Fine-grained Synthesis of Unrestricted Adversarial Examples

Omid Poursaeed[1,2] Tianxing Jiang[1] Harry Yang[3] Serge Belongie[1,2] Ser-Nam Lim[3]

[1]Cornell University [2]Cornell Tech [3]Facebook AI

Abstract

We propose a novel approach for generating unrestricted adversarial examples by manipulating fine-grained aspects of image generation. Unlike existing unrestricted attacks that typically hand-craft geometric transformations, we learn stylistic and stochastic modifications leveraging state-of-the-art generative models. This allows us to manipulate an image in a controlled, fine-grained manner without being bounded by a norm threshold. Our model can be used for both targeted and non-targeted unrestricted attacks. We demonstrate that our attacks can bypass certified defenses, yet our adversarial images look indistinguishable from natural images as verified by human evaluation. Adversarial training can be used as an effective defense without degrading performance of the model on clean images. We perform experiments on LSUN and CelebA-HQ as high resolution datasets to validate efficacy of our proposed approach.

1. Introduction

Adversarial examples, inputs resembling real samples but maliciously crafted to mislead machine learning models, have been studied extensively in the last few years. Most of the existing papers, however, focus on norm-constrained attacks and defenses, in which the adversarial input lies in the ϵ-neighborhood of a real sample using the L_p distance metric (commonly with $p = 0, 2, \infty$). For small ϵ, the adversarial input is quasi-indistinguishable from the natural sample. For an adversarial image to fool the human visual system, it is sufficient to be norm-constrained; but this condition is not necessary. Moreover, defenses tailored for norm-constrained attacks can fail on other subtle input modifications [15]. This has led to a recent surge of interest on unrestricted adversarial attacks in which the adversary is not restricted by a norm threshold. These methods typically hand-craft transformations to capture visual similarity. Spatial transformations [15, 56, 1], viewpoint or pose changes [2], inserting small patches [7], among other methods, have been proposed to generate unrestricted adversarial examples.

In this paper, we focus on fine-grained manipulation of images for unrestricted adversarial attacks. Building upon the Style-GAN model [26] which disentangles fine and coarse-grained variations of images, we manipulate stylistic and stochastic latent variables in order to mislead a classification model. Loss of the target classifier is used to learn subtle variations to create adversarial examples. The pre-trained generative model constrains the search space to natural-looking images. We verify that we do not deviate from the space of realistic images with a user study using Amazon Mechanical Turk. Finally, we demonstrate that our proposed attack can break certified defenses, revealing new vulnerabilities of existing approaches. Adversarial training can be used as an effective defense, and unlike training on norm-bounded adversarial examples, it does not decrease accuracy on clean images. We elaborate on the proposed approach in Section 3.

2. Related Work

2.1. Norm-constrained Adversarial Examples

Most of the existing works on adversarial attacks and defenses focus on norm-constrained adversarial examples: for a given classifier $F : \mathbb{R}^n \rightarrow \{1, \ldots, K\}$ and an image $x \in \mathbb{R}^n$, the adversarial image $x' \in \mathbb{R}^n$ is created such that $\|x - x'\|_p < \epsilon$ and $F(x) \neq F(x')$. Common values for p are $0, 2, \infty$, and ϵ is chosen small enough so that the perturbation is imperceptible. Various algorithms have been proposed for creating x' from x. Optimization-based methods solve a surrogate optimization problem based on the classifier's loss and the perturbation norm. In their pioneering paper on adversarial examples, Szegedy *et al.* [48] use box-constrained L-BFGS [16] to minimize the surrogate loss function. Carlini and Wagner [9] propose stronger optimization-based attacks for L_0, L_2 and L_∞ norms using better objective functions and the Adam optimizer [28]. Deep-Fool is introduced in [36] as a non-targeted attack optimized for the L_2 distance. It iteratively computes a minimal norm adversarial perturbation for a given image by linearly approximating the decision function. Gradient-based methods use gradient of the classifier's loss with respect

to the input image. Fast Gradient Sign Method (FGSM) [18] uses a first-order approximation of the function for faster generation, and is optimized for the L_∞ norm. Projected Gradient Descent (PGD) [35] is an iterative variant of FGSM which provides a strong first-order attack by using multiple steps of gradient ascent and projecting perturbed images to an ϵ-ball centered at the input. Other variants of FGSM are proposed in [13, 29]. Jacobian-based Saliency Map Attack (JSMA) [39] is a greedy algorithm that modifies pixels one at a time. It uses the gradients to compute a saliency map, picks the most important pixel and modifies it to increase likelihood of the target class. Li *et al.* [32] introduce a gradient transformer module to generate regionally homogeneous perturbations. They claim state-of-the-art attack results, which are independent of input images and can be transferred to black-box models. Generative attack methods [4, 40, 55] use an auxiliary network to learn adversarial perturbations, which provides benefits such as faster inference and more diversity in the synthesized images.

Several methods have been proposed for defending against adversarial attacks. These approaches can be broadly categorized to *empirical* defenses which are empirically robust to adversarial examples, and *certified* defenses which are provably robust to a certain class of attacks. One of the most successful empirical defenses is *adversarial training* [18, 29, 35] which augments training data with adversarial examples generated as the training progresses. Adversarial logit pairing [24] is a form of adversarial training which constrains logit predictions of a clean image and its adversarial counterpart to be similar. Many empirical defenses attempt to defeat adversaries using a form of input pre-processing or by manipulating intermediate features or gradients [31, 19, 57, 44, 33, 58]. Few approaches have been able to scale up to high-resolution datasets such as ImageNet [57, 33, 58, 42, 24]. Most of the proposed heuristic defenses were later broken by stronger adversaries [9, 51, 3]. Athalye *et al.* [3] show that many of these defenses fail due to an issue they term *obfuscated gradients*, which occurs when the defense method is designed to mask information about the model's gradients. They propose workarounds to obtain approximate gradients for adversarial attacks. Vulnerabilities of empirical defenses have led to increased interest in certified defenses, which provide a guarantee that the classifier's prediction is constant within a neighborhood of the input. Several certified defenses have been proposed [54, 43, 14, 50] which typically do not scale to ImageNet. Cohen *et al.* [10] use randomized smoothing with Gaussian noise to obtain provably L_2-robust classifiers on ImageNet.

