Improving Adversarial Robustness Using Proxy Distributions

Vikash Sehwag[⋆†], Saeed Mahloujifar[⋆], Tinashe Handina[⋆], Sihui Dai[⋆], Chong Xiang[⋆]
Mung Chiang[⋄], Prateek Mittal[⋆]

[⋆]Princeton University, [⋄]Purdue University

Abstract

We focus on the use of *proxy* distributions, i.e., approximations of the underlying distribution of the training dataset, in both understanding and improving the adversarial robustness in image classification. While additional training data helps in adversarial training, curating a very large number of real-world images is challenging. In contrast, proxy distributions enable us to sample a potentially unlimited number of images and improve adversarial robustness using these samples. We first ask the question: when does adversarial robustness benefit from incorporating additional samples from the proxy distribution in the training stage? We prove that the difference between the robustness of a classifier on the proxy and original training dataset distribution is upper bounded by the conditional Wasserstein distance between them. Our result confirms the intuition that samples from a proxy distribution that closely approximates training dataset distribution should be able to boost adversarial robustness. Motivated by this finding, we leverage samples from state-of-the-art generative models, which can closely approximate training data distribution, to improve robustness. In particular, we improve robust accuracy by up to 6.1% and 5.7% in l_∞ and l_2 threat model, and certified robust accuracy by 6.7% over baselines not using proxy distributions on the CIFAR-10 dataset. Since we can sample an unlimited number of images from a proxy distribution, it also allows us to investigate the effect of an increasing number of training samples on adversarial robustness. Here we provide the first large scale empirical investigation of accuracy vs robustness trade-off and sample complexity of adversarial training by training deep neural networks on 2K to 10M images.

1 Introduction

To achieve robustness against adversarial examples, adversarial training remains the most effective technique (Madry et al., 2018; Zhang et al., 2019; Pang et al., 2021). However, it is largely used with limited training samples available in current image datasets (such as only 50,000 training images in the CIFAR-10 dataset) where it suffers from multiple challenges, such as poor generalization on the test set (Madry et al., 2018) and accuracy vs robustness trade-off (Raghunathan et al., 2020). Recent works have demonstrated that more training data can improve the performance of adversarial training (Schmidt et al., 2018b; Uesato et al., 2019; Deng et al., 2020). However, this approach runs into the challenge of curating a large set of real-world images for training. We circumvent this challenge using proxy distributions, distributions that closely approximate the underlying distribution of the original training dataset. Note that the proxy distributions are modeled using only available training images in current datasets. Proxy distributions when modeled with deep neural networks, such as Generative adversarial networks (GAN), allow us to generate a potentially unlimited number of high-fidelity images (Goodfellow et al., 2014; Gui et al., 2020; Ho et al., 2020).

Work in progress. [†]Corresponding author: `vvikash@princeton.edu`
A short version is published at ICLR 2021 Workshop on Security and Safety in Machine Learning Systems.

Our key goal is to improve the adversarial robustness of deep neural networks by taking advantage of samples from such proxy distributions.

Note that proxy distributions are only an approximation of the underlying distribution of the training data. This begs the question of whether robustness achieved on samples from the proxy distribution will transfer to the original dataset. A more generic version of this question is *how much the robustness of a classifier transfers between data distributions?* We prove that the difference in the robustness of a classifier on two distributions is *tightly* upper bounded by the conditional Wasserstein distance between them. We refer to Wasserstein distance conditioned on class labels as conditional Wasserstein distance. This result confirms the intuition that samples from a proxy distribution that closely approximates training data distribution should be able to improve robustness on the latter. We also empirically validate this claim where we show that adversarial training only on samples from the proxy distribution indeed achieves non-trivial robustness on samples from the original distribution.

