ASSESSING NEURAL NETWORK ROBUSTNESS VIA ADVERSARIAL PIVOTAL TUNING OF REAL IMAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to assess the robustness of image classifiers to a diverse set of manipulations is essential to their deployment in the real world. Recently, semantic manipulations of real images have been considered for this purpose, as they may not arise using standard adversarial settings. However, such semantic manipulations are often limited to *style*, *color* or *attribute* changes. While expressive, these manipulations do not consider the full capacity of a pretrained generator to affect adversarial image manipulations. In this work, we aim at leveraging the full capacity of a pretrained image generator to generate highly detailed, diverse and photorealistic image manipulations. Inspired by recent GAN-based image inversion methods, we propose a method called Adversarial Pivotal Tuning (APT). APT first finds a pivot latent space input to a pretrained generator that best reconstructs an input image. It then adjusts the weights of the generator to create small, but semantic, manipulations which fool a pretrained classifier. Crucially, APT changes both the input and the weights of the pretrained generator, while preserving its expressive latent editing capability, thus allowing the use of its full capacity in creating semantic adversarial manipulations. We demonstrate that APT generates a variety of semantic image manipulations, which preserve the input image class, but which fool a variety of pretrained classifiers. We further demonstrate that classifiers trained to be robust to other robustness benchmarks, are not robust to our generated manipulations and propose an approach to improve the robustness towards our generated manipulations.

1 INTRODUCTION

Significant progress has been made in developing classifiers that work reliably in a broad range of data distributions Akhtar et al. (2021) and which are robust to corruption methods. To assess those classifiers, several benchmarks have been proposed. A large body of work considers robustness against adversarial l_p-bounded pixel-space perturbations. Since such perturbations act on raw pixels, they do not result in *semantic manipulations* such as changes in lighting conditions.

Recently, a new generation of models that can generate highly expressive and photorealistic images has gotten much attention, these include DALL-E-2 Ramesh et al. (2022), RQ-VAE Lee et al. (2022), Stable-Diffusion Rombach et al. (2021), among others. In particular, these models can be used to manipulate existing images with a high degree of detail and expressivity. In the context of neural network robustness, a recent line of work considers the ability to use generative models to generate class-preserving semantic adversarial manipulations Song et al. (2018); Xu et al. (2020); Poursaeed et al. (2021); Gowal et al. (2020), overcoming the limitations of the abovementioned pixel-space perturbations. However, such manipulations are often restricted to specific style or color changes. While such manipulations are challenging, they fall short of covering the entire space of possible class-preserving semantic manipulations.

In this paper, we aim to address this shortcoming by asking the following research question: can we leverage the full expressive power of a pretrained image generator to perform more general, highly detailed, photorealistic image manipulations for assessing the robustness of image classifiers? Given a pretrained classifier C and a pretrained generator G, we wish to perform manipulations on a given set of images such that: (i) the resulting images are within the original dataset distribution, (ii) the

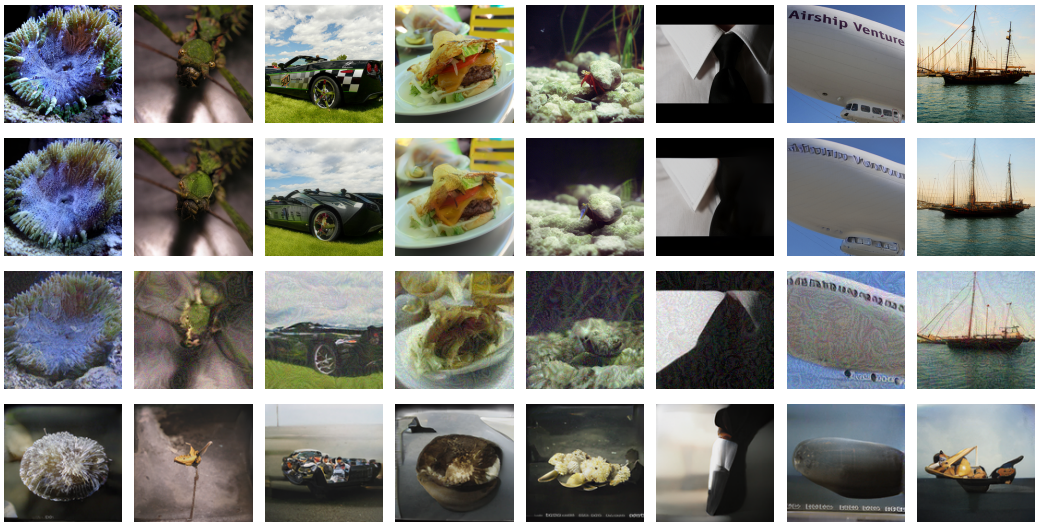

Figure 1: **Generated manipulations. Row 1** shows the input images. **Row 2** shows the images resulting from our manipulations. **Row 3** shows the result of Lin et al. (2020), using pixel-space adversarial manipulations applied to StyleGAN-XL's reconstructions. **Row 4** shows the result of Lin et al. (2020) with latent space manipulates applied using StyleGAN-XL. Our method manipulates images in a non-trivial but class-preserving manner, using the full capacity of a pretrained StyleGAN generator. For example, it removes the eye of the mantis (second column), changes the type of race car (third column), changes the color of the crab tail (fifth column), removes the text in a spaceship (seventh column) and removes some of the ropes (eighth column). All of these are class-preserving examples that fool a pretrained PRIME-ResNet50 Modas et al. (2021) classifier. In contrast, Lin et al. (2020) either generates noisy and less realistic images (row 3) or images which differ significantly semantically and which do not preserve the input class (row 4).

manipulations are class-preserving, (iii) they fool the target classifier C, and (iv) they are highly expressive, i.e., the full capacity of the generator G is used to perform such manipulations.