2.2. Unrestricted Adversarial Examples

For an image to be adversarial, it needs to be visually indistinguishable from real images. One way to achieve this is by applying subtle geometric transformations to the input image. Spatially transformed adversarial examples are introduced in [56] in which a flow field is learned to displace pixels of the image. They use sum of spatial movement distance for adjacent pixels as a regularization loss to minimize the local distortion introduced by the flow field. Similarly, Alaifari *et al.* [1] iteratively apply small deformations, found through a gradient descent step, to the input in order to obtain the adversarial image. Engstrom *et al.* [15] show that simple translations and rotations are enough for fooling deep neural networks. This remains to be the case even when the model has been trained using appropriate data augmentation. Alcorn *et al.* [2] manipulate pose of an object to fool deep neural networks. They estimate parameters of a 3D renderer that cause the target model to misbehave in response to the rendered image. Another approach for evading the norm constraint is to insert new objects in the image. Adversarial Patch [7] creates an adversarial image by completely replacing part of an image with a synthetic patch. The patch is image-agnostic, robust to transformations, and can be printed and used in real-world settings. Song *et al.* [46] search in the latent (z) space of AC-GAN [38] to find generated images that can fool a classifier, and show results on MNIST [30], SVHN [37] and CelebA [34] datasets. Since the z space is not interpretable, their method has no control over the generation process. On the other hand, our method can manipulate real or synthesized images in a fine-grained, controllable manner. Existence of on-the-manifold adversarial examples is also shown in [17], which considers the task of classifying between two concentric n-dimensional spheres. A challenge for creating unrestricted adversarial examples and defending against them is introduced in [6] using the simple task of classifying between birds and bicycles.

2.3. Fine-grained Image Generation

With recent improvements in generative models, they are able to generate high-resolution and realistic images. Moreover, these models can be used to learn and disentangle various latent factors for image synthesis. Style-GAN is proposed in [26] which disentangles high-level attributes and stochastic variations of generated images in an unsupervised manner. The model learns an intermediate latent space from the input latent code, which is used to adjust style of the image. It also injects noise at each level of the generator to capture stochastic variations. Singh *et al.* introduce Fine-GAN [45], a generative model which disentangles the background, object shape, and object appearance to hierarchically generate images of fine-grained object categories. Layered Recursive GAN is proposed in [60], and generates image background and foreground separately and recursively without direct supervision. Stacking is used in [12, 22, 27, 41, 63] to generate images in

a coarse to fine manner. Conditional fine-grained generation has been explored in several papers. Bao *et al.* [5] introduce a Conditional VAE-GAN for synthesizing images in fine-grained categories. Modeling an image as a composition of label and latent attributes, they vary the fine-grained category label fed into the generative model, and randomly draw values of a latent attribute vector. AttnGAN [59] uses attention-driven, multi-stage refinement for fine-grained text-to-image generation. Hong *et al.* [20] present a hierarchical framework for semantic image manipulation. Their model first learns to generate the pixel-wise semantic label maps from the initial object bounding boxes, and then learns to generate the manipulated image from the predicted label maps. A multi-attribute conditional GAN is proposed in [53], and can generate fine-grained face images based on the specified attributes.

3. Approach

Most of the existing works on unrestricted adversarial attacks rely on geometric transformations and deformations [15, 56, 1], which are oblivious to latent factors of variation. In this paper, we leverage disentangled latent representations of images for unrestricted adversarial attacks. Style-GAN [26] is a state-of-the-art generative model which learns to disentangle high-level attributes and stochastic variations in an unsupervised manner. More specifically, stylistic variations are represented by *style* variables and stochastic details are captured by *noise* variables. Changing the noise only affects low-level details, leaving the overall composition and high-level aspects such as identity intact. This allows us to manipulate the noise variables such that variations are barely noticeable by the human eye, yet the synthesized image can fool a pre-trained classifier. The style variables affect higher level aspects of image generation. For instance, when the model is trained on bedrooms, style variables from the top layers control viewpoint of the camera, middle layers select the particular furniture, and bottom layers deal with colors and details of materials. This allows us to manipulate images in a controlled manner, providing an avenue for fine-grained unrestricted attacks.

Formally, we can represent Style-GAN with a non-linear mapping function f and a synthesis network g. The mapping function is an 8-layer MLP which takes a latent code $\mathbf{z}$, and produces an intermediate latent vector $\mathbf{w} = f(\mathbf{z})$. This vector is then specialized by learned affine transformations A to styles $\mathbf{y} = (\mathbf{y}_s, \mathbf{y}_b)$. Style variables in turn control adaptive instance normalization operations [21] after each convolutional layer of the synthesis network g. Noise inputs are single-channel images consisting of un-correlated Gaussian noise that are fed to each layer of the synthesis network. Learned per-feature scaling factors B are then used to generate noise variables $\boldsymbol{\eta}$ which are added to the output of convolutional layers. The synthesis network takes style $\mathbf{y}$ and noise $\boldsymbol{\eta}$ as input, and generates an image $\mathbf{x} = g(\mathbf{y}, \boldsymbol{\eta})$. We then pass the generated image to a pre-trained classifier F. We seek to slightly modify $\mathbf{x}$ so that F can no longer classify it correctly. We achieve this through perturbing the style and noise tensors, which control different aspects of image generation in a fine-grained manner. More specifically, we initialize adversarial style and noise variables as $\mathbf{y}_{\mathbf{adv}}^{(\mathbf{0})} = \mathbf{y}$ and $\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{0})} = \boldsymbol{\eta}$ respectively. These adversarial tensors are then iteratively updated in order to fool the classifier. Loss of the classifier determines the update rule, which in turn depends on the type of attack. As common in the literature, we consider two types of attacks: non-targeted and targeted.

3.1. Non-targeted Attacks

In order to generate non-targeted adversarial examples, we need to change the model's original prediction. Starting from initial values $\mathbf{y}_{\mathbf{adv}}^{(\mathbf{0})} = \mathbf{y}$ and $\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{0})} = \boldsymbol{\eta}$, we perform gradient ascent in the style and noise spaces of the generator to find values that maximize the classifier's loss. At time step t, the update rule for the style and noise variables is:

$$\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t+1})} = \mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})} + \epsilon \cdot \text{sign}(\nabla_{\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), c_{\mathbf{x}})) \tag{1}$$

$$\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t+1})} = \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})} + \delta \cdot \text{sign}(\nabla_{\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), c_{\mathbf{x}})) \tag{2}$$

in which $J(\cdot, \cdot)$ represents the classifier's loss function (e.g. cross-entropy), $c_{\mathbf{x}}$ is the ground-truth class for $\mathbf{x} = g(\mathbf{y}, \boldsymbol{\eta})$, and $\epsilon, \delta \in \mathbb{R}$ are step sizes. Note that $F(\cdot)$ gives the probability distribution over classes. This formulation resembles Iterative-FGSM [29]; however, the gradients are computed with respect to the noise and style variables of the synthesis network. Alternatively, as proposed in [29] we can use the least-likely predicted class $ll_{\mathbf{x}} = \arg\min(F(\mathbf{x}))$ as our target:

$$\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t+1})} = \mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})} - \epsilon \cdot \text{sign}(\nabla_{\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), ll_{\mathbf{x}})) \tag{3}$$

$$\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t+1})} = \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})} - \delta \cdot \text{sign}(\nabla_{\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), ll_{\mathbf{x}})) \tag{4}$$

We found the latter approach more effective in practice. We use $(\epsilon, \delta) = (0.004, 0.2)$ and $(0.004, 0.1)$ in the experiments on LSUN and CelebA-HQ respectively. We perform multiple steps of gradient descent (usually 2 to 10) until the classifier is fooled. Unlike I-FGSM that generates high-frequency noisy perturbations in the pixel space, our pre-trained generative model constrains the space of generated images to realistic ones.

