes alters the high-level semantics of generated images in a way that is perceptually salient to humans [17]. Such perturbations can introduce ambiguity in certain image attributes, and visibly distort the original concept, resulting in UAEs that are often semantically vague and of poor quality. These UAEs could thus be perceptually very different from the original examples.

In order to address these issues, we propose AdvDiffuser, a novel generative unrestricted adversarial attack based on diffusion models [13]. Diffusion models draw inspiration from non-equilibrium thermodynamics, which define a

[*] Equal contribution. Correspondence to Xitong Gao.

[1] AdvDiffuser is open source and available at https://github.com/lafeat/advdiffuser.

Markov process of noise-adding image diffusion steps, and then learns to reverse the diffusion process to generate data samples from noisy images. This enables trained diffusion models to sample the data distribution with high fidelity and diversity. In Section 3.1, we utilize and modify the backward denoising process of pre-trained diffusion models, and inject small adversarial perturbations that can attack the defending model successfully. Diffusion models are trained with a denoising objective, and therefore, they can effectively remove conspicuous adversarial noise while retaining the ability to attack, resulting in naturally appearing UAEs. To achieve more realistic outcomes, we introduce adversarial inpainting, which leverages masks derived from gradient-based class activation mapping (GradCAM) [35]. It tunes the denoising strength of each pixel based on object saliency, ensuring that regions containing important objects undergo smaller modifications. As AdvDiffuser perturbs images at the pixel level, it produces perceptual perturbations that are considerably smaller when compared to those generated by GAN-based methods. The final UAEs produced by our method are therefore more natural and imperceptible than those synthesized by either gradient- or GAN-based approaches. In addition to its image-conditioned attacks, AdvDiffuser offers another advantage over other unrestricted adversarial attacks as it has the ability to craft an infinite number of synthetic yet natural adversarial examples. This can potentially provide more comprehensive robustness training and evaluation for future defense techniques.

We summarize our contribution as follows:

- To the best of our knowledge, our work is the first to investigate natural adversarial example synthesis with diffusion models. Along with its image-conditioned attack ability, and it is also the first that can generate an infinite number of synthetic yet natural adversarial examples.
- We propose adversarial inpainting to introduce CAM-based sample conditioning, resulting in diverse and high-quality outputs while preserving the semantics of the reference images.
- AdvDiffuser can successfully deceive the top-ranked robust models in RobustBench [6] with high success rates (close to 100%). The generated examples closely resemble the original distribution. Our perturbations are both more effective and less perceptible, with better LPIPS, FID and SSIM distance metrics than the current state-of-the-art unrestricted adversarial attacks.

2. Preliminaries & Related Work

2.1. Adversarial Examples

A successful adversarial attack [39] occurs when an attacker adds a small but potentially imperceptible perturbation to the original image, in order to mislead the targeted model into giving incorrect outputs. Since the discovery of adversarial examples, such attacks raised major security concerns [40, 31] in computer vision and machine learning communities. On the other hand, these techniques have also been utilized to improve transfer learning [41], deep learning interpretability [32], safeguarding privacy [27, 26], federated learning [45], among other applications. Formally, let us assume a defending classifier $f \colon \mathcal{I} \to \mathbb{R}^K$, which takes an input image $\mathbf{x}$ from a test dataset $\mathcal{D}_\text{test} \subset \mathcal{I} \triangleq \mathbb{R}^{C\times H\times W}$ and evaluates the correct classification result $y = \operatorname{argmax} f(\mathbf{x}) \in \mathcal{C}$ for $\mathbf{x}$, then the attacker searches for the adversarial example $\hat{\mathbf{x}}$ that satisfies:

$$\operatorname{argmax} f(\hat{\mathbf{x}}) \neq y \quad \text{and} \quad \operatorname{dist}(\mathbf{x}, \hat{\mathbf{x}}) \leq \delta. \tag{1}$$

Here, the condition $\operatorname{argmax} f(\hat{\mathbf{x}}) \neq y$ indicates a successful deception of the classifier f by the adversarial example $\hat{\mathbf{x}}$, and $\operatorname{dist}(\mathbf{x}, \hat{\mathbf{x}}) \leq \delta$ places a bound on a distance metric, $\operatorname{dist}(\mathbf{x}, \hat{\mathbf{x}})$, which measures the distance between the original image $\mathbf{x}$ and the adversarial $\hat{\mathbf{x}}$.

In traditional ℓ_p-bounded attacks, we let $\operatorname{dist}(\mathbf{x}, \hat{\mathbf{x}}) = \|\mathbf{x} - \hat{\mathbf{x}}\|_p$ to bound $\hat{\mathbf{x}}$ within a small δ-ball of ℓ_p-distance away from the original image $\mathbf{x}$. Many algorithms have been proposed to find such adversarial examples, such as the fast gradient-sign method (FGSM) [9], projected gradient descent (PGD) [23], Carlini-Wagner attack [3], and other more effective variants [47, 48].

2.2. Unrestricted Adversarial Examples

Since the ℓ_p-norm distance is inadequate to capture how human perceive perturbation accurately, recent years have seen an upsurge of interest in unrestricted adversarial examples (UAEs). UAEs are images that satisfy the distribution, which humans can correctly categorize but wrongly classify by the classifier.

First line of approaches exploits prescribed image transformations which appear natural to search for UAEs. Xiao *et al.* [44] generate adversarial examples with spatial warping transformations. By switching to the LAB color space, Ali *et al.* [37] optimizes in the AB channel for adversarial examples while keeping the luminance component constant, varying the range of the perturbation in different regions.

The idea of training generative models to generate adversarial attacks has been proposed by many papers [1, 43, 15]. This approach, however, typically suffers from limited attack success rates, An alternative approach is suggested in [46, 38, 15, 50], which leverages generative models pre-trained on natural images to produce adversarial examples by perturbing the latent representation. This technique may produce lowered visual quality UAEs, which may not faithfully match the original data distribution.