We focus on the robustness of ImageNet classifiers and use the recently proposed StyleGAN-XL (Sauer et al., 2022) generator, as it offers the ability to effectively manipulate style and content semantically. Our approach, *Adversarial Pivotal Tuning (APT)*, is inspired by Pivotal Tuning Inversion (PTI), a recent GAN inversion approach proposed by Roich et al. (2021). Given an input image x, we first perform latent optimization to find the input pivot latent vector w_p that results in the closest (but imperfect) reconstruction to x. We subsequently optimize the StyleGAN weights with the following objectives: (1) reconstructing image x, (2) fooling the classifier C, and (3) ensuring the generated image appears real to the pretrained StyleGAN discriminator, so that the generated image remains within the real image distribution. To ensure that the manipulations are class-preserving, we bound the maximum distance between input and generated image and stop the optimization when this distance is reached.

Our method enjoys a number of advantages: First, by generating images which are close to input images and which appear realistic to a pretrained discriminator D, they are likely to be of high quality and fidelity as well as class-preserving. Second, by using objective (2), images are likely to fool the classifier. Third, and most importantly, our manipulations are optimized over the entire space of StyleGAN parameters (both in the latent space and in its weights). By applying this optimization after an initial latent optimization stage, we ensure that the editing capabilities of StyleGAN are preserved, thus allowing for fully expressive manipulations.

We use our generated manipulations to assess the robustness of a variety of pretrained classifiers with a diverse range of architectures, as well as classifiers specifically trained to be robust against common corruptions and perturbations. Our results show a significant performance drop on our image manipulations, and that these adversarial manipulations are transferable, indicating that the tested classifiers are not robust to them. Visually (see Fig. 1), we observe a wide variety of image manipulations going beyond style transfer or changes in specific attributes. We subsequently consider an approach to improve the robustness to our manipulations through adversarial training on images that have been manipulated using our approach.

2 Related Work

Semantic Adversarial Robustness. The majority of current literature considers adversarial robustness to pixel-space manipulations where the l_p norm is bounded Szegedy et al. (2013); Fletcher (2013). For a comprehensive review, see Akhtar et al. (2021). We focus on approaches that semantically manipulate the input image, resulting in a naturally looking adversarial manipulation.

One set of works considers a specific class of semantic manipulations. These include geometric changes Xiao et al. (2018); Alaifari et al. (2018); Engstrom et al. (2017), view changes Alcorn et al. (2018), manipulating intermediate classifier features Dunn et al. (2020); Laidlaw et al. (2020); Xu et al. (2020), and inserting patches Brown et al. (2017). Hendrycks et al. (2021) consider an image filtering approach of natural images. Other works consider manipulation of style, texture or color statistics, where the structure of the image is fixed. Hosseini & Poovendran (2018) convert images to HSV and change the hue and saturation. Bhattad et al. (2019) consider unrestricted image perturbations by either manipulating the image color or texture in an adversarial manner. Shamsabadi et al. (2020) selectively modify the image's color within a chosen range that appears natural to humans.

Another set of works considers adversarial manipulation of facial attributes. Joshi et al. (2019) control a binary attribute such as glasses. Qiu et al. (2020) propose an attribute-conditional generative model producing adversarial examples which differ from the input image by one attribute. One can also consider deepfakes Tolosana et al. (2020) as class-preserving semantic manipulations.

Another line of work considers the use of pretrained generative models. Song et al. (2018) search the latent space of a pretrained AC-GAN to find inputs that fool a given classifier. Unlike our method, they do not manipulate real images, resulting in less realistic generations and less faithful matching of the real image distribution—a result of AC-GAN's mode-dropping. Xu et al. (2020) consider an autoencoder-based manipulation of real images, but it is restricted to style changes.

Gowal et al. (2020) demonstrate an approach for adversarial training with samples generated by StyleGAN. However, it only manipulates a subset of the latent space variables, limiting the set of manipulations to coarse image changes. Moreover, our approach considers higher-resolution ImageNet samples while their approach is limited to low-resolution faces or MNIST digits. Lin et al. (2020) project images to a pretrained StyleGAN's latent space and adversarially manipulate their style code. Similarly, Poursaeed et al. (2021) manipulate both the style and noise latent vectors of StyleGAN. Our work takes a step further and manipulates not only the latent space of StyleGAN, but also its weights while preserving its editing capabilities. We thus enable the full utilization of StyleGAN's capacity to create highly expressive semantic manipulations, as shown in Fig. 1.

GAN Inversion and Image Manipulation. Our work is inspired by recent pretrained GAN inversion methods for effective manipulation of images. Some works optimize the latent space of a pretrained GAN Lipton & Tripathi (2017); Creswell & Bharath (2018); Abdal et al. (2019); Karras et al. (2020) or use an encoder to find the latent input for a given image, such that the input image is effectively reconstructed Perarnau et al. (2016); Luo et al. (2017); Guan et al. (2020). In the context of StyleGAN, Abdal et al. (2020) have shown that optimizing over StyleGAN's latent input space $\mathcal{W}$ results in unfaithful reconstructions. When considering optimization over the $\mathcal{W}+$ space, latent manipulations are inferior compared to the same manipulations over StyleGAN's $\mathcal{W}$ space. To this end, Roich et al. (2021) proposed to directly update StyleGAN's weights, following an initial latent optimization step. Unlike these methods, our object is not to invert an input image, but rather to fool a classifier.