3.2. Targeted Attacks

Generating targeted adversarial examples is more challenging as we need to change the prediction to a specific

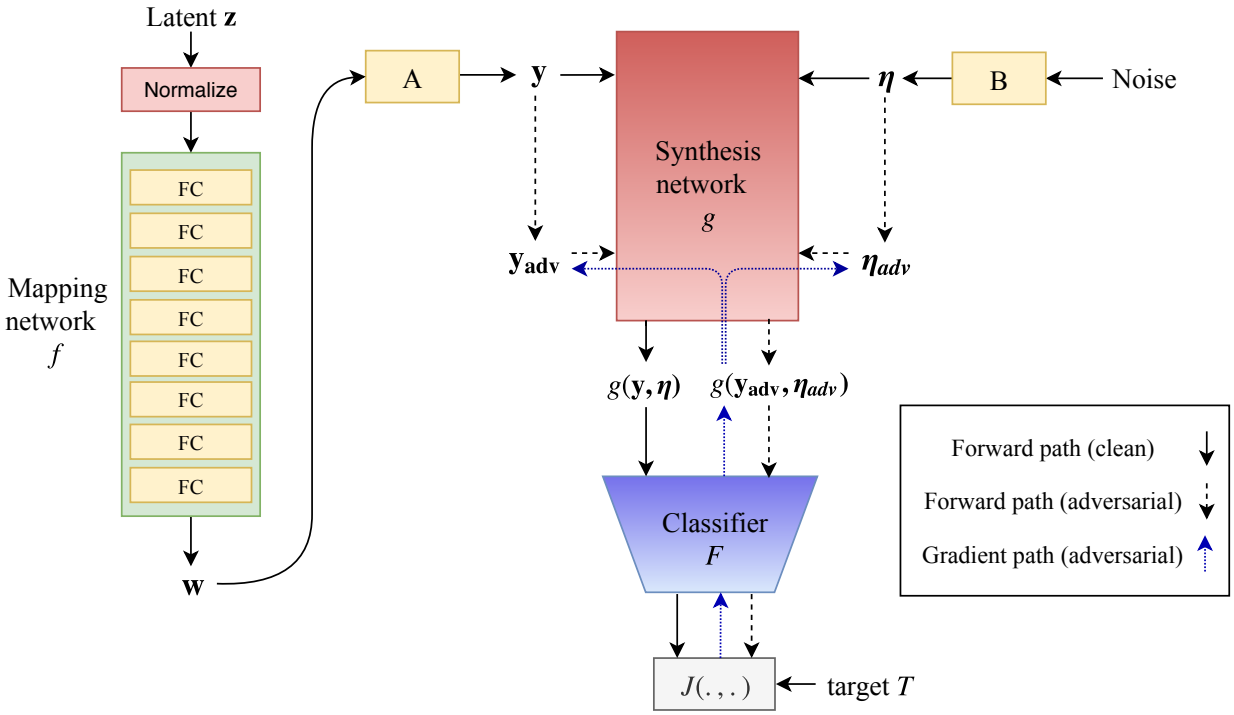


Figure 1: Model architecture. Style ($\mathbf{y}$) and noise ($\boldsymbol{\eta}$) variables are used to generate images $g(\mathbf{y}, \boldsymbol{\eta})$ which are fed to the classifier F. Adversarial style and noise tensors are initialized with $\mathbf{y}$ and $\boldsymbol{\eta}$ and iteratively updated using gradients of the loss function J.

class T. In this case, we perform gradient descent to minimize the classifier's loss with respect to the target:

$$\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t+1})} = \mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})} - \epsilon \cdot \text{sign}(\nabla_{\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), T)) \tag{5}$$

$$\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t+1})} = \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})} - \delta \cdot \text{sign}(\nabla_{\boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})}} J(F(g(\mathbf{y}_{\mathbf{adv}}^{(\mathbf{t})}, \boldsymbol{\eta}_{\mathbf{adv}}^{(\mathbf{t})})), T)) \tag{6}$$

We use $(\epsilon, \delta) = (0.005, 0.2)$ and $(0.004, 0.1)$ in the experiments on LSUN and CelebA-HQ respectively. In practice 3 to 15 updates suffice to fool the classifier. Updating only the noise tensor results in finer adversarial changes, while only using the style variable creates coarser stylistic changes. We can also have a detailed control over the generation process by manipulating specific layers of the synthesis network. Note that we only control deviation from the initial latent variables, and do not impose any norm constraint on generated images.

3.3. Input-conditioned Generation

Generation can also be conditioned on real input images by embedding them into the latent space of Style-GAN. We first synthesize images similar to the given input image I by optimizing values of $\mathbf{y}$ and $\boldsymbol{\eta}$ such that $g(\mathbf{y}, \boldsymbol{\eta})$ is close to I. More specifically, we minimize the perceptual distance [23] between $g(\mathbf{y}, \boldsymbol{\eta})$ and I. We can then proceed similar to equations 3–6 to perturb these tensors and generate the adversarial image. Realism of synthesized images depends on inference properties of the generative model. In practice, generated images resemble input images with high fidelity especially for CelebA-HQ images.

4. Results and Discussion

We provide qualitative and quantitative results demonstrating our proposed approach. Experiments are performed on LSUN [61] and CelebA-HQ [25]. LSUN contains 10 scene categories each with around one million labeled images and 20 object categories. We use all the 10 scene classes as well as two object classes: *cars* and *cats*. We consider this dataset since it is used in Style-GAN, and is well suited for a classification task. For the scene categories, a 10-way classifier is trained based on Inception-v3 [47] which achieves an accuracy of 87.7% on LSUN's test set. The two object classes also appear in ImageNet [11], a richer dataset containing 1000 categories. Therefore, for experiments on cars and cats we use an Inception-v3 model trained on ImageNet. This allows us to explore a broader set of categories in our attacks, and is particularly helpful for targeted adversarial examples. Note that there are multiple classes representing cars and cats in ImageNet, so we identify and group those classes. CelebA-HQ is a high-quality version of the CelebA dataset [34] consisting of 30,000 face images at 1024×1024 resolution. We consider the gender

classification task, and use the classifier provided by Karras *et al.* [26]. This is a binary task for which targeted and non-targeted attacks are similar.