The gradient-based unrestricted attacks [19, 20] searches for UAEs with distance metrics other than the traditional ℓ_p-

norm. This approach results in stronger adversarial perturbation that are still difficult to perceive by humans. Laidlaw *et al*. [19] performs projected gradient descent (PGD) [23] with LPIPS [49], which adopts deep features as a perceptual metric. Geometry-aware attacks [20] further use validation models to find the smallest perturbation bound for ℓ_p attacks. However, selecting proxy models and distance metrics require subjective prior knowledge to generate adversarial examples that appear realistic.

2.3. Diffusion Models

For the first time, Ho *et al*. [13] demonstrate that diffusion models can generate images of higher quality and diversity than GANs. Their approach defines a Markov chain comprising T forward diffusion steps, $\mathbf{x}_{1:T}$, from an original image $\mathbf{x}_0$. Each step $t \in [1:T]$ produces a latent variable $\mathbf{x}_t$ which gradually introduce Gaussian noise to an original image $\mathbf{x}_0$ with a predefined monotonically increasing variance schedule $\beta_{1:T}$. More specifically, start with a sample image from the training data set, $\mathbf{x}_0 \in \mathcal{D}_{\text{train}}$, and we can sample the latent variable $\mathbf{x}_t$ using the following forward process for $t \in [1:T]$:

$$q(\mathbf{x}_t \mid \mathbf{x}_{t-1}) = \mathcal{N}\big(\sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t \mathbf{I}\big). \tag{2}$$

As $T \to \infty$, $\mathbf{x}_T$ resembles an isotropic Gaussian distribution. Satisfying the Markov property, we can evaluate x_t directly from x_0 with the following closed form equation, where $\overline{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ and $\alpha_t = 1 - \beta_t$:

$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}\big(\sqrt{\overline{\alpha}_t}\mathbf{x}_0, (1-\overline{\alpha}_t)\mathbf{I}\big), \tag{3}$$

Subsequently, diffusion models learn the reverse process for any step t, which predicts $\mathbf{x}_{t-1}$ by removing Gaussian noise from a given latent variable $\mathbf{x}_t$:

$$\mathbf{x}_{t-1} \sim \mathcal{N}\left(\frac{1}{\sqrt{\alpha_t}}\left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\overline{\alpha}_t}}\epsilon_{\boldsymbol{\theta}}(\mathbf{x}_t, t, y)\right), \overline{\beta}_t \mathbf{I}\right), \tag{4}$$

with $\overline{\beta}_t = \frac{1-\overline{\alpha}_{t-1}}{1-\overline{\alpha}_t}\beta_t$. Additionally, $\epsilon_{\boldsymbol{\theta}}$ denotes the diffusion model with parameters $\boldsymbol{\theta}$, and a conditional model can further accept a target label y as input. These parameters can be trained to minimize the variational lower bound $\mathbb{E}_{q(\mathbf{x}_{0:T})} \log(q(\mathbf{x}_{1:T} \mid \mathbf{x}_0)/p_\theta(\mathbf{x}_{0:T}))$ which in turn optimizes the sample negative log-likelihood $-\mathbb{E}_{q(\mathbf{x}_0)} \log p_{\boldsymbol{\theta}}(\mathbf{x}_0)$.

Building on top of this, Improved DDPM [24] learns the variance schedule to enhance sample quality and increase sampling efficiency. Dhariwal *et al*. [7] enhance it further with classifier guidance, to generate class-conditioned examples. This leverages the gradient of softmax cross-entropy loss of a classifier to guide image synthesis. Inspired by this idea, Liu *et al*. [21] extend it to image- and text-based guidance, while Choi *et al*. [5] use reference images as guidance to further enable image translation, editing and inpainting applications. Ho *et al*. [14] propose to train conditional diffusion models, eliminating the need to use classifiers.

Diffusion models have numerous applications in various domains. For example, Dall-E [28] and stable diffusion [30] produce professional artistic paintings with user-specified text prompts. DiffPure [25] uses diffusion models to purify adversarial adversarial examples to make downstream vision models more robust. Furthermore, there are numerous techniques that apply diffusion models to natural language processing, signal processing, and time-series data modeling.

3. Method

Figure 1 provides a high-level overview of the AdvDiffuser algorithm. The algorithm initiates by computing the Grad-CAM [35] of the image under attack, utilizing the defending model and the ground-truth label to form a mask of the salient object. Afterward, it iteratively uses a pretrained diffusion model, to denoise the latent image $\mathbf{x}_{t-1}$. Subsequently, an ℓ_2-bounded PGD attack is performed on the image. Following this, AdvDiffuser then interpolates between the resulting attack image and a noised original image, using the precomputed mask. By repeating the T-step denoising process, it thus forms a process to add adversarial perturbations while removing unnatural components from the injected noise. As a result, we can generate adversarial examples that are semantically close to the originals, yet containing shape-based adversarial perturbations exhibiting detailed diversity.

3.1. Adversarial Guidance

We introduce *adversarial guidance*, to generate natural adversarial examples using diffusion models. This involves iteratively solving the following optimization problem:

$$\begin{aligned} &\hat{\mathbf{x}}_{t-1} = \operatorname{argmax}_{\mathbf{x}} \mathsf{L}(f(\mathbf{x}), y), \text{where} \\ &\quad \operatorname{dist}(\mathbf{x}, \tilde{\mathbf{x}}_{t-1}) \leq \varepsilon_t, \text{and} \\ &\quad \tilde{\mathbf{x}}_{t-1} \sim p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\hat{\mathbf{x}}_t). \end{aligned} \tag{5}$$

At each step, the process initiates by denoising the previously perturbed latent variable $\tilde{\mathbf{x}}_{t-1}$, and subsequently introducing adversarial perturbation that fools the defending classifier f. It thus forms a saddle-point solution which attempts to minimize the negative log-likelihood $-\log p_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_t)$ with the diffusion model, while simultaneously increasing the adversarial loss $\mathsf{L}(f(\mathbf{x}), y)$ of the defending classifier f, where $y = \operatorname{argmax} f(\mathbf{x})$ is the predicted label.