3 Adversarial Pivotal Tuning

We now describe our proposed approach for generating adversarial images for a pretrained classifier C, utilizing the full expressive capacity of StyleGAN-XL. Given a collection of images from ImageNet-1k, we first filter those misclassified by C. In order to fool C on a correctly classified image, we wish to semantically manipulate it to be misclassified by C. Simple color jittering, rotation, translations and semantically generated manipulations such as style, texture or specific attribute change, can result in misclassification but remain limited in scope and realism. We therefore suggest a new method, Adversarial Pivotal Tuning(APT), that learns non-trivial and highly non-linear image

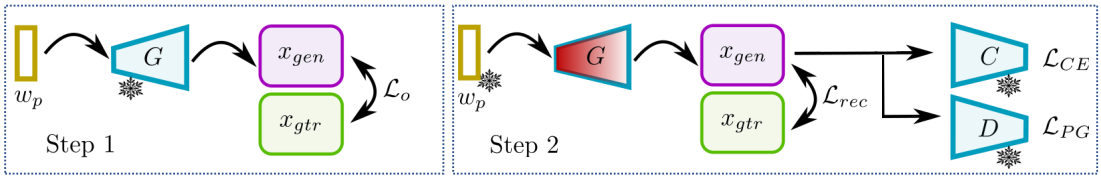

Figure 2: **The Adversarial Pivotal Tuning (APT) framework.** In the first step, we optimize a style code w_p using standard latent optimization $\mathcal{L}_o$ from Eq. (1), while keeping the generator G frozen. The loss is computed between the ground-truth image x_{gtr} and the generated image x_{gen}. In the second step, we freeze w_p and finetune G (shown in red) using the three objectives from Eq. (8); a reconstruction objective $\mathcal{L}_{rec}$, the projected GAN objective using the discriminator D, $\mathcal{L}_{PG}$, and our fooling objective $\mathcal{L}_{CE}$ using the classifier C. A $*$ is used to indicate a frozen component.

manipulations, while simultaneously ensuring the generated image stays within the data manifold. An overview of the APT method is shown in Fig. 2.

We use StyleGAN-XL Sauer et al. (2022), a generative model trained on ImageNet-1K. The generator G consists of a mapping network G_m and a synthesis network G_s. The mapping network maps a random Gaussian latent variable $z \in \mathbb{R}^{64}$ along with a one-hot class label c to the style code $w \in \mathbb{R}^{27\times512} = \mathcal{W}$. The synthesis network subsequently maps w and a noise vector n to an RGB image $\hat{x} \in \mathbb{R}^{3\times H\times W}$ of height $H = 256$ and width $W = 256$. This generator is subsequently trained to fool a set of discriminators $\{D_l\}$ using the Projected GAN objective Sauer et al. (2021).

The first step of our method, aims at identifying a latent code w (and noise vector n) that minimizes the reconstruction error between a generated image x_{gen} and a given input image x_{gtr}, for a pre-trained generator, in a similar manner to GAN inversion methods. This is done using the process of latent optimization over w_p, n:

$$\arg\min_{w,n} \mathcal{L}_{\text{LPIPS}}\left(x_{gtr}, G_s\left(w, n; \theta\right)\right) + \lambda_n \mathcal{L}_n(n) \tag{1}$$

Here, $x_{gen} = G_s(w, n; \theta)$ is the image produced by a *pre-trained* synthesis network G_s parameterized by weights θ. We follow Roich et al. (2021) and Karras et al. (2020) in using a noise regularization term $\mathcal{L}_n$ and use λ_n as a hyperparameter. The optimization is performed in $\mathcal{W}$ space and $\mathcal{L}_{\text{LPIPS}}$ is the perceptual distance introduced in Zhang et al. (2018).

In the second step, we modify the image x, to fool the classifier C, utilizing the full capacity of StyleGAN. We note that $G_s(w_p, n; \theta)$, i.e., the initial estimate for the reconstruction of x, should not be far from the adversarially manipulated image $\hat{y}$ we wish to generate.

We first consider the reconstruction objective, as in Eq. (2), as we wish our manipulated image to be close to the input image x. Similarly to Roich et al. (2021), the generator weights are adjusted and regularized to restrict changes to a local region in the latent space, while the latent code $w_p \in \mathcal{W}$ and noise n are fixed, leading to better reconstruction:

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}_{rec}(x, G_s(w_p, n; \theta)), \tag{2}$$

where $\hat{\theta}$ represents the new fine-tuned weights. The reconstruction loss is defined as follows:

$$\mathcal{L}_{rec} = L_{pt} + \mathcal{L}_R \tag{3}$$

$$\mathcal{L}_R = \mathcal{L}_{\text{LPIPS}}\left(x_r, x_r^*\right) + \lambda_{L2}^R \mathcal{L}_{L2}\left(x_r, x_r^*\right) \tag{4}$$

$$\mathcal{L}_{pt} = \mathcal{L}_{\text{LPIPS}}\left(x, x_p^*\right) + \lambda_{L2}^P \mathcal{L}_{L2}\left(x, x_p^*\right) \tag{5}$$

where x_p^* is generated using the modified weights as $G_s(w_p, n; \hat{\theta})$. A locality regularization term ($\mathcal{L}_R$) is applied by restricting changes to a local region in the latent space. Specifically, setting $w_r = w_p + \alpha \frac{w_z - w_p}{\|w_z - w_p\|_2}$, where, in each iteration, z is sampled from a normal distribution and w_z is obtained by applying the mapping network G_m to z and class c of the input image x and α is a hyperparameter. We then generate x_r and x_r^* using the initial and modified weights, $G_s(w_r, n; \theta)$ and $G_s(w_r, n; \hat{\theta})$ respectively.