In order to synthesize a variety of adversarial examples, we use different random seeds in Style-GAN to obtain various values for $\mathbf{z}, \mathbf{w}, \mathbf{y}$ and $\boldsymbol{\eta}$. Style-based adversarial examples are generated by initializing $\mathbf{y_{adv}}$ with the value of $\mathbf{y}$, and iteratively updating it as in equation 3 (or 5) until the resulting image $g(\mathbf{y_{adv}}, \boldsymbol{\eta})$ fools the classifier F. Noise-based adversarial examples are created similarly using $\boldsymbol{\eta_{adv}}$ and the update rule in equation 4 (or 6). We can also combine the effect of style and noise by simultaneously updating $\mathbf{y_{adv}}$ and $\boldsymbol{\eta_{adv}}$ in each iteration, and feeding $g(\mathbf{y_{adv}}, \boldsymbol{\eta_{adv}})$ to the classifier. In this case, the effect of style usually dominates since it creates coarser modifications. To make sure the iterative process always converges in reasonable number of steps, we measure the number of updates required to fool the classifier on 1000 randomly-selected images. In the case of non-targeted attacks on LSUN, 3.7 ± 1.8 and 6.8 ± 3.6 (mean $\pm$ std) updates are required for noise-based and style-based examples respectively. For targeted attacks, we first randomly sample a target class different from the ground-truth label for each image. In this case, the number of updates required for noise-based and style-based attacks are 4.5 ± 1.7 and 9.1 ± 4.2 respectively. For the CelebA-HQ dataset, 6.2 ± 4.1 and 7.3 ± 3.0 updates are needed for noise and style manipulation respectively. While using different step sizes makes a fair comparison difficult, we generally found it easier to fool the model by manipulating the noise.

Figure 2 illustrates generated adversarial examples for non-targeted and targeted attacks on LSUN. Original image $g(\mathbf{y}, \boldsymbol{\eta})$, noise-based image $g(\mathbf{y}, \boldsymbol{\eta_{adv}})$ and style-based image $g(\mathbf{y_{adv}}, \boldsymbol{\eta})$ are shown. As we observe, adversarial images look almost indistinguishable from natural images. This also holds in targeted attacks even when original and target classes are very dissimilar. Manipulating the noise variable results in very subtle, imperceptible changes. Varying the style leads to coarser changes such as different colorization, pose changes, and even removing or inserting new objects in the scene. We can also control granularity of changes by selecting specific layers of the model. Manipulating top layers, corresponding to coarse spatial resolutions, results in high-level changes. Lower layers, on the other hand, modify finer details. In the first two columns of Figure 2, we only modify top 6 layers (out of 18) to generate adversarial images. The middle column corresponds to changing layers 7 to 12, and the last two columns use the bottom 6 layers.

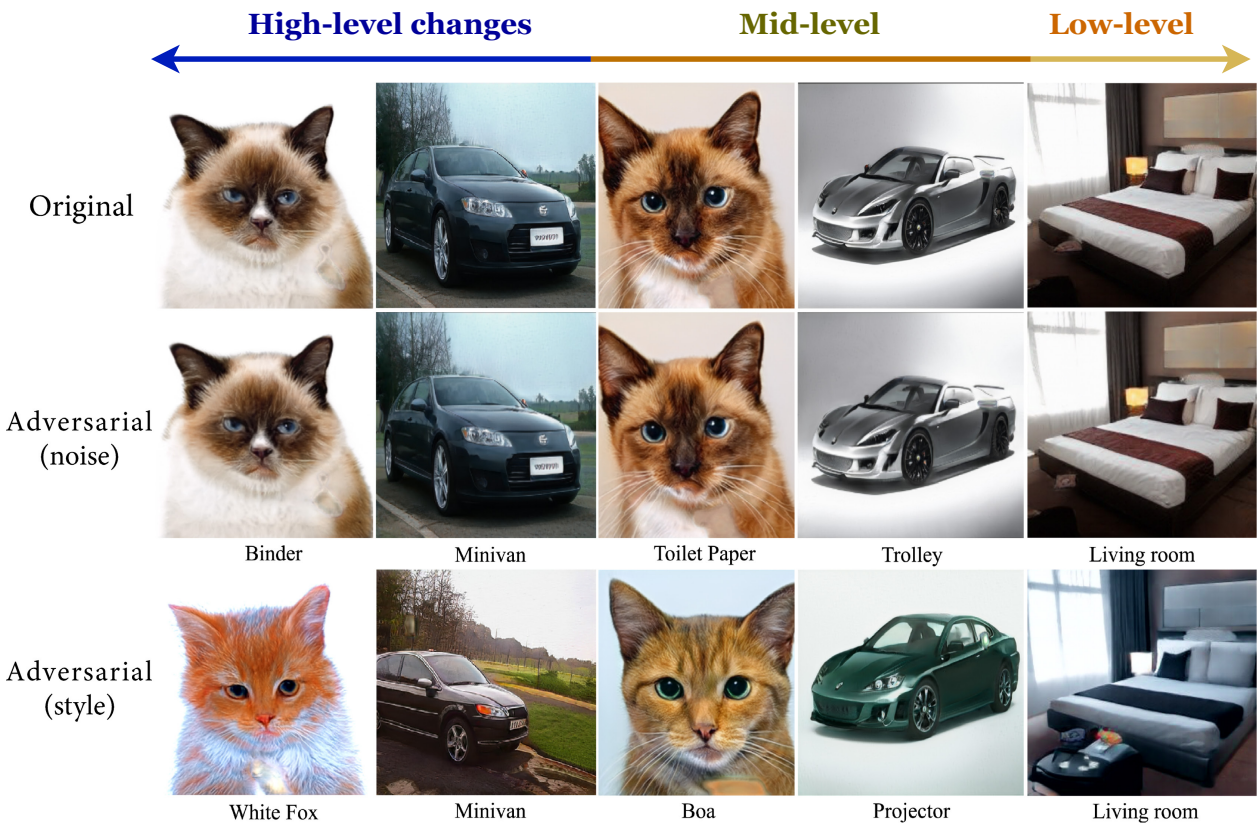
Figure 3 depicts adversarial examples on CelebA-HQ gender classification. Males are classified as females and vice versa. As we observe, various facial features are altered by the model yet the identity is preserved. Similar to LSUN images, noise-based changes are more subtle than style-based ones, and we observe a spectrum of high-level, mid-level and low-level changes. Figure 4 illustrates adversarial examples conditioned on real input images using the procedure described in Section 3.3. Synthesized images resemble inputs with high fidelity, and set the initial values in our optimization process. In some cases, we can notice how the model is altering masculine or feminine features. For instance, women's faces become more masculine in columns 2 and 4, and men's beard is removed in column 3 of Figure 3 and column 1 of Figure 4.