To improve adversarial robustness on current datasets, we aim to leverage samples from a proxy distribution which *closely* approximates the underlying distribution of these datasets. Here we propose to simultaneously train on both the original training set and a set of additional images sampled from the proxy distribution. We use state-of-the-art generative models as a model for proxy distribution since they can closely approximate the data distribution from only a limited number of training samples (Karras et al., 2020; Ho et al., 2020; Gui et al., 2020). Our experimental results demonstrate that the use of samples from the proxy distribution improves robust accuracy by up to 6.1% and 5.7% in l_∞ and l_2 threat models, respectively, over baselines not using proxy distributions on the CIFAR-10 dataset. In the category of not using any extra real-world data, our models achieve the first rank on RobustBench (Croce et al., 2020), a standardized benchmark for adversarial training. We also improve the certified robust accuracy (Cohen et al., 2019) by up to 6.7% with randomized smoothing on the CIFAR-10 dataset. Intriguingly, we achieve better certified robust accuracy with proxy distribution samples than using an additional set of 500K curated real-world images (Carmon et al., 2019).

Since access to proxy distributions gives us the ability to generate a large number of images, this allows us to delve deeper into the relationship of adversarial robustness with the number of training samples on the scale of image classification with deep neural networks. Next, we work solely with a proxy distribution, where we train multiple networks with an increasing amount of training images (we use 2K-10M training images) and test each on another fixed set of images from it. Our goal is to analyze multiple intriguing properties of the adversarial training w.r.t. the number of training samples. We first analyze the sample complexity of adversarial training. Earlier works (Wei and Ma, 2019; Bhagoji et al., 2019; Schmidt et al., 2018b) have derived sample complexity bounds in simplified settings such as classifying a mixture of Gaussian distributions. On the scale of deep neural networks for image classification, we empirically demonstrate that more data continues to improve adversarial robustness. Next, we analyze the accuracy vs robustness trade-off: earlier works showed that robustness in adversarial training is achieved at a cost of clean accuracy of deep neural networks (Raghunathan et al., 2020; Javanmard et al., 2020). We demonstrate that increase in the number of training samples significantly reduces this trade-off for deep neural networks. To the best of our knowledge, this is the *first* empirical investigation into adversarial training of deep neural networks on millions of images on the scale of CIFAR-10 dataset.

Contributions. We make the following key contributions.

- We provide a theoretical understanding as well as an empirical validation of the transfer of adversarial robustness between data distributions. In particular, we provide a tight upper bound on the difference between the robustness of a classifier on two data distributions.

- We propose to combine adversarial training with proxy distributions. By leveraging additional images sampled from proxy distributions, we improve robust accuracy by up to 6.1% and 5.7% in l_∞ and l_2 threat models, respectively, and certified robust accuracy by 6.7% on the CIFAR-10 dataset.

- We provide the *first* large scale empirical investigation of accuracy vs robustness trade-off and sample complexity of adversarial training by training deep neural networks on 2K to 10M images.

2 Integrating proxy distributions in adversarial training

We first provide a brief overview of adversarial training in deep neural networks. Before integrating proxy distributions in the adversarial training, we first delve deeper into the question of whether adversarial robustness will transfer from the proxy distribution to the original training data distribution. After answering it affirmatively, we integrate proxy distributions in both adversarial training and randomized smoothing, where the latter is used to achieve certified adversarial robustness.

Notation. We represent the input space by $\mathcal{X}$ and corresponding label space as $\mathcal{Y}$. We assume that the data is sampled from a joint distribution, i.e., $\mathbb{D}_{\mathcal{X}\times\mathcal{Y}}$. We refer to the underlying distribution of current image datasets, such as CIFAR-10, as D. While D is unknown, we assume that a limited set of training and test images are available from this distribution. We denote the proxy distribution as $\tilde{D}$. Unlike D, $\tilde{D}$ is known as it is modeled using a generative model. We denote the neural network for classification by $f : X \to Z$, parameterized by θ, which maps input images to output probability vectors (z). We represent the cross-entropy loss function, which is used to train the classifier, as $\ell(.)$, where $\ell(\theta, x, y) = \langle -log(f_\theta(x)), y\rangle$. For a set S sampled from a distribution D, we use $\hat{S}$ to denote the empirical distribution with respect to set S.

Formulation of adversarial training. The key objective in adversarial training is to minimize the training loss on adversarial examples obtained with adversarial attacks, such as projected gradient descent (PGD) (Madry et al., 2018) based attacks, under the following formulation.

$$\underset{\theta}{min} \underset{(x,y)\sim D}{\mathbb{E}} L_{adv}(\theta, x, y, \Omega), \qquad L_{adv}(\theta, x, y, \Omega) = \ell(\theta, PGD(x, \Omega), y)$$

where Ω is the threat model which includes the magnitude of adversarial perturbations (ϵ), number of attack steps, and the size of each step.