To optimize (5), we adopt the projected gradient descent (PGD) [23] attack to find an approximate solution $\mathbf{z}_I$ for a reference image $\mathbf{z}_0$, by iterating $i \in [0:I-1]$:

$$\mathbf{z}_{i+1} = \mathcal{P}_{\mathbf{z}_0,\varepsilon}(\mathbf{z}_i + \operatorname{sign}(\nabla_{\mathbf{z}_i} \mathsf{L}(f(\mathbf{z}_i, y)))), \tag{6}$$

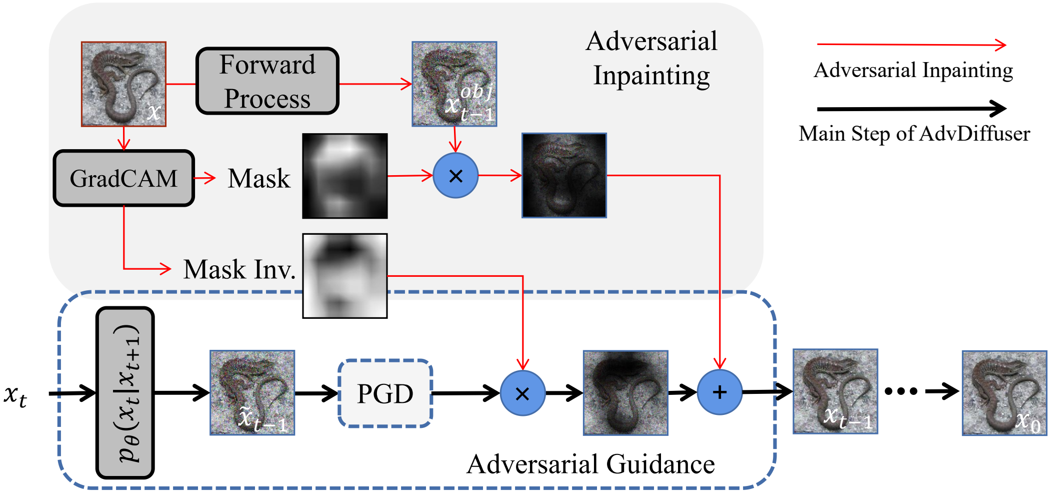


Figure 1: An overview of the AdvDiffuser algorithm for generating unrestricted adversarial examples.

and $\mathcal{P}_{\hat{\mathbf{x}}_t,\varepsilon}(\mathbf{z})$ represents the projection of $\mathbf{z}$ into ε-ball of ℓ_2-distance. We further use the normalized softmax cross-entropy (SCE) loss [47] as the maximization objective function L, instead of the conventional SCE loss, as it is shown to be more effective at generating successful attacks than alternative surrogate losses. We let $\mathbf{z}_I = \mathrm{PGD}(\mathbf{z}_0, f, \varepsilon, I)$ denote the above process, where $\mathbf{z}_0, \hat{\mathbf{x}}_t, \varepsilon$ can be assigned $\tilde{\mathbf{x}}_{t-1}, \mathbf{z}_I, \varepsilon_t$ respectively to solve (6).

Finally, let $\varepsilon_t = \sigma\beta_t$, where $\sigma \in [0, 1]$ adjusts the strength of the adversarial guidance. This means that the adversarial perturbation injected by (6) is always smaller than the noise scale used by the diffusion model, and decreasing w.r.t. the variance schedule to ensure naturalness of the synthesized samples.

3.2. Adversarial Inpainting

In the previous section, we outline how AdvDiffuser is capable of synthesizing adversarial examples from scratch using (5). Here, we continue by introducing *adversarial inpainting*. This technique allows for the creation of natural-looking adversarial examples based on reference images. The process ensures that the generated image closely resembles the reference image, while also manipulating aspects such as background textures, shapes, or objects, which the defending classifier may view as containing irrelevant features. The goal is to produce an image that can successfully deceive the defending classifier while preferably preserving the salient object in the original image.

The process starts with identifying salient regions in the reference image $\mathbf{x}_0$ of ground-truth label y with gradient-weighted class activation mapping (Grad-CAM) [35]. Grad-CAM helps to localize class-specific regions of the corresponding object of y based on the defending classifier f. The localization is then further normalized into $[0, 1]$ to become a mask for the salient object:

$$\mathbf{m} = \mathrm{GradCAM}(f, \mathbf{x}_0, y). \tag{7}$$

Inspired by the inpainting technique [22], in each denoising step t, we evaluate the following

$$\begin{aligned}
&\mathbf{x}_{t-1} = \mathbf{m} \odot \mathbf{x}_{t-1}^{\mathrm{obj}} + (\mathbf{1} - \mathbf{m}) \odot \hat{\mathbf{x}}_{t-1}, \text{where} \\
&\mathbf{x}_{t-1}^{\mathrm{obj}} \sim \mathcal{N}\big(\sqrt{\overline{\alpha}_t}\mathbf{x}_0, (1 - \overline{\alpha}_t)\mathbf{I}\big), \\
&\hat{\mathbf{x}}_{t-1} = \mathrm{PGD}(\tilde{\mathbf{x}}_{t-1}, f, \sigma\beta_{t-1}, I), \\
&\tilde{\mathbf{x}}_{t-1} \sim p_{\boldsymbol{\theta}}(\mathbf{x}_{t-1}|\mathbf{x}_t),
\end{aligned} \tag{8}$$

and recall that $\tilde{\mathbf{x}}_{t-1}$ can be sampled using (4) on $\mathbf{x}_t$.