Secondly, we wish to fool the classifier C. That is, the cross entropy loss for the classifier's prediction on the manipulated image should be high. In practice, we observed a more stable optimization

when minimizing the cross entropy between the classifier's prediction and an incorrect class label chosen at random. Lastly, we utilize StyleGAN's pretrained discriminators $\{D_l\}$ to distinguish between real and synthetic images, and enforce that the manipulated image appears real. Following Sauer et al. (2021), we consider the following objective:

$$\mathcal{L}_{PG} = \sum_l \log\left(1 - D_l\left(G_s(w_p, n; \theta)\right)\right), \tag{6}$$

where the weights of each D_l are fixed. We then finetune G's weights θ with the following objective:

$$\mathcal{L}_{APT} = \mathcal{L}_{rec} + \lambda_{CE}\mathcal{L}_{CE}(c_{any}, C(G_s(w_p, n; \theta))) \tag{7}$$

$$+ \lambda_{PG}\mathcal{L}_{PG} \tag{8}$$

where c_{any} is a randomly chosen class different from the true class, and $\mathcal{L}_{CE}$ is the cross entropy loss. As both the classifier C and discriminators $\{D_l\}$ are fixed, the generated image is changed to match the reference image as closely as possible, while deviating only slightly to change the class predicted by the classifier. Given a desired maximum distance d, we consider generated images for which $\mathcal{L}_{pt} \leq d$, where d is a hyperparameter, and stop the optimization whenever $\mathcal{L}_{pt} \geq d$. We note that, unlike traditional frameworks that use a maximum l_p norm to bound adversarial examples, we consider $\mathcal{L}_{pt}$ which uses both a pixel-based distance and a perceptual distance.

3.1 Implementation Details

For the latent optimization step (Eq. (1)), we use the hyperparameters described by Karras et al. (2020). We run the optimization for $1k$ iterations. Our GAN generated images are of 256^2 resolution. Given a generated image, we follow Modas et al. (2021) in center cropping the image to 224^2 resolution and normalizing it using standard ImageNet statistics, before being classified by our pretrained classifier. The hyperparameters in Eq. (8) are: $\lambda^P_{L2} = \lambda^R_{L2} = 0.1$, $\lambda_{CE} = 0.01$ and $\lambda_{PG} = 0.005$ for all experiments. For Eq. (8), we use the Adam optimizer with a learning rate of $3 \cdot 10^{-4}$. For evaluation, we follow the official codebase of the baselines. When finetuning on APT generated images, we follow our standard training configuration, but with learning rate of 0.001.

4 Experiments

We begin by assessing the degree to which our generated images (i) are within the ImageNet distribution, (ii) represent the same class as the corresponding input images (i.e., the manipulation is class-preserving), (iii) fool a target classifier, i.e., the classifier misclassifies the generated images, and (iv) exhibit a wide variety of semantic changes.

To this end, we consider a collection of pretrained classifiers including those specifically trained to be robust to different robustness benchmark datasets such as ImageNet-C, ImageNet-A, and ImageNet-R. More specifically, we consider PRIME-ResNet50 Modas et al. (2021) which is trained using new augmentation techniques for enhanced robustness, and FAN-VIT Zhou et al. (2022), a Vision Transformer with no MLP layers that is highly robust to unseen natural images.

Additionally, to test transferability, we use the adversarially generated samples using PRIME-ResNet50 and FAN-VIT as classifiers, and test them on other architectures: (1) ResNet50 He et al. (2015), (2) MAE He et al. (2021), a generalizable and scalable asymmetric encoder-decoder architecture, (3) RegNet-Y Goyal et al. (2022), a ResNet-type model with a regulatory model to extract complementary features, and (4) data2vec Baevski et al. (2022), a self-supervised transformer that predicts contextualized latent representations in a self-distillation setup for any modality. We then explore a training regime for improving model robustness to our APT manipulations. Lastly, we conduct an ablation study, illustrating the necessity of the different components for generating our samples, and investigating the effect of different hyperparameters.

4.1 Adversarially Generated Manipulations

We evaluate our generated samples with respect to properties (i)-(iv) above.

Table 1: User studies. (Q2). We conduct a user study to determine, for a real and generated samples, if the majority of 25 annotators considered the class to have changed or not. Similarly, for the pretrained classifier (Classif), we consider if its classification changed. We consider 40 such samples from the ImageNet-1k validation set extracted using a pretrained PRIME-ResNet50 classifier. (Q3). For our generated samples, real images and those of Lin et al. (2020), we display each image and ask the user to assign the corresponding label. The percentage of correct responses corresponding to the real image's class is shown.

	Classif (same cl.)	Classif (diff cl.)
Human (same cl.)	25	13
Human (diff cl.)	1	1

(Q2)

dual-L	dual-P	Ours	Real
12.5%	42.5%	**90.0%**	95.0%

(Q3)

Table 2: Accuracy (Acc) and mean softmax probability of the labeled class (Conf) on the ImageNet validation set (Real) and corresponding Generated images with APT using PRIME-ResNet50 Modas et al. (2021) and FAN-VIT Zhou et al. (2022) as classifiers respectively. Each model is evaluated on the generated images for which it was also used as the classifier in the APT generation.