Unlike perturbation-based attacks, L_p distances between original and adversarial images are large, yet they are visually similar. Moreover, we do not observe high-frequency perturbations in the generated images. The model learns to modify the initial input without leaving the manifold of realistic images. Note that the classifiers are trained on millions of images, yet they are easily fooled by these subtle on-the-manifold changes. This poses serious concerns about robustness of deep neural networks, and reveals new vulnerabilities of them. Additional examples and higher-resolution images are provided in the supplementary material.

4.1. User Study

Norm-constrained attacks provide visual realism by L_p proximity to a real input. To verify that our unrestricted adversarial examples are realistic and correctly classified by an oracle, we perform human evaluation using Amazon Mechanical Turk. In the first experiment, each adversarial image is assigned to three workers, and their majority vote is considered as the label. The user interface for each worker contains nine images, and shows possible labels to choose from. We also include the label "Other" for images that workers think do not belong to any specific class. We use 2400 noise-based and 2400 style-based adversarial images from the LSUN dataset, containing 200 samples from each class (10 scene classes and 2 object classes). The results indicate that 99.2% of workers' majority votes match the ground-truth labels. This number is 98.7% for style-based adversarial examples and 99.7% for noise-based ones. As we observe in Figure 2, noise-based examples do not deviate much from the original image, resulting in easier prediction by a human observer. On the other hand, style-based images show coarser changes, which in a few cases result in unrecognizable images or false predictions by the workers.

We use a similar setup in the second experiment but for classifying real versus fake (generated). We also include 2400 real images as well as 2400 unperturbed images generated by Style-GAN. 74.7% of unperturbed images are labeled by workers as real. This number is 74.3% for noise-based adversarial examples and 70.8% for style-based ones, indicating less than 4% drop compared with unperturbed images generated by Style-GAN.



(a) Non-targeted

(b) Targeted

Figure 2: Unrestricted adversarial examples on LSUN for a) non-targeted and b) targeted attacks. Predicted classes are shown under each image. First two columns correspond to manipulating top 6 layers of the synthesis network. The middle column manipulates layers 7 to 12, and the last two columns correspond to the bottom 6 layers.

Figure 3: Unrestricted adversarial examples on CelebA-HQ gender classification. From top to bottom: original, noise-based and style-based adversarial images. Males are classified as females and vice versa. First two columns correspond to manipulating top 6 layers of the synthesis network. The middle three columns manipulate layers 7 to 12, and the last two columns correspond to the bottom 6 layers.

Figure 4: Input-conditioned adversarial examples on CelebA-HQ gender classification. From top to bottom: input, generated and style-based images. Males are classified as females and vice versa.

4.2. Evaluation on Certified Defenses

Several approaches have been proposed in the literature to defend against adversarial examples, which can be broadly divided into *empirical* and *certified* defenses. Empirical defenses are heuristic methods designed to mitigate effects of perturbations, and certified defenses provide provable guarantees on model's robustness. Almost all of these methods consider norm-constrained attacks. Most of the empirical defenses were later broken by stronger adversaries [8, 3]. This has led to a surge of interest in provable defenses. However, most certified defenses are not scalable to high-resolution datasets. Cohen *et al.* [10] propose

the first certified defense at the scale of ImageNet. Using randomized smoothing with Gaussian noise, their defense guarantees a certain top-1 accuracy for perturbations with L_2 norm less than a specific threshold.

We demonstrate that our unrestricted attacks can break the state-of-the-art certified defense on ImageNet. We use 400 noise-based and 400 style-based adversarial images from the object categories of LSUN, and group all relevant ImageNet classes as the ground-truth. Our adversarial examples are evaluated against a randomized smoothing classifier based on ResNet-50 using Gaussian noise with standard deviation of 0.5 [10]. Table 1 shows accuracy of the model on clean and adversarial images. As we observe, the accuracy drops on adversarial inputs, and the certified defense is not effective against our attack. Note that we stop updating adversarial images as soon as the model is fooled. If we keep updating for more iterations afterwards, we can achieve even stronger attacks. Our adversarial examples are learned on Inception-v3, yet they can fool a defended model based on ResNet-50. This indicates that these inputs are transferable to other models, showing their potential for black-box attacks. Considering the variety of methods used for creating unrestricted adversarial examples, designing effective defenses against them is a challenging task. We believe this can be an interesting direction for future research.

	Accuracy
Clean	63.1%
Adversarial (style)	21.7%
Adversarial (noise)	37.8%

Table 1: Accuracy of a certified classifier equipped with randomized smoothing on adversarial images.

4.3. Adversarial Training

Adversarial training increases robustness of models by injecting adversarial examples into training data [18, 35, 29]. This approach makes the classifier robust to perturbations similar to those used in training; however, it can still be vulnerable to black-box adversarial inputs transferred from other models [49]. To mitigate this issue, Ensemble Adversarial Training is proposed in [49] to augment training data with perturbations transferred from other pre-trained models. The main drawback of adversarial training is that it degrades performance of the classifier on clean images [35]. Various regularizers have been proposed to tackle this issue [62, 52].

We show that while adversarial training makes the model robust to our unrestricted adversarial inputs, it does not degrade accuracy on clean images. We perform adversarial training by incorporating generated images in training the LSUN classifier. 400k clean images as well as 50k noise-based and 50k style-based adversarial inputs are used to train the classifier. Same number of samples are used across all scene categories. Table 2 shows accuracy of the strengthened and original classifiers on clean and adversarial test images. Similar to norm-constrained perturbations, adversarial training is an effective defense against our unrestricted attack. However, accuracy of the model on clean test images remains almost the same after adversarial training. This is in contrast to training with norm-bounded adversarial inputs, which hurts classifier's performance on clean images. This is due to the fact that unlike perturbation-based inputs, our generated images live on the manifold of realistic images as constrained by the generative model.

	Adv. Trained	Original
Clean	87.6%	87.7%
Adversarial (noise)	81.2%	0.0%
Adversarial (style)	76.9%	0.0%

Table 2: Accuracy of adversarially trained and original classifiers on clean and adversarial test images.