3.3. The AdvDiffuser Algorithm

We provide a complete algorithmic overview of AdvDiffuser in Algorithm 1. The algorithm accepts a diffusion model $\epsilon_{\boldsymbol{\theta}}$, an attacked classifier f, an optional reference image $\mathbf{x}$, a ground-truth label y, adversarial guidance scale σ, adversarial iterations I, and a noise schedule $\beta_{1:T}$ as input. If a reference image is specified, it evaluates the salient object mask $\mathbf{m}$. For each diffusion step t, the algorithm iteratively denoises the latent variable $\hat{\mathbf{x}}_t$ using a conditional diffuser for the target y. After that, it injects a small adversarial perturbation and constructs $\mathbf{z}_I$ with a PGD attack. It then preserves the salient object by an interpolation between the noised image $\mathbf{x}_{t-1}^{\mathrm{obj}}$ and $\mathbf{z}_I$ using the mask $\mathbf{m}$. Eventually, it produces the natural adversarial example $\hat{\mathbf{x}}_0$ after completing all the steps.

4. Experimental Results

This section begins by describing the experimental setting, comparison methodology, and the evaluation metrics. We then provide quantitative and qualitative comparisons

Algorithm 1 The overall algorithm of AdvDiffuser.

```
function AdvDiffuser(diffusion model $\epsilon_{\boldsymbol{\theta}}$, attacked classifier $f$, optional reference image $\mathbf{x}$, ground-truth label $y$, adversarial guidance scale $\sigma$, adversarial iterations $I$, noise schedule $\beta_{1:T}$ )
    $\hat{\mathbf{x}}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I}); \mathbf{m} \leftarrow \mathbf{0}$
    if $\mathbf{x}$ exists then
        $\mathbf{m} \leftarrow \mathrm{GradCAM}(\mathbf{x}, f, y)$
    end if
    for $t \in [T, T-1, \dots, 1]$ do
        $\mathbf{z}_0 \leftarrow \frac{1}{\sqrt{\alpha_t}}\left(\hat{\mathbf{x}}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}}\epsilon_{\boldsymbol{\theta}}(\hat{\mathbf{x}}_t, t, y)\right)$
        for $i \in [0, 1, \dots, I-1]$ do
            if $\arg\max f(\mathbf{z}_i) \neq y$ then
                break
            end if
            $\mathbf{z}_{i+1} \leftarrow \mathcal{P}_{\mathbf{z}_0,\sigma\beta_t}(\mathbf{z}_i + \mathrm{sign}(\nabla_{\mathbf{z}_i}\mathsf{L}(f(\mathbf{z}_i, y))))$
        end for
        $\mathbf{x}^{\mathrm{obj}}_{t-1} \sim \mathcal{N}(\sqrt{\overline{\alpha}_{t-1}}\mathbf{x}, (1-\overline{\alpha}_{t-1})\mathbf{I})$
        $\hat{\mathbf{x}}_{t-1} \leftarrow \mathbf{m} \odot \mathbf{x}^{\mathrm{obj}}_{t-1} + (\mathbf{1}-\mathbf{m}) \odot \mathbf{z}_I$
    end for
    return $\hat{\mathbf{x}}_0$
end function
```

against the current SOTAs on the stealthiness of introduced perturbations and the degree of realism of synthesized examples. Finally, we provide ablation and sensitivity analyses for its functioning components and hyperparameters.

4.1. Experimental settings

Dataset and models. For diffusion models that can generate synthetic examples, we adopt pre-trained conditional DDPM models with classifier-free guidance from OpenAI for ImageNet[2] and our reproduction for CIFAR-10 and CelebA. We adopt the white-box assumption for all UAE attack algorithms, which allows them to directly evaluate gradient information using the defending model.

Hyperparameters. We let the number of diffusion steps $T = 100$ and $T = 400$ respectively for CIFAR-10 and ImageNet datasets. For adversarial guidance scale, we chose $\sigma = 0.1$ for CIFAR-10 and $\sigma = 0.4$ for ImageNet. We also let the adversarial attack iterations $I = 1$ and 25 respectively for the two datasets. For additional experimental settings, please refer to the Appendix.

4.2. Synthetic Adversarial Examples from Scratch

We begin by comparing AdvDiffuser against AC-GAN [38] on their respective abilities to generate adversarial examples from scratch. To conduct this comparison, we use the same robust gender classifier, as trained adversarially in [38]. It has a natural accuracy of 97.3% and a robust accuracy of 76.5% under the $\ell_\infty = 8/255$ PGD-50 attack. As shown in Table 1, AdvDiffuser outperforms AC-GAN in terms of success rate, FID [12] score, and speed of sample generation. Figure 2 shows randomly sampled UAEs with the respective methods. As evinced by the comparison, it further shows that AdvDiffuser can generate cohesive face images, whereas AC-GAN may fail to produce images with realistic face features. We further provide samples of adversarial examples synthesized from scratch for ImageNet models, as shown in Figure 3. Please refer to Appendix A for the detailed configurations of the experiment.

Table 1: Comparsion between CelebA UAEs generated from scratch with AdvDiffuser and AC-GAN.

Metrics	AC-GAN	AdvDiffuser
Attack success rate (%)	91.1	**99.1**
Speed (seconds/image)	23.6	**12.8**
Fréchet inception distance (FID)	15.6	**8.4**

4.3. Unrestricted Adversarial Examples

For image-dependent UAE synthesis, we compare AdvDiffuser with the current SOTA, Geometry-aware (GA) attacks [20], the 1$^{\text{th}}$ place winner of the 2021 CVPR competition [4]. and it comprises and naturally subsumes two sub-attacks, the GA-PGD which uses the PGD attack [23], and GA-FSA with feature space attack (FSA) [46]. For black-box transferability, the GA attack uses validation models to determine optimal perturbation budgets. In the case of white-box attacks, such validation models are not necessary, and we use the same perturbation budget increments.