Model	Real (Acc)	Real (Conf)	APT (Acc)	APT (Conf)
PRIME Resnet-50	77.1%	69.7	54.0%	23.4
FAN-VIT	83.6%	62.4	62.0%	44.7

Table 3: Transferability of APT generated samples. For the ImageNet-1k validation set, we consider samples generated to fool a PRIME-Resnet50 Modas et al. (2021) (PRIME) and a FAN-VIT Zhou et al. (2022) (FAN) pretrained classifier. We then test the accuracy (Acc) and mean softmax probability of the labelled class (Conf) on those samples. The left column indicates the classifier on which we tested the accuracy of real or generated samples. * indicates the accuracy and confidence for samples generated and tested using the same classifier.

Model	Real (Acc)	Real (Conf)	PRIME (Acc)	PRIME (Conf)	FAN (Acc)	FAN (Conf)
PRIME	77.1%	69.7	54.0%*	23.4*	60.1%	52.3
FAN-VIT	83.6%	62.4	70.9%	52.0	62.0%*	44.7*
Resnet-50	75.3%	68.6	60.9%	48.3	59.7%	51.5
MAE-ViT-B	83.2%	76.7	70.4%	62.6	61.2%	56.8
MAE-ViT-L	86.0%	78.8	71.8%	64.4	62.2%	58.1
MAE-ViT-H	87.0%	80.1	73.8%	65.4	62.3%	58.4
Regnet-320	83.1%	83.7	62.1%	63.8	54.5%	60.6
Regnet-1280	83.7%	77.2	61.3%	64.5	53.0%	61.3
data2vec	83.5%	77.7	70.6%	65.0	61.6%	59.1

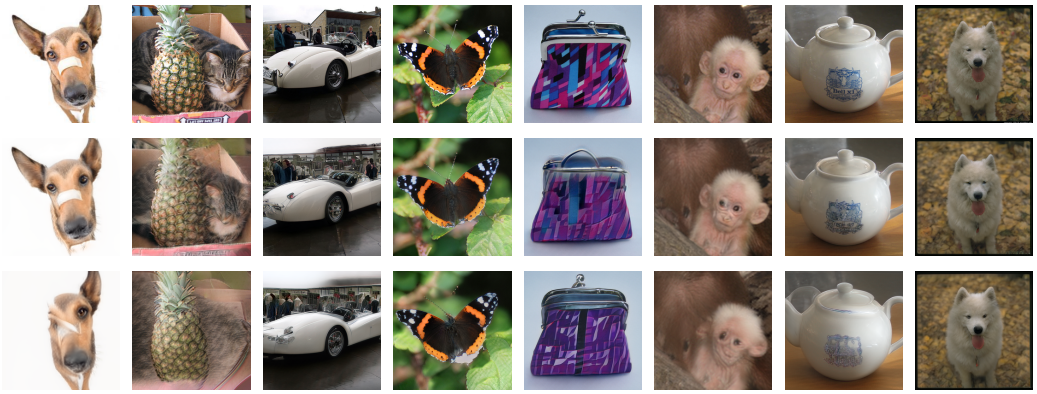

Figure 3: **Generated manipulations.** The top row shows input images, the middle row shows APT manipulations for a ResNet-50 classifier, and the bottom row shows APT manipulations from a FAN-VIT classifier. Column 1-4+7 illustrates similar manipulations for both classifiers, column 5-6 shows texture and spatial manipulations, the last column showcase a fooling image without a clear APT manipulation.

Fidelity and Diversity. To measure (i), i.e., whether samples lie on the ImageNet manifold, we are interested in measuring both *Fidelity*—whether the generated samples are high quality—and *Diversity*—whether the generated samples capture the diversity of the original real dataset. To this end, we consider the FID score Heusel et al. (2017), which was introduced as a metric to capture both *Fidelity* and *Diversity*. For a pretrained PRIME-ResNet50 classifier Modas et al. (2021), we consider three groups of images: (1) 3k images chosen at random from the ImageNet validation set, (2) their corresponding adversarial manipulations generated using APT, (3) 3k images chosen at random from the ImageNet training set. First, we evaluate the FID score between (1) and (2). As can be seen in Table 4, the value is lower than the other groups, indicating that the distributions are close. To evaluate the FID against non-matching groups of real images, we consider the FID between (1) and (3) and between (2) and (3).

As can be seen in Table 4, the FID value between (2) and (3) is only slightly higher than that of (1) and (3), indicating that our generated distribution matches the training image distribution in only a slightly worse manner than real validation images. We note that the trace of the covariance matrices contributes to the vast majority of the score, likely due to the low number of samples available for the validation set. To convince ourselves that this is the case, we also report the FID between the training set and their corresponding adversarial manipulations (43k) to be 6.62 indicating that the real and generated images are similar.

Table 4: **Top three rows:** FID scores using a PRIME-Resnet50 for our generated manipulations in comparison to manipulations generated by Lin et al. (2020) using pixel space manipulations (dual-P) on StyleGAN-XL's reconstructions or latent-space manipulations (dual-L). The same set of input images is used. **Fourth row:** FID scores for our generated samples using a FAN-VIT classifier. The FID score between real validation and training images from ImageNet ((1) & (3)) is 25.99.

	(1) & (2)	(2) & (3)
Ours (PRIME)	**19.87**	**23.72**
dual-P (PRIME)	63.63	92.86
dual-L (PRIME)	50.94	61.62
Ours (FAN-VIT)	20.01	24.24

As a point of comparison, we consider the generated manipulations (2) on the same set of images using Lin et al. (2020), either using pixel-space adversarial manipulations applied on StyleGAN-XL's reconstructions (Lin et al. (2020) dual-P) or with latent-space manipulates applied using StyleGAN-XL (Lin et al. (2020)-L). As can be seen, the generated samples are of a much worse FID score in comparison to our samples, indicating that they are of much lower generation quality and do not match ImageNet's real image distribution.