5. Conclusion and Future Work

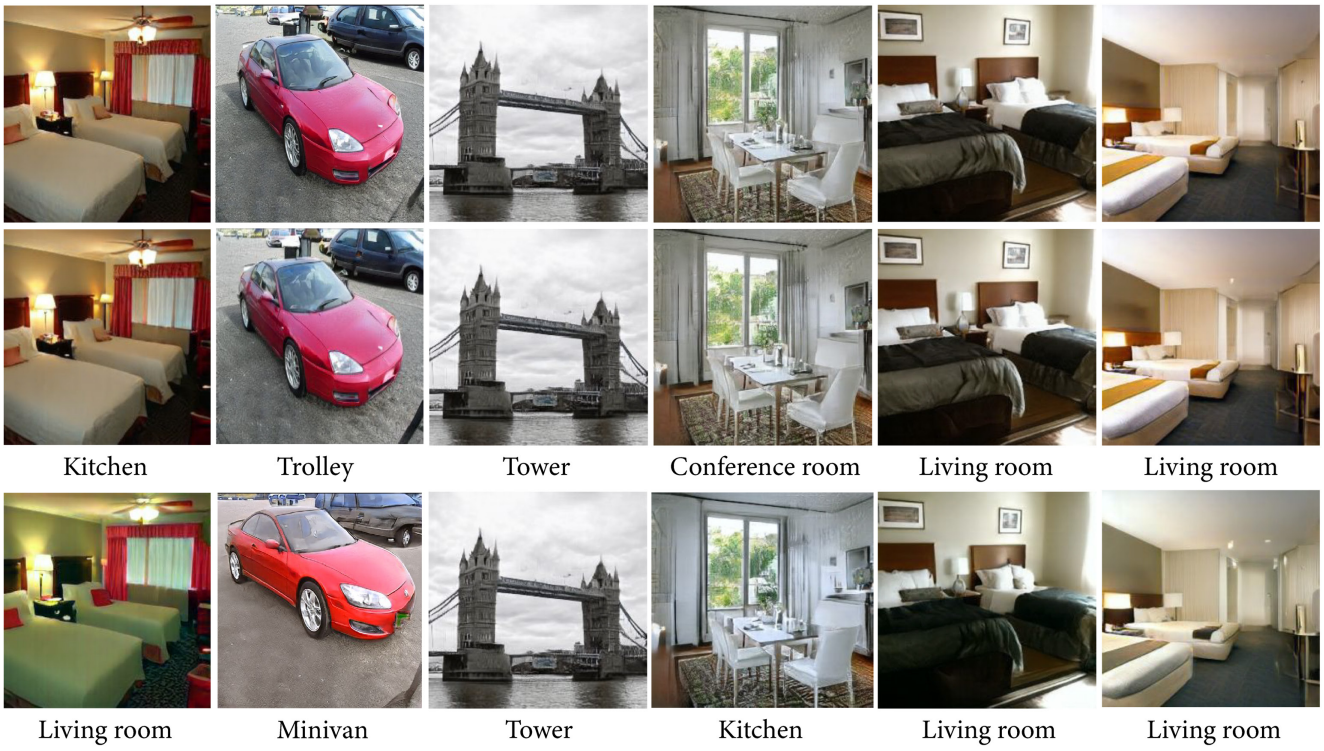
We present a novel approach for creating unrestricted adversarial examples leveraging state-of-the-art generative models. Unlike existing works that rely on hand-crafted transformations, we learn stylistic and stochastic changes to mislead pre-trained models. Loss of the target classifier is used to perform gradient descent in the style and noise spaces of Style-GAN. Subtle adversarial changes can be crafted using noise variables, and coarser modifications can be created through style variables. We demonstrate results in both targeted and non-targeted cases, and validate visual realism of synthesized images through human evaluation. We show that our attacks can break state-of-the-art defenses, revealing vulnerabilities of current norm-constrained defenses to unrestricted attacks. Moreover, while adversarial training can be used to make models robust against our adversarial inputs, it does not degrade accuracy on clean images.

The area of unrestricted adversarial examples is relatively under-explored. Not being bounded by a norm threshold provides its own pros and cons. It allows us to create a diverse set of attack mechanisms; however, fair comparison of relative strength of these attacks is challenging. It is also unclear how to even define provable defenses. While several papers have attempted to interpret norm-constrained attacks in terms of decision boundaries, there has been less effort in understanding the underlying reasons for models' vulnerabilities to unrestricted attacks. We believe these can be promising directions for future research. We also plan to further explore transferability of our approach for black-box attacks in the future.

6. Appendix

We provide additional examples and higher-resolution images in the following. Figure 5 illustrates adversarial examples on CelebA-HQ gender classification, and Figure 6 shows additional examples on the LSUN dataset. Higher-resolution versions for some of the adversarial images are shown in Figure 7, which particularly helps to distinguish subtle differences between original and noise-based images.

Figure 5: Unrestricted adversarial examples on CelebA-HQ gender classification. From top to bottom: Original, noise-based and style-based adversarial images. Males are classified as females and vice versa.



(a) Non-targeted

(b) Targeted

Figure 6: Unrestricted adversarial examples on LSUN for a) non-targeted and b) targeted attacks. From top to bottom: original, noise-based and style-based images.