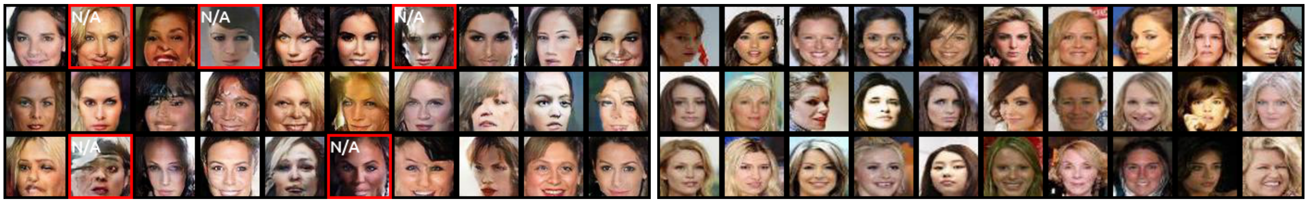
4.3.1 CIFAR-10

We use a normally trained WideResNet-28-10 model (Standard) as baseline for CIFAR-10, and incorporated the top three most robust models in ℓ_2 perturbations from the RobustBench leaderboard [6]. These models are two WideResNet-70-16 models from [29]. In the former model (Rebuffi *et al.* A), external data was employed for its training, while the latter (Rebuffi *et al.* B) used images generated by DDPMs trained on existing training data. Additionally, we include a WideResNet-70-16 model by Gowal *et al.* [10]. As shown in Figure 4, our attack method can generate an adversarial sample similar to the original image but with diverse features. Table 2 provides the attack success rates on the respective models. In Appendix D we further compare the attack methods under DiffPure, a defense mechanism which leverages diffusion models to purify adversarial perturbations.

4.3.2 ImageNet

For the attacked network on the ImageNet test set, we use models produced by [34]: a WideResNet-50-2 (Salmon *et*

[2] https://github.com/openai/guided-diffusion.



(a) AC-GAN. (b) AdvDiffuser.

Figure 2: Adversarial examples generated from scratch (not cherry-picked) by AC-GAN (a) and AdvDiffuser (b) on CelebA. The defending model is an adversarially-trained robust gender classifier. Images generated are females faces and the classifier predict as male. Red-bordered images denote the model fails to generate cohesive faces.

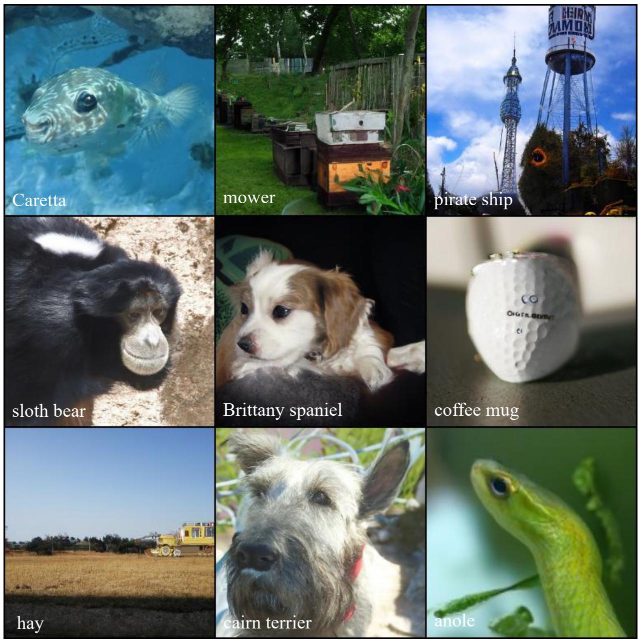


Figure 3: Adversarial examples generated from scratch (not cherry-picked) using AdvDiffuser for Engstrom *et al.* [8]. We include the predicted labels by the model.

Table 2: Attack success rates (%) on CIFAR-10. For reference, we provide the best known robust accuracy of these models with an ℓ_2 perturbation bound of 0.5.

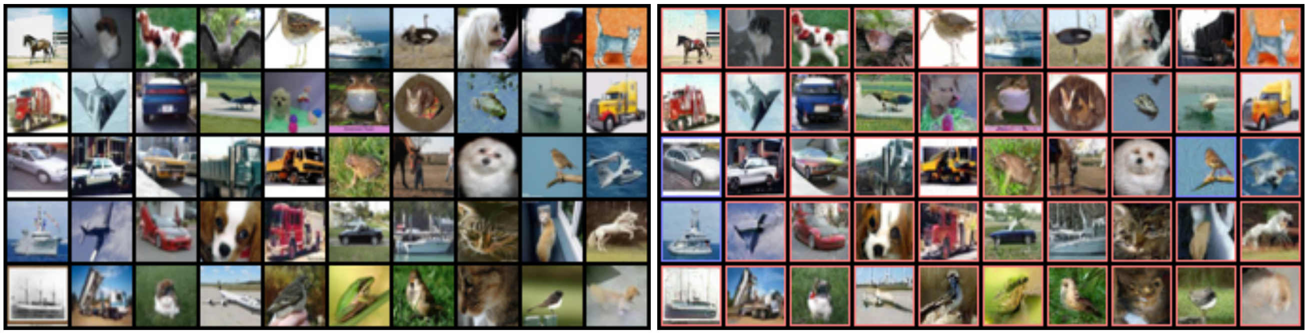
Model	$\ell_2 = 0.5$	AdvDiffuser	GA-PGD
Standard [6]	0.00	**0.27**	1.39
Rebuffi *et al.* A [29]	82.32	**9.81**	83.01
Gowal *et al.* [11]	80.53	**11.10**	80.73
Rebuffi *et al.* B [29]	80.42	**11.77**	80.15