2.1 Understanding transfer of adversarial robustness between data distributions

As explained above, our goal is to use samples from a proxy distribution $\tilde{D}$ instead of the actual distribution D. We consider two distributions D and $\tilde{D}$ supported on $X \times Y$ which is the space of labeled examples. We use a classifier $h : X \to Y$ to predict class label of an input sample. We first define the average robustness of a classifier on a distribution followed by the definition of conditional Wasserstein distance, a measure of the distance between two labeled distributions.

Definition 1 (Average Robustness). *We define average robustness for a classifier f on a distribution D according to a distance metric d as follows:*

$$\mathsf{Rob}_d(h, D) = \underset{(x,y)\leftarrow D}{\mathbb{E}}[\inf_{h(x')\neq y} d(x', x)].$$

This definition refers to the expectation of the distance to the closest adversarial example for each sample. In contrast to robust accuracy, which measures whether an adversarial example exists within a given distance, we calculate the distance to the closest adversarial example. Having this definition, we can now explain our goal using the following decomposition of adversarial error. Let L be a learning algorithm (e.g. adversarial training). In robust learning from a proxy distribution, we are interested in bounding the average robustness of the classifier obtained by L, on distribution D, when the training set is a set S of n labeled examples sampled from a proxy distribution $\tilde{D}$. In particular we want to provide a lower bound on the following quantity

$$\mathop{\mathbb{E}}_{\substack{S\leftarrow\tilde{D}^n\\ h\leftarrow L(S)}}[\mathsf{Rob}_d(h,D)].$$

In order to understand this quantity better, suppose h is a classifier trained on a set S that is sampled from $\tilde{D}^n$, using algorithm L. We decompose $\mathsf{Rob}_d(h,D)$ to three quantities

$$\mathsf{Rob}_d(h,D) = \big(\mathsf{Rob}_d(h,D) - \mathsf{Rob}_d(h,\tilde{D})\big) + \big(\mathsf{Rob}_d(h,\tilde{D}) - \mathsf{Rob}_d(h,\hat{S})\big) + \mathsf{Rob}_d(h,\hat{S}).$$

Using this decomposition, by linearity of expectation and triangle inequality we can bound $\mathbb{E}_{\substack{S\leftarrow\tilde{D}^n\\ h\leftarrow L(S)}}[\mathsf{Rob}_d(h,D)]$ from below by

$$\underbrace{\mathop{\mathbb{E}}_{\substack{S\leftarrow\tilde{D}^n\\ h\leftarrow L(S)}}[\mathsf{Rob}_d(h,\hat{S})]}_{\textbf{Empirical robustness}} - \underbrace{\Big|\mathop{\mathbb{E}}_{\substack{S\leftarrow\tilde{D}^n\\ h\leftarrow L(S)}}[\mathsf{Rob}_d(h,\tilde{D}) - \mathsf{Rob}_d(h,\hat{S})]\Big|}_{\textbf{Generalization penalty}} - \underbrace{\Big|\mathop{\mathbb{E}}_{\substack{S\leftarrow\tilde{D}^n\\ h\leftarrow L(S)}}[\mathsf{Rob}_d(h,D) - \mathsf{Rob}_d(h,\tilde{D})]\Big|}_{\textbf{Distribution-shift penalty}}.$$

As the above inequality suggests, in order to bound the average robustness, we need to bound both the generalization penalty and the distribution shift penalty. Indeed, if D and $\tilde{D}$ were identical, we were in the standard robust learning setting and we had to only deal with the generalization penalty. The generalization penalty has been studied before in multiple works (Cullina et al., 2018; Montasser et al., 2019; Schmidt et al., 2018a) where distribution-independent bounds on robust generalization of adversarial training on VC classes are provided. Hence, we mostly focus on bounding the distribution shift penalty. Our goal is to provide a bound on the distribution-shift penalty that is independent of the classifier in hand and is only related to the properties of the distributions. With this goal, we define a notion of distance between two distributions.