Class Preservation. To measure (ii), we consider, for a pretrained PRIME-ResNet50 classifier, whether generated samples are class-preserving. We conduct user studies consisting of 25 users and 40 samples from ImageNet's validation set and their corresponding samples generated with APT (Q1-3) and Lin et al. (2020) dual-P/L (Q3). We then conduct the following assessments:

- Q1: We display each generated sample in isolation, and ask how realistic it is, on a Likert scale of 1 (strongly disagree) to 5 (strongly agree).
- Q2: For each generated sample, we first display the real image and the associated class to the user. We then display the generated sample and ask whether the class is preserved. Additionally, we consider if the pretrained classifier misclassifies the generated image.
- Q3: We display each image and ask the user to assign a corresponding label. The label is considered correct if it corresponds to the real image's ground truth label. This is performed separately for real images, for our generates samples and for those generated by Lin et al. (2020).

For Q1, we report the mean score on the Likert-scale to be 3.55. For Q2, as seen in Table 1(a), user almost always state that the class is preserved, except in 5% of the cases, whereas 32.5% of these images fool the classifier. For Q3, as can be seen in Table 1(b), the generated samples by Song et al. (2018)-L/P exhibit significant loss of class identity. For APT and the real images, the users correctly predict 90% and 95% of the images respectively, suggesting that our method yields realistic and class-preserving images.

Classifier Fooling. To test the degree to which a target classifier is fooled, we measure its accuracy on 3k images from the ImageNet1k validation set and on the corresponding images generated by APT. Each class contains 3 randomly sampled images from the validation set. Additionally, we measure the average decrease of the softmax probability for the real class, to as-

Table 6: ImageNet-C (CC) accuracy for PRIME-Resnet50 before and after finetuning.

Method	CC	Noise			Blur				Weather				Digital			
		Gauss.	Shot	Impulse	Defoc.	Glass	Motion	Zoom	Snow	Frost	Fog	Bright.	Contr.	Elastic	Pixel.	JPEG
Before	54.5	59.7	58.6	58.0	47.6	39.0	48.4	46.0	47.0	50.3	53.8	71.5	58.2	56.3	59.5	62.2
After	54.4	59.4	58.4	58.6	48.4	38.8	50.5	49.3	44.7	48.6	60.5	71.2	58.7	55.6	55.3	58.4

sess the decrease in confidence of the classifier on the real class. As shown in Table 2, the accuracy drops by as much as 23.1%, down to a level comparable to the ImageNet-C accuracy. Similarly, in Table 3 we assess whether our APT samples are transferable. That is, whether images generated by APT using PRIME-ResNet50 and FAN-VIT classifiers fool other classifiers. Our fooling samples are transferable and the performance on different classifiers also drops. We observe that the accuracy on FAN-VIT APT images drops further and that in general higher confidence leads to larger drops of confidence on the generated images.

Table 5: Average accuracy and confidence on APT samples using PRIME-ResNet50 before and after fine-tuning.

	Accuracy	Confidence
Before	54.0	23.4
After	57.5	42.0

Diversity of Manipulations. In addition to our diverse manipulations shown in Fig. 1, in Fig. 3, we show for the same images, model-dependent APT manipulations which fool either a ResNet50 or a FAN-VIT classifier. Interestingly, for FAN-VIT, other manipulations like texture or spatial transformations (column 5-6) are more present, in addition to more prevalent versions of the same manipulation as for the ResNet50 classifier (column 1-4), or no clear manipulation (column 8).

4.2 Improving Robustness to Generates Samples

We now consider whether APT can be used to improve robustness. To this end, for a PRIME-ResNet50 classifier C, we use APT to manipulate 50k images from the ImageNet training set, fine-tune the classifier on these resulting images, resulting in classifier $C_{finetune}$. We then consider 3k images from the ImageNet validation set, and generate APT samples for both C and $C_{finetune}$. In Table 5, we observe that accuracy on APT generated images increases by 3.7% after fine-tuning.

Performance on Corruption Benchmarks. We consider the performance of our APT-based fine-tuning on previously proposed corruption-based robustness benchmarks. We consider the standard robustness benchmark of ImageNet-C Hendrycks & Dietterich (2019) which is generated by applying simple corruptions to ImageNet, such as color jittering, noise, and blurring. While fine-tuning improves robustness to APT samples, as can be seen in Table 6, the performance remains almost the same on ImageNet-C. We note however that some corruptions such as Fog, Motion, and zoom blur have improved, while noise and digital transformations did not. This is likely due to the fact that APT manipulations do not include these types of manipulations.

4.3 Ablation Study

First, we consider the effect of removing each component our APT objective (Eq. (8)), using a PRIME-ResNet50 classifier. Fig. 4 illustrates examples of generated images with one of the components removed: $\mathcal{L}_{rec}$ (reconstruction), $\mathcal{L}_{PG}$ (discriminator realness) and $\mathcal{L}_{CE}(c_{any}; C(G_\theta(w_p)))$ (fooling objective). Lastly, we consider applying the optimization of Eq. (8) while modifying the latent space and leaving the generator's parameters fixed. When the reconstruction or the discriminator realness components are removed, we observe worse image quality. To measure the effect of each setup we record the number of images that fool the classifier, and observe that when the fooling objective is removed, only one out of the seven shown samples fool the pretrained classifier, whereas all seven samples fool the classifier otherwise.