al. A) and a ResNet-50 (Salmon *et al.* B), the current most robust convolutional neural networks on the RobustBench leaderboard [6], and a ResNet-50 trained with standard PGD adversarial training (Engstrom *et al.*) [8]. As a baseline, we also use a normally trained WideResNet-28-10 (Standard). We use an identical subset of the ImageNet test set as GA-attack [20]. This subset contains 1000 randomly selected images. GA-attack variants are generally effective against defenses. However, as depicted in Figure 5, they transform the overall color of the image to an extent, causing significant color shifts. On the other hand, the perturbations created by GA-PGD are easily noticeable in areas with low information (*e.g.* the background sky). In contrast, our UAEs are more realistic. AdvDiffuser not only enjoys higher success rates than the two GA variants, but they are more difficult to identify and have higher SSIM, lower LPIPS and FID scores, as shown in Table 3. Figure 6 depicts the FID, average ℓ_∞, LPIPS, and SSIM distance metrics w.r.t. attack success rates as we vary the strength of each attack. This figure shows that AdvDiffuser consistently outperforms the competition since it results in minor changes across all metrics, except for the ℓ_∞ distance metric. We expect the ℓ_∞ distance as its not the goal of our optimization, and Figure 5 shows that ℓ_∞ bounded attacks produce noticeable artefacts. In addition, it is not pertinent to the perceptual metrics that we consider.

Finally, we amplify and show the perturbations added by the respective attacks in Figure 7. Our findings show that the perturbations are consistent with "shape-specific" changes that are in line with the natural image distribution. We also demonstrate that our UAEs can maintain the original semantic content of the image even under significant perturbations. This observation validates the idea that incorporating the backward denoising process and adversarial guidance generates perturbations that adhere more closely to the clean image distribution. In contrast, we observe that GA-PGD creates UAEs with high-frequency noise that has a visible "texture" bias and thus may appear less natural.

4.4. Robustness against Unseen Threat Models

Rebuffi *et al.* [29] demonstrate that diffusion models as a data-augmentation technique can improve adversarial training. Inspired by their discovery, we explore the potential for AdvDiffuser to dynamically generate adversarial examples for the model to perform adversarial training. Yet unlike existing adversarial training techniques that consider ℓ_p robustness, we do not train our model with explicit assumptions on the threat model. We seek to evaluate the effectiveness of the different approaches using various threat models. These include the conventional ℓ_∞ and ℓ_2 attacks,



(a) Original Images.

(b) Unrestricted Adversarial Examples produced by AdvDiffuser.

Figure 4: Comparing the original images (a) from CIFAR-10, with their respective unrestricted adversarial examples (b) produced with AdvDiffuser. Images with a red / blue border indicate successful / failed attacks.

Table 3: Comparing attacks on ImageNet defending models. For reference, we provide their respective best known robustness within $\ell_\infty = 4/255$ from [6].

Attacker	Accuracy	LPIPS	SSIM	FID
	Standard [6]			
$\ell_\infty = 4/255$	0.0	-	-	-
AdvDiffuser	**0.0**	**0.03**	**0.99**	**20.9**
GA-PGD	0.0	0.27	0.73	38.8
GA-FSA	0.0	0.30	0.66	63.7
	Salman *et al.* A [34]			
$\ell_\infty = 4/255$	38.1	-	-	-
AdvDiffuser	**0.5**	**0.05**	**0.97**	**26.7**
GA-PGD	2.5	0.24	0.80	49.5
GA-FSA	5.5	0.34	0.60	69.4
	Salman *et al.* B [34]			
$\ell_\infty = 4/255$	34.9	-	-	-
AdvDiffuser	**0.2**	**0.05**	**0.97**	**27.2**
GA-PGD	5.6	0.24	0.80	48.9
GA-FSA	4.0	0.34	0.59	67.3
	Engstrom *et al.* [8]			
$\ell_\infty = 4/255$	29.2	-	-	-
AdvDiffuser	**0.6**	**0.05**	**0.98**	**25.9**
GA-PGD	1.0	0.34	0.59	49.2
GA-FSA	4.6	0.24	0.79	66.9

JPEG corruption [16], ReColorAdv [18], Lagrangian perceptual attack (LPA) [19], and spatially-transformed adversarial attack (StAdv) [44]. We carry out a series of experiments on CIFAR-10 in Table 4. Note that models trained with traditional ℓ_2 bounds are not robust against attacks with unseen threat models. In stark contrast, all of our defenses gain certain degree of robustness against all threat models.