Definition 2 (Conditional Wasserstein distance)**.** *For two labeled distributions D and $\tilde{D}$ supported on $X \times Y$, we define conditional wasserstein distance according to a distance metric d as follows:*

$$\mathsf{CWD}_d(D,\tilde{D}) = \mathop{\mathbb{E}}_{(\cdot,y)\leftarrow D}\left[\inf_{J\in\mathcal{J}(D|y,\tilde{D}|y)} \mathop{\mathbb{E}}_{(x,x')\leftarrow J}[d(x,x')]\right]$$

where $\mathcal{J}(D,\tilde{D})$ is the set of joint distributions whose marginals are identical to D and $\tilde{D}$.

Conditional Wasserstein distance between the two distributions is simply the expectation of Wasserstein distance between conditional distributions for each class. Note that Wasserstein distance is indeed used as a metric to measure the quality of generative models (Heusel et al., 2017). Now, we are ready to state our main theorem that bounds the distribution shift penalty for any learning algorithm based only on the Wasserstein distance of the two distributions.

Theorem 1 (Bounding distribution-shift penalty)**.** *Let D and $\tilde{D}$ be two labeled distributions supported on $X \times Y$ with identical label distributions, i.e., $\forall y^* \in Y, \Pr_{(x,y)\leftarrow D}[y = y^*] = \Pr_{(x,y)\leftarrow \tilde{D}}[y = y^*]$. Then for any classifier $h : X \to Y$*

$$|\mathsf{Rob}_d(h, \tilde{D}) - \mathsf{Rob}_d(h, D)| \leq \mathsf{CWD}_d(D, \tilde{D}).$$

Theorem 1 shows how one can bound the distribution-shift penalty. Importantly, it gives a method of measuring the quality of a proxy distribution for robust training. Note that, despite the generalization penalty, the distribution-shift penalty does not decrease when more data is provided. Hypothetically, we can have many samples from the proxy distribution which reduces the effect of generalization penalty and makes the distribution-shift penalty the dominant factor in reducing robustness. Although there are interesting theoretical results showing the effect of sample complexity on the generalization penalty, these bounds do not apply to neural networks and we take an empirical approach to show that the generalization penalty indeed approaches zero when more samples from the proxy distribution is provided (See Section 3.4). This shows the importance of Theorem 1 in understanding robustness we get by training on proxy distributions. In other words, this theorem enables us to switch our attention from robust generalization to creating high quality generative models for which the underlying distributions is close to the original distribution.

A natural question is what happens when we combine the original distribution with the proxy distribution. For example, one might have access to a generative model but they want to combine the samples from the generative model with some samples from the original distribution and train a robust classifier on the aggregated dataset. The following corollary answers this question.

Corollary 2. *Let D and $\tilde{D}$ be two labeled distributions supported on $X \times Y$ with identical label distributions and let $\bar{D} = p \cdot D + (1-p) \cdot \tilde{D}$ be the weighted mixture of D and $\tilde{D}$. Then for any classifier $h : X \to Y$*

$$|\mathsf{Rob}_d(h, \bar{D}) - \mathsf{Rob}_d(h, D)| \leq (1-p) \cdot \mathsf{CWD}_d(D, \tilde{D}).$$

Note that the value of p is usually very small as the number of data from proxy distribution is usually much higher than the original distribution. This shows that including (or not including) the data from original distribution should not have a large effect on the obtained bound on distribution-shift penalty.

Finally, we show a theorem that shows our bound on distribution-shift penalty is tight. The following theorem shows that one cannot obtain a bound on the distribution-shift penalty for a specific classifier that is *always* better than our bound.

Theorem 3 (Tightness of Theorem 1)**.** *For any distribution D supported on $X \times Y$, any classifier h, any homogeneous distance d and any $\epsilon \leq \mathsf{Rob}_d(h, D)$, there is a labeled distribution $\tilde{D}$ such that*

$$\mathsf{Rob}_d(h, D) - \mathsf{Rob}_d(h, \tilde{D}) = \mathsf{CWD}(D, \tilde{D}) = \epsilon.$$

Note that Theorem 3 only shows the tightness of Theorem 1 for a specific classifier. But there might exist a learning algorithm L that incurs a much better bound in the expectation. Namely, there might exist L such that for any two distributions D and $\tilde{D}$ we have

$$\Big| \mathop{\mathbb{E}}_{\substack{S \leftarrow \tilde{D}^n \\ h \leftarrow L(S)}} [\mathsf{Rob}_d(h, D) - \mathsf{Rob}_d(h, \tilde{D})] \Big| \ll \mathsf{CWD}(D, \tilde{D}).$$

We leave finding such an algorithm as an open question.