Reconstruction vs. fooling trade-off. The maximum distance d allowed between the input and generated sample before the optimization is stopped (see Section 3) is an important hyperparameter which must be chosen carefully. Increasing this distance may allow for more expressive adversarial manipulations, but this may also result in a change of label for the image. Empirically, we found that $d = 0.2$ avoids a change of class. We investigate the effect of varying this value in $\{0.2, 0.3, 0.4\}$ and show example generations in Fig. 5. We note that the images tend to lose detail with higher values of d, which stems from the fact the reconstruction is poorer. Nonetheless, more diverse manipulations are possible, such as the removal of the antennae on the butterfly.

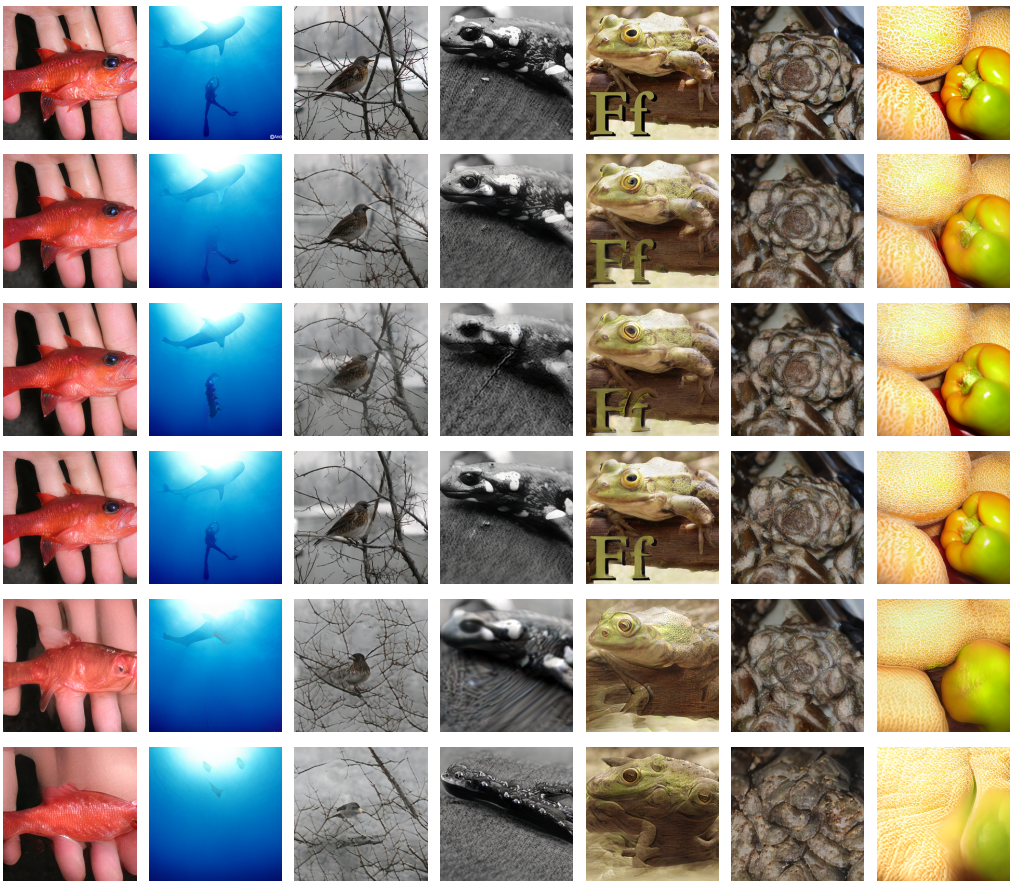

Figure 4: **APT sample generation ablation.** The first row shows the input images. The second row shows APT samples generated using our full objective $\mathcal{L}_{APT}$, all of which fool the pretrained PRIME-ResNet50 classifier. In the third row, we consider $\mathcal{L}_{APT}$ without the reconstruction loss ($\mathcal{L}_{rec}$). In the fourth row, we consider $\mathcal{L}_{APT}$ without the fooling objective ($\mathcal{L}_{CE}$). Without $\mathcal{L}_{CE}$, all but one sample fool the classifier. In the fifth row, we consider $\mathcal{L}_{APT}$ without the discriminator loss ($\mathcal{L}_{PG}$). The sixth row considers $\mathcal{L}_{APT}$ where only the latent space is optimised (generator's parameters fixed), resulting in loss of class preservation.

Figure 5: **APT generation for various distance d cutoff values**. The leftmost image shows the input image. We increase the maximum distance d to 0.2, 0.3 and 0.4 respectively, for a PRIME-ResNet50 classifier.

Figure 6: Generated samples by Song et al. (2018) using StyleGAN-XL. Our generated samples are more realistic and class preserving.

Randomly generated samples. We consider randomly generated samples which are not manipulations of real images. To this end, we adapt the method of Song et al. (2018) to use StyleGAN-XL. We perform latent optimization so as to fool a PRIME-Resnet50 classifier. The results are visualized in Fig. 6 and the image classes are the same as in Fig. 1. We note that the images are not class-preserving.

5 CONCLUSION

We have presented Adversarial Pivotal Tuning, a framework for generating highly expressive adversarial manipulations of real images. In a sense, we break with the common assumption that robustness benchmarks are not model specific, or in other words, allow for conducting a new type of robustness study tailored around fooling a particular classifier specifically well. This is achieved by leveraging the full capacity of StyleGAN-XL in generating highly detailed and diverse manipulations.