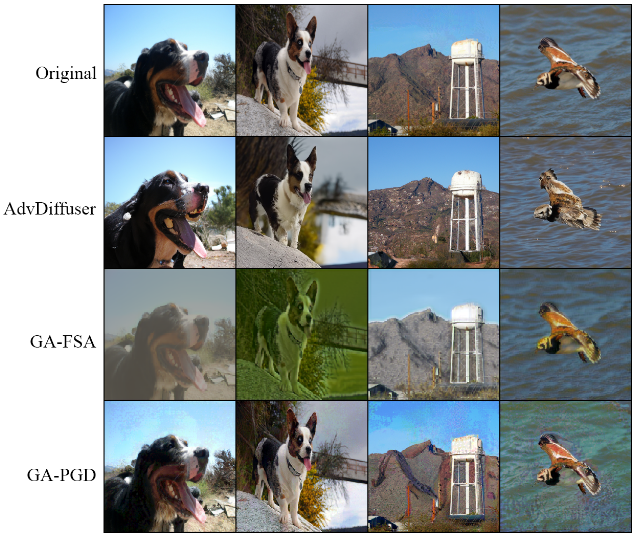


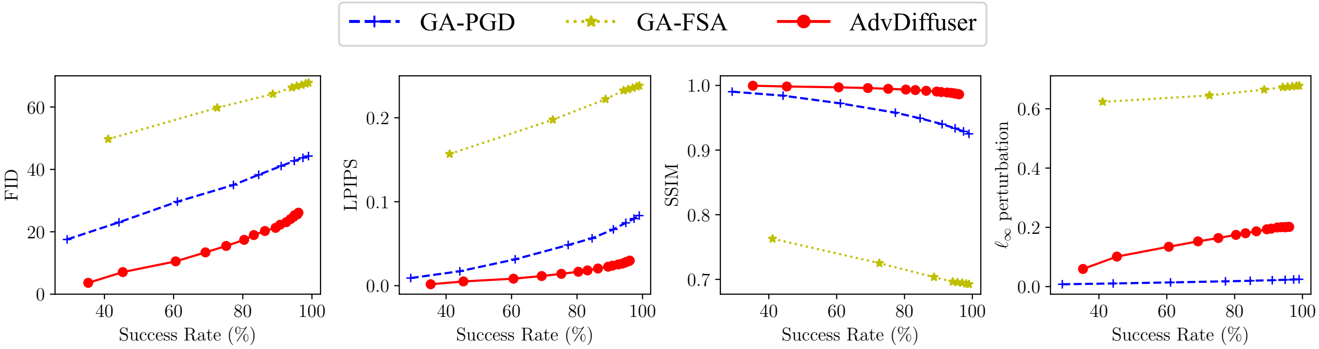
Figure 5: Unrestricted adversarial examples generated by the different attack methods for ImageNet. The defender is Salman *et al.* A [34].

5. Addtional Results

Finally, we provide additional results in the Appendix. Appendix A provides detailed experimental configurations. In Appendix B, we examine hyperparameters and components introduced in AdvDiffuser through sensitivity and ablation analyses. Appendix D presents our results for DiffPure [25] defenses, which use diffusion models to remove adversarial perturbations from images. Lastly, in Appendix E, we provide additional UAE examples for ImageNet.

6. Conclusion

Using the diffusion model, we introduce a novel technique, AdvDiffuser, for synthesizing an unlimited number of natural adversarial examples. By steering the latent variable during the denoising process with adversarial guidance, we



(a) Attack success rates *vs*. FID. (b) Attack success rates *vs*. LPIPS. (c) Attack success rates *vs*. SSIM. (d) Attack success rates *vs*. ℓ_∞.

Figure 6: Comparing traditional and perceptual distance metrics and the SSIM metric w.r.t. attack success rates, as we vary the strengths of each attack. The defending model is Engstrom *et al*. [8] trained on ImageNet. AdvDiffuser consistently shows better perceptual distances than the competing algorithms for successful attacks.

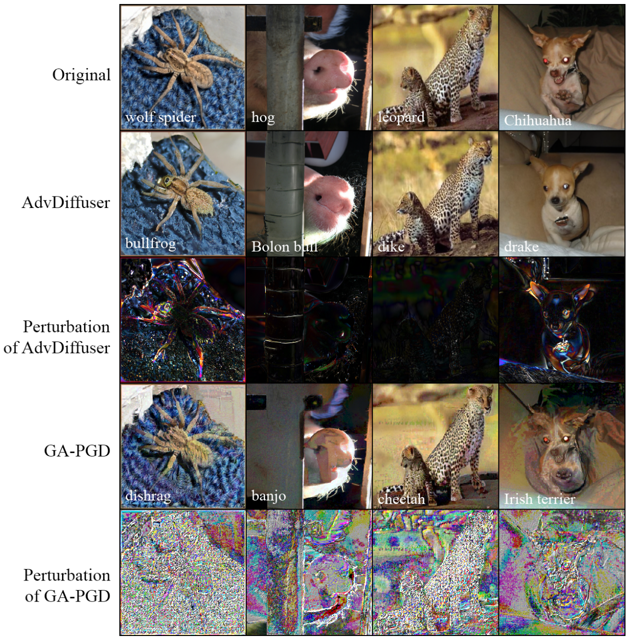


Figure 7: Visualization of adversarial perturbations generated by different attacks. The attacked model is Salman *et al*. A [34]. We include the predicted labels by the model.

can enable diffusion models to generate natural yet powerful adversarial examples. Our experimental results show that existing robust models are unable to defend against these attacks. Moreover, our UAEs outperform prior works while being more natural and less undetectable. They exhibit smaller perceptual distance while enjoying higher success rates. In contrast to ℓ_p adversarial training, Adversarial training with AdvDiffuser shows that models can acquire robustness against threat models unseen during training. We hope that our work will inspire and bring more attention to the topic of attacks involving unrestricted adversarial examples, as well as the defenses against them.

Table 4: Accuracies (%) under attacks of ℓ_p and unseen threat models on adversarially trained CIFAR-10 models. The baseline architecture achieves a 94.7% accuracy on clean test set with clean training. Here, "UAE", "SAE" and "Mixed" are all adversarially trained with AdvDiffuser. "UAE" uses image-conditioned UAEs, "SAE" generates adversarial examples from scratch, and finally "Mixed" uses both to train the robust model. As "Engstrom *et al*." [8] is trained with ℓ_2 adversarial training, it performs well on ℓ_∞ and ℓ_2 adversarial perturbations but poorly on novel threat models.

Models	UAE	SAE	Mixed	Engstrom *et al.*
Clean	81.8	61.2	67.0	**90.2**
$\ell_\infty = 8/255$	8.0	22.8	12.3	**33.4**
$\ell_2 = 1$	7.8	25.4	13.6	**38.6**
JPEG	3.0	**17.6**	8.6	6.4
ReColorAdv	**39.0**	34.4	35.0	38.0
LPA	**35.6**	30.2	30.6	0.0
StAdv	1.5	**6.8**	2.4	0.4

Acknowledgments

This work is supported in part by National Natural Science Foundation of China (No. 92267105) Guangdong Basic and Applied Basic Research Foundation (No. 2020B1515130004), Guangdong Special Support Plan (No. 2021TQ06X990), and Basic Research Program of Shenzhen (Nos. JCYJ20190812160003719, JCYJ20200109115418592, JCYJ20220531100804009).