2.2 Improving adversarial robustness using proxy distributions

Now we focus on improving robustness on original training data distribution (D). As Theorem 1 states, robust training on a close proxy distribution ($\tilde{D}$) can generalize to the training data distribution. Therefore, to improve robustness on D, we propose to augment original training set with samples from $\tilde{D}$. In particular, we use the following adversarial training formulation.

$$\underset{\theta}{min} \left[\gamma * \underset{(x,y)\sim D}{\mathbb{E}} L_{adv}(\theta, x, y, \Omega) + (1-\gamma) * \underset{(x,y)\sim \tilde{D}}{\mathbb{E}} L_{adv}(\theta, x, y, \Omega) \right], \quad \text{where } \gamma \in [0, 1]$$

We approach it as an empirical risk minimization problem where we estimate the loss on original training data distribution using available samples in its training set. Similarly, we estimate loss on the proxy distribution using a set of synthetic images sampled from it. Since increasing the number of synthetic samples helps (Section 3.4), we use a significantly large number of synthetic samples than available training samples from D.

Next, we aim to also improve certified robustness on the training data distribution (D). We use randomized smoothing (Cohen et al., 2019) to certify robustness as it provides better performance and scalability than alternative techniques (Wong et al., 2018; Zhang et al., 2020). Similar to our modification in adversarial training, we propose to train on samples from both proxy distribution ($\tilde{D}$) and original training data distribution (D).

$$\underset{\theta}{min} \left[\gamma * \underset{(x,y)\sim D}{\mathbb{E}} L_{smooth}(\theta, x, y) + (1-\gamma) * \underset{(x,y)\sim \tilde{D}}{\mathbb{E}} L_{smooth}(\theta, x, y) \right]$$

where $\gamma \in [0, 1]$, $L_{smooth}(\theta, x, y) = \ell(\theta, x, y) + \beta D_{kl}(f_\theta(x), f_\theta(x + \delta))$, $\delta \sim \mathrm{N}(0, \sigma^2 I)$, and $D_{kl}(., .)$ is the KL-divergence.

We use the combination of both cross-entropy loss over unperturbed images and KL-divergence loss on images perturbed with Gaussian noise. This loss function, originally proposed for stability training (Zheng et al., 2016), performs better than using $\ell(\theta, x + \delta, y)$ as L_{smooth} (Carmon et al., 2019), where $\ell(.)$ is the cross-entropy loss function.

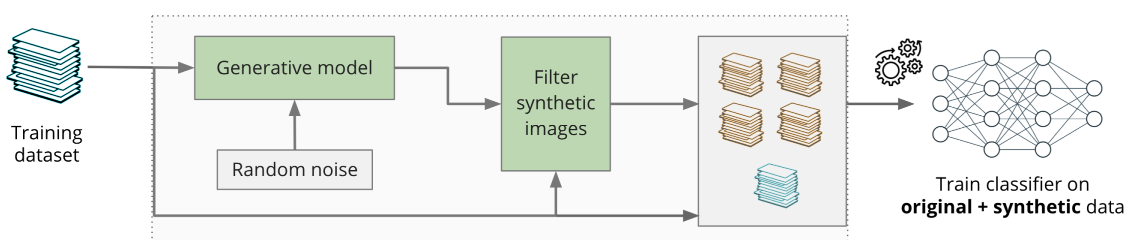


Figure 1: **Overview of our approach.** Sketch of our proposed approach to improve adversarial robustness using proxy distributions. Using only training images from a given dataset, we first train a generative model and sample a significantly large amount of synthetic images from it. Next, we filter poor quality synthetic images from it. Finally, we robustly train a classifier on the combined set of filtered synthetic samples and original training samples.