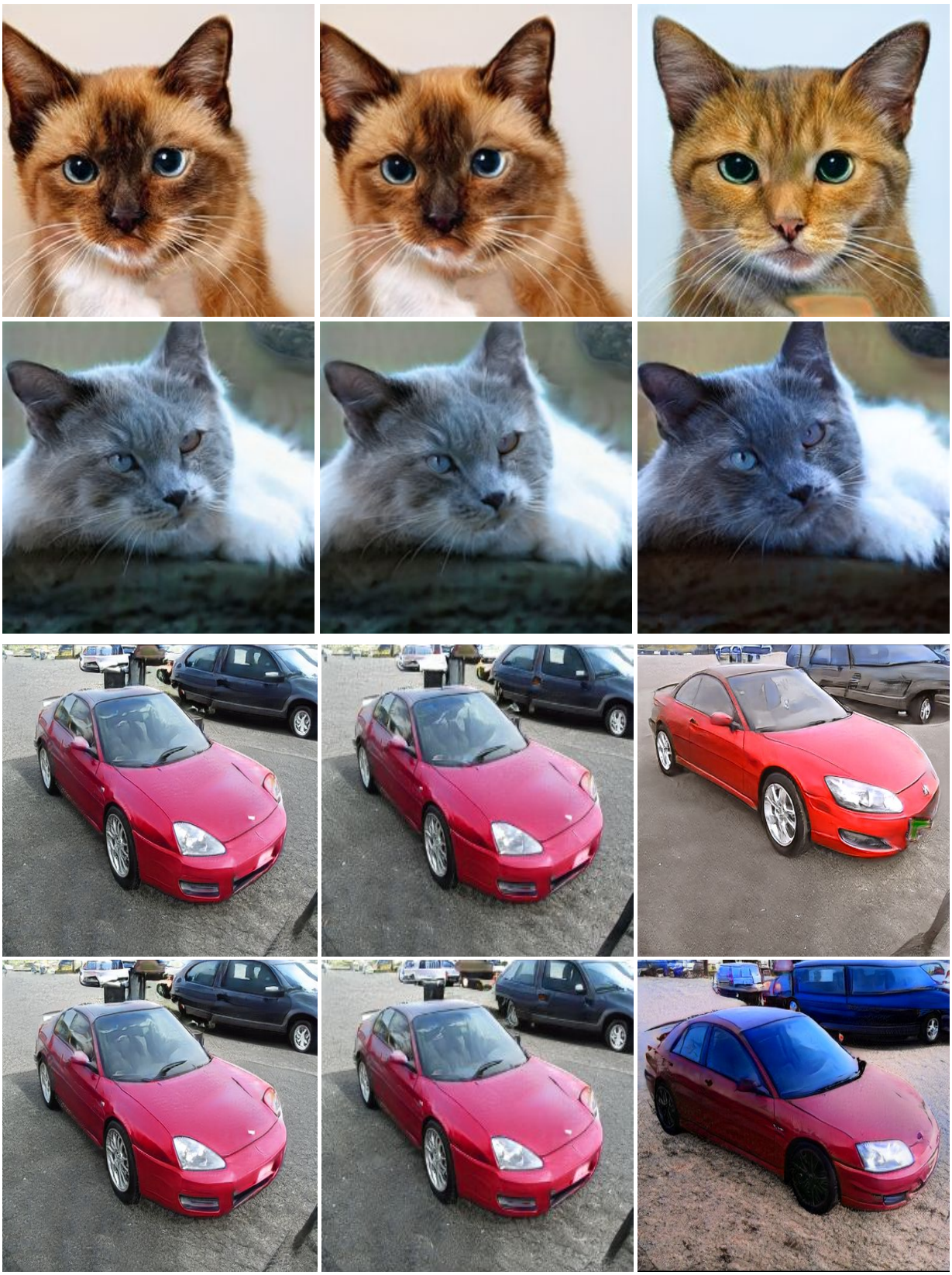


Figure 7: High resolution versions of adversarial images. From left to right: original, noise-based and style-based images.

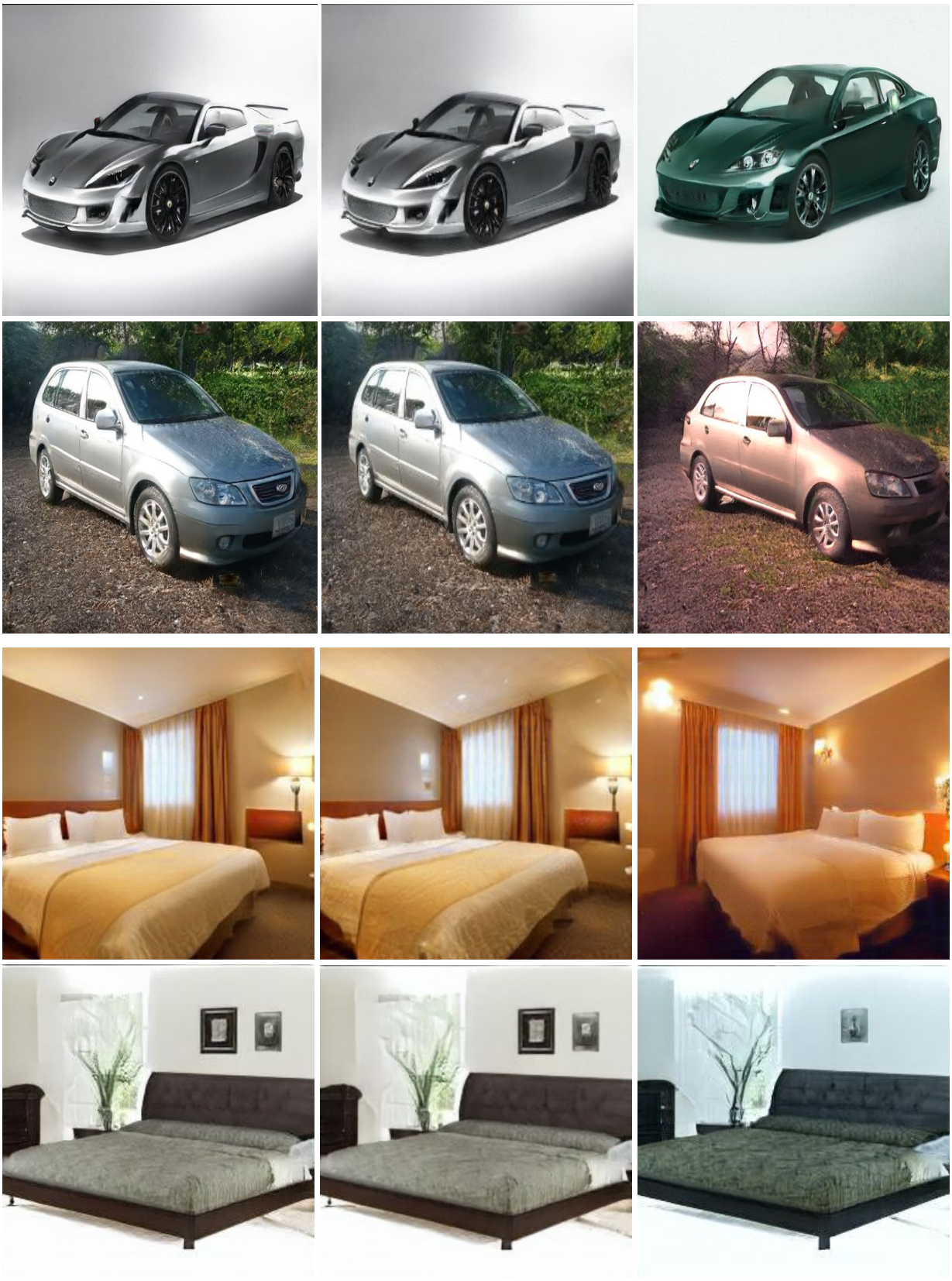


Figure 7: High resolution versions of adversarial examples. From left to right: original, noise-based and style-based images.

References

[1] Rima Alaifari, Giovanni S Alberti, and Tandri Gauksson. Adef: An iterative algorithm to construct adversarial deformations. *arXiv preprint arXiv:1804.07729*, 2018. 1, 2, 3

[2] Michael A Alcorn, Qi Li, Zhitao Gong, Chengfei Wang, Long Mai, Wei-Shinn Ku, and Anh Nguyen. Strike (with) a pose: Neural networks are easily fooled by strange poses of familiar objects. *arXiv preprint arXiv:1811.11553*, 2018. 1, 2

[3] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. *arXiv preprint arXiv:1802.00420*, 2018. 2, 7

[4] Shumeet Baluja and Ian Fischer. Learning to attack: Adversarial transformation networks. In *AAAI*, 2018. 2

[5] Jianmin Bao, Dong Chen, Fang Wen, Houqiang Li, and Gang Hua. Cvae-gan: fine-grained image generation through asymmetric training. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2745–2754, 2017. 3

[6] Tom B Brown, Nicholas Carlini, Chiyuan Zhang, Catherine Olsson, Paul Christiano, and Ian Goodfellow. Unrestricted adversarial examples. *arXiv preprint arXiv:1809.08352*, 2018. 2

[7] Tom B Brown, Dandelion Mané, Aurko Roy, Martín Abadi, and Justin Gilmer. Adversarial patch. *arXiv preprint arXiv:1712.09665*, 2017. 1, 2