References

[1] Shumeet Baluja and Ian Fischer. Adversarial transformation networks: Learning to generate adversarial examples. *arXiv:1703.09387*, 2017. 2

[2] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples, 2018. 1

[3] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (S&P)*, pages 39–57, 2017. 2, 12

[4] Yuefeng Chen, Xiaofeng Mao, Yuan He, Hui Xue, Chao Li, Yinpeng Dong, Qi-An Fu, Xiao Yang, Wenzhao Xiang, Tianyu Pang, et al. Unrestricted adversarial attacks on ImageNet competition. *abs/2110.09903*, 2021. 1, 5

[5] Jooyoung Choi, Sungwon Kim, Yonghyun Jeong, Youngjune Gwon, and Sungroh Yoon. ILVR: Conditioning method for denoising diffusion probabilistic models. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14347–14356, 2021. 3

[6] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 5, 6, 7

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021. 3, 12

[8] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (Python library), 2019. 6, 7, 8, 13, 14

[9] Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 2

[10] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv:2010.03593*, 2020. 5

[11] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 4218–4233. Curran Associates, Inc., 2021. 6

[12] 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. 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020. 1, 3

[14] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. 3, 12

[15] Surgan Jandial, Puneet Mangla, Sakshi Varshney, and Vineeth Balasubramanian. AdvGAN++: Harnessing latent layers for adversary generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, 2019. 2

[16] Daniel Kang, Yi Sun, Dan Hendrycks, Tom Brown, and Jacob Steinhardt. Testing robustness against unforeseen adversaries. *arXiv:1908.08016*, 2019. 7

[17] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4396–4405, 2019. 1

[18] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *NeurIPS*, 2019. 7

[19] C Laidlaw, S Singla, and S Feizi. Perceptual adversarial robustness: Defense against unseen threat models. In *International Conference on Learning Representations (ICLR)*, 2021. 1, 2, 3, 7

[20] Fangcheng Liu, Chao Zhang, and Hongyang Zhang. Towards transferable unrestricted adversarial examples with minimum changes. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023. 1, 2, 3, 5, 6, 12

[21] Xihui Liu, Dong Huk Park, Samaneh Azadi, Gong Zhang, Arman Chopikyan, Yuxiao Hu, Humphrey Shi, Anna Rohrbach, and Trevor Darrell. More control for free! Image synthesis with semantic diffusion guidance. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 289–299, 2023. 3

[22] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using denoising diffusion probabilis-

tic models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11461–11471, 2022. 4

[23] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 5

[24] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8162–8171. PMLR, 18–24 Jul 2021. 3, 12

[25] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16805–16827. PMLR, 17–23 Jul 2022. 3, 7, 13

[26] Tianrui Qin, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. APBench: A unified benchmark for availability poisoning attacks and defenses. *arXiv preprint arXiv:2308.03258*, 2023. 2

[27] Tianrui Qin, Xitong Gao, Juanjuan Zhao, Kejiang Ye, and Cheng-Zhong Xu. Learning the unlearnable: Adversarial augmentations suppress unlearnable example attacks. *arXiv preprint arXiv:2303.15127*, 2023. 2

[28] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[29] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness, 2021. 5, 6

[30] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2022. 3

[31] Ishai Rosenberg, Asaf Shabtai, Yuval Elovici, and Lior Rokach. Adversarial machine learning attacks and defense methods in the cyber security domain. *ACM Computing Surveys (CSUR)*, 54(5):1–36, 2021. 2

[32] Andrew Ross and Finale Doshi-Velez. Improving the adversarial robustness and interpretability of deep neural networks by regularizing their input gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 2

[33] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. 14

[34] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust ImageNet models transfer better? In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 3533–3545. Curran Associates, Inc., 2020. 5, 7, 8

[35] 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 (ICCV)*, Oct 2017. 2, 3, 4

[36] Ali Shahin Shamsabadi, Changjae Oh, and Andrea Cavallaro. Semantically adversarial learnable filters. *IEEE Transactions on Image Processing*, 30:8075–8087, 2021. 1

[37] A. Shahin Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro. ColorFool: Semantic adversarial colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1148–1157, Los Alamitos, CA, USA, jun 2020. IEEE Computer Society. 1, 2

[38] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. 1, 2, 5

[39] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[40] Florian Tramèr, Pascal Dupré, Gili Rusak, Giancarlo Pellegrino, and Dan Boneh. AdVersarial: Perceptual Ad blocking meets adversarial machine learning. In *Proceedings of the 2019 ACM SIGSAC Conference on Computer and Communications Security*, pages 2005–2021, 2019. 2

[41] Francisco Utrera, Evan Kravitz, N. Benjamin Erichson, Rajiv Khanna, and Michael W. Mahoney.

Adversarially-trained deep nets transfer better: Illustration on image classification. In *International Conference on Learning Representations*, 2021. 2

[42] Zhou Wang. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 1

[43] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 3905–3911, 2018. 2

[44] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*. 2, 7

[45] Chencheng Xu, Zhiwei Hong, Minlie Huang, and Tao Jiang. Acceleration of federated learning with alleviated forgetting in local training. In *International Conference on Learning Representations*, 2022. 2

[46] Qiuling Xu, Guanhong Tao, Siyuan Cheng, and Xiangyu Zhang. Towards feature space adversarial attack by style perturbation. 35(12):10523–10531, 2021. 2, 5

[47] Yunrui Yu, Xitong Gao, and Cheng-Zhong Xu. LAFEAT: Piercing through adversarial defenses with latent features. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5735–5745, 2021. 2, 4, 12

[48] Yunrui Yu, Xitong Gao, and Cheng zhong Xu. MORA: Improving ensemble robustness evaluation with model reweighing attack. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. 2

[49] 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 (CVPR)*, June 2018. 1, 3

[50] Zhengli Zhao, Dheeru Dua, and Sameer Singh. Generating natural adversarial examples. In *International Conference on Learning Representations*, 2018. 1, 2

[51] Zhengyu Zhao, Zhuoran Liu, and Martha Larson. Towards large yet imperceptible adversarial image perturbations with perceptual color distance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020. 1