We have demonstrated that current robust classifiers, be it ResNets or Vision Transformers, are vulnerable to this new type of attack. As it turns out, it is possible to also boost performance by using the same framework to create training images as an additional type of augmentation. We have shown that APT can successfully be applied both as a way to fool classifiers and as a training framework to improve robustness. We envision this setup will inspire a new line of robustness research that will improve classifiers' robustness to models capable of generating photo-realistic images.

6 ETHICAL STATEMENT

Image manipulation methods are key to developing robust and well performing computer visions systems. At the same time, they also have a potential for being used to circumvent automated systems that have been deployed to catch illegal or otherwise questionable content. It is our belief that research on this topic should be done actively in the open to ensure that the technology to develop robust systems is widely disseminated.

7 REPRODUCIBILITY STATEMENT

We aim to make all the work presented reproducible by providing details of the architecture, in Section 3, along with the experimental setup described in Section 4. We will publicly release the code used along with the generated images and trained models.

REFERENCES

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan: How to embed images into the stylegan latent space? In *Proceedings of the IEEE international conference on computer vision*, pp. 4432–4441, 2019.

Rameen Abdal, Yipeng Qin, and Peter Wonka. Image2stylegan++: How to edit the embedded images? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8296–8305, 2020.

Naveed Akhtar, Ajmal Mian, Navid Kardan, and Mubarak Shah. Advances in adversarial attacks and defenses in computer vision: A survey. *IEEE Access*, 9:155161–155196, 2021.

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

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.

Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. *arXiv preprint arXiv:2202.03555*, 2022.

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

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

Antonia Creswell and Anil Anthony Bharath. Inverting the generator of a generative adversarial network. *IEEE transactions on neural networks and learning systems*, 30(7):1967–1974, 2018.

Isaac Dunn, Laura Hanu, Hadrien Pouget, Daniel Kroening, and Tom Melham. Evaluating robustness to context-sensitive feature perturbations of different granularities. *arXiv preprint arXiv:2001.11055*, 2020.

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.

Roger Fletcher. *Practical methods of optimization*. John Wiley & Sons, 2013.

Sven Gowal, Chongli Qin, Po-Sen Huang, Taylan Cemgil, Krishnamurthy Dvijotham, Timothy Mann, and Pushmeet Kohli. Achieving robustness in the wild via adversarial mixing with disentangled representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1211–1220, 2020.

Priya Goyal, Quentin Duval, Isaac Seessel, Mathilde Caron, Ishan Misra, Levent Sagun, Armand Joulin, and Piotr Bojanowski. Vision models are more robust and fair when pretrained on uncurated images without supervision, 2022. URL https://arxiv.org/abs/2202.08360.

Shanyan Guan, Ying Tai, Bingbing Ni, Feida Zhu, Feiyue Huang, and Xiaokang Yang. Collaborative learning for faster stylegan embedding. *arXiv preprint arXiv:2007.01758*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *CoRR*, abs/1512.03385, 2015. URL http://arxiv.org/abs/1512.03385.

Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. *arXiv:2111.06377*, 2021.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *Proceedings of the International Conference on Learning Representations*, 2019.

Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15262–15271, 2021.

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. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pp. 6629–6640, Red Hook, NY, USA, 2017. Curran Associates Inc. ISBN 9781510860964.

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

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

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8110–8119, 2020.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *arXiv preprint arXiv:2006.12655*, 2020.

Doyup Lee, Chiheon Kim, Saehoon Kim, Minsu Cho, and Wook-Shin Han. Autoregressive image generation using residual quantization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11523–11532, 2022.

Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. *Advances in Neural Information Processing Systems*, 33:3487–3498, 2020.

Zachary C Lipton and Subarna Tripathi. Precise recovery of latent vectors from generative adversarial networks. *arXiv preprint arXiv:1702.04782*, 2017.

Junyu Luo, Yong Xu, Chenwei Tang, and Jiancheng Lv. Learning inverse mapping by autoencoder based generative adversarial nets. In *International Conference on Neural Information Processing*, pp. 207–216. Springer, 2017.

Apostolos Modas, Rahul Rade, Guillermo Ortiz-Jiménez, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Prime: A few primitives can boost robustness to common corruptions. *arXiv preprint arXiv:2112.13547*, 2021.

Guim Perarnau, Joost Van De Weijer, Bogdan Raducanu, and Jose M Álvarez. Invertible conditional gans for image editing. *arXiv preprint arXiv:1611.06355*, 2016.

Omid Poursaeed, Tianxing Jiang, Harry Yang, Serge Belongie, and Ser-Nam Lim. Robustness and generalization via generative adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15711–15720, 2021.

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

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.

Daniel Roich, Ron Mokady, Amit H Bermano, and Daniel Cohen-Or. Pivotal tuning for latent-based editing of real images. *ACM Trans. Graph.*, 2021.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

Axel Sauer, Kashyap Chitta, Jens Müller, and Andreas Geiger. Projected gans converge faster. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Axel Sauer, Katja Schwarz, and Andreas Geiger. Stylegan-xl: Scaling stylegan to large diverse datasets, 2022. URL https://arxiv.org/abs/2202.00273.

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

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

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.

Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.

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.

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

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Daquan Zhou, Zhiding Yu, Enze Xie, Chaowei Xiao, Anima Anandkumar, Jiashi Feng, and Jose M. Alvarez. Understanding the robustness in vision transformers. In *International Conference on Machine Learning (ICML)*, 2022.