[8] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pages 3–14. ACM, 2017. 7

[9] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pages 39–57. IEEE, 2017. 1, 2

[10] Jeremy M Cohen, Elan Rosenfeld, and J Zico Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019. 2, 7, 8

[11] 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. 4

[12] Emily L Denton, Soumith Chintala, Rob Fergus, et al. Deep generative image models using a laplacian pyramid of adversarial networks. In *Advances in neural information processing systems*, pages 1486–1494, 2015. 2

[13] 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. 2

[14] Krishnamurthy Dvijotham, Robert Stanforth, Sven Gowal, Timothy Mann, and Pushmeet Kohli. A dual approach to scalable verification of deep networks. *arXiv preprint arXiv:1803.06567*, 104, 2018. 2

[15] Logan Engstrom, Brandon Tran, Dimitris Tsipras, Ludwig Schmidt, and Aleksander Madry. A rotation and a translation suffice: Fooling cnns with simple transformations. *arXiv preprint arXiv:1712.02779*, 2017. 1, 2, 3

[16] Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013. 1

[17] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018. 2

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 2, 8

[19] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. *arXiv preprint arXiv:1711.00117*, 2017. 2

[20] Seunghoon Hong, Xinchen Yan, Thomas S Huang, and Honglak Lee. Learning hierarchical semantic image manipulation through structured representations. In *Advances in Neural Information Processing Systems*, pages 2708–2718, 2018. 3

[21] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1501–1510, 2017. 3

[22] Xun Huang, Yixuan Li, Omid Poursaeed, John Hopcroft, and Serge Belongie. Stacked generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5077–5086, 2017. 2

[23] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European conference on computer vision*, pages 694–711. Springer, 2016. 4

[24] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018. 2

[25] 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. 4

[26] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. *arXiv preprint arXiv:1812.04948*, 2018. 1, 2, 3, 5

[27] Mohammadhadi Kiapour, Shuai Zheng, Robinson Piramuthu, and Omid Poursaeed. Generating a digital image using a generative adversarial network, Sept. 19 2019. US Patent App. 15/923,347. 2

[28] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 1

[29] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016. 2, 3, 8

[30] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989. 2

[31] Xin Li and Fuxin Li. Adversarial examples detection in deep networks with convolutional filter statistics. In *Proceedings*

of the IEEE International Conference on Computer Vision, pages 5764–5772, 2017. 2

[32] Yingwei Li, Song Bai, Cihang Xie, Zhenyu Liao, Xiaohui Shen, and Alan L Yuille. Regional homogeneity: Towards learning transferable universal adversarial perturbations against defenses. *arXiv preprint arXiv:1904.00979*, 2019. 2

[33] Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1778–1787, 2018. 2

[34] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pages 3730–3738, 2015. 2, 4

[35] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 2, 8

[36] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: a simple and accurate method to fool deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2574–2582, 2016. 1

[37] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. Reading digits in natural images with unsupervised feature learning. 2011. 2

[38] Augustus Odena, Christopher Olah, and Jonathon Shlens. Conditional image synthesis with auxiliary classifier gans. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2642–2651. JMLR. org, 2017. 2

[39] Nicolas Papernot, Patrick McDaniel, Somesh Jha, Matt Fredrikson, Z Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *2016 IEEE European Symposium on Security and Privacy (EuroS&P)*, pages 372–387. IEEE, 2016. 2

[40] Omid Poursaeed, Isay Katsman, Bicheng Gao, and Serge Belongie. Generative adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4422–4431, 2018. 2

[41] Omid Poursaeed, Vladimir G Kim, Eli Shechtman, Jun Saito, and Serge Belongie. Neural puppet: Generative layered cartoon characters. *arXiv preprint arXiv:1910.02060*, 2019. 2

[42] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8571–8580, 2018. 2

[43] Aditi Raghunathan, Jacob Steinhardt, and Percy S Liang. Semidefinite relaxations for certifying robustness to adversarial examples. In *Advances in Neural Information Processing Systems*, pages 10877–10887, 2018. 2

[44] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. *arXiv preprint arXiv:1805.06605*, 2018. 2

[45] Krishna Kumar Singh, Utkarsh Ojha, and Yong Jae Lee. Finegan: Unsupervised hierarchical disentanglement for fine-grained object generation and discovery. *arXiv preprint arXiv:1811.11155*, 2018. 2

[46] Yang Song, Rui Shu, Nate Kushman, and Stefano Ermon. Constructing unrestricted adversarial examples with generative models. In *Advances in Neural Information Processing Systems*, pages 8312–8323, 2018. 2

[47] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2818–2826, 2016. 4

[48] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[49] Florian Tramèr, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick McDaniel. Ensemble adversarial training: Attacks and defenses. *arXiv preprint arXiv:1705.07204*, 2017. 8

[50] Yusuke Tsuzuku, Issei Sato, and Masashi Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. In *Advances in Neural Information Processing Systems*, pages 6541–6550, 2018. 2

[51] Jonathan Uesato, Brendan O'Donoghue, Aaron van den Oord, and Pushmeet Kohli. Adversarial risk and the dangers of evaluating against weak attacks. *arXiv preprint arXiv:1802.05666*, 2018. 2

[52] Vikas Verma, Alex Lamb, Christopher Beckham, Aaron Courville, Ioannis Mitliagkis, and Yoshua Bengio. Manifold mixup: Encouraging meaningful on-manifold interpolation as a regularizer. *arXiv preprint arXiv:1806.05236*, 2018. 8

[53] Lipeng Wan, Jun Wan, Yi Jin, Zichang Tan, and Stan Z Li. Fine-grained multi-attribute adversarial learning for face generation of age, gender and ethnicity. In *2018 International Conference on Biometrics (ICB)*, pages 98–103. IEEE, 2018. 3

[54] Eric Wong and J Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. *arXiv preprint arXiv:1711.00851*, 2017. 2

[55] Chaowei Xiao, Bo Li, Jun-Yan Zhu, Warren He, Mingyan Liu, and Dawn Song. Generating adversarial examples with adversarial networks. *arXiv preprint arXiv:1801.02610*, 2018. 2

[56] Chaowei Xiao, Jun-Yan Zhu, Bo Li, Warren He, Mingyan Liu, and Dawn Song. Spatially transformed adversarial examples. In *International Conference on Learning Representations*, 2018. 1, 2, 3

[57] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017. 2

[58] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. *arXiv preprint arXiv:1812.03411*, 2018. 2

[59] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1316–1324, 2018. 3

[60] Jianwei Yang, Anitha Kannan, Dhruv Batra, and Devi Parikh. Lr-gan: Layered recursive generative adversarial networks for image generation. *arXiv preprint arXiv:1703.01560*, 2017. 2

[61] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015. 4

[62] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. *arXiv preprint arXiv:1710.09412*, 2017. 8

[63] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 5907–5915, 2017. 2