3 Experimental results

We first describe our experimental setup followed by the choice of generative model for the proxy distribution. Next, we demonstrate the gains in both clean and robust accuracy using samples from the proxy distribution. We also demonstrate that filtering poor quality synthetic images further improve the gains. In the end, we explore different intriguing properties of adversarial training with increasing number of training samples.

3.1 Common experimental setup across experiments

We use network architectures from the ResNet family, namely ResNet-18 (He et al., 2016) and variants of WideResNet (Zagoruyko and Komodakis, 2016). We primarily work with the CIFAR-10 dataset and its two-class subset, i.e., an easier problem of binary classification between class-1 (*automobile*) and class-9 (*truck*). We refer to the latter as CIFAR-2. We consider both l_∞ and l_2 threat models. We use a perturbation budget (ϵ) of 8/255 and 127/255 for the former and latter threat model, respectively. We perform adversarial training using a 10-step projected gradient descent attack (PGD-10) and benchmark test set robustness with the much stronger AutoAttack (Croce and Hein, 2020)[1]. When using proxy distributions in training, we set $\gamma = 0.5$ in favor of simplicity. We train each network using stochastic gradient descent and cosine learning rate with weight decay of 5×10^{-4} and 100 epochs. We use two key metrics to evaluate performance of trained models: *clean accuracy* and *robust accuracy*. While the former refers to the accuracy on unmodified test set images, the latter refers to the accuracy on adversarial examples generated from test set images. We also measure *certified robust accuracy* when we use randomized smoothing. We use $\beta = 6$, $\sigma = 0.25$, 100 samples for selection, and $10,000$ samples for estimation in randomized smoothing, as described in Cohen et al. (2019).

Table 1: **Comparing different generative models.** Comparing quality of synthetic samples generated from StyleGAN and DDPM model. We report FID and Inception score which are standard metrics to evaluate quality of synthetic samples[2]. We also measure generalization to a held out test set of 100K synthetic images and CIFAR-10 test set, when a ResNet-18 network is adversarially trained only on 1M synthetic images from the respective model.

Model	FID	Inception score	On synthetic data		On CIFAR-10	
			Clean	Robust	Clean	Robust
StyleGAN (Karras et al., 2020)	**2.92**	**10.24**	**94.1**	78.4	84.5	51.8
DDPM (Ho et al., 2020)	3.17	9.46	88.5	**79.2**	**85.4**	**73.6**

3.2 Choice of generative model for the proxy distribution

We work with two state-of-the-art generative models, namely StyleGAN (Karras et al., 2020) and DDPM (Ho et al., 2020). While the former is a generative adversarial network (GAN), the latter is a probabilistic model based on the diffusion process. We sample 10M labeled images from the conditional StyleGAN and another set of 6M images from the DDPM model[3]. We use a smaller

[1] We don't report numbers with PGD attacks as AutoAttack already captures them while also making it easier to compare with other works (Croce et al., 2020).

[2] We provide numbers reported in the original publication of each method.

[3] It is a pre-sampled set of images made available by Nakkiran et al. (2021).

number of samples from the latter as the cost of generating each sample from it is significantly higher than the former. Note that these samples are drawn from an unconditional DDPM model, thus class labels aren't available. We label these images using LaNet (Wang et al., 2019a) network, which achieves 99.03% clean accuracy on the CIFAR-10 dataset without using any additional data. We discuss the effect of different labeling strategies later in this section. To avoid any leakage of the test set in generated samples, both generative models are trained only on the training set of the CIFAR-10 dataset.

StyleGAN vs DDPM. Our key question is that which generative model is more useful in improving robustness on the CIFAR-10 dataset? We judge this by adversarial training on only synthetic samples (1M in total) from the generative model and evaluating the performance on the test set of the CIFAR-10 dataset. We use PGD-4 attack in both training and evaluating robustness.

We note that existing metrics to evaluate quality of synthetic samples, such as Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Salimans et al., 2016), *fails to answer* this question. While samples from the StyleGAN model achieve better FID and Inception score, adversarial training on them achieves lower performance on the CIFAR-10 test set than samples from the DDPM model (Table 1). For example, training on samples from the latter model achieve 73.6% robust accuracy while samples from the former achieve only 51.8% robust accuracy on the CIFAR-10 test set. We hypothesize that DDPM model generates more hard to learn instances which helps most in learning but are less photorealistic, thus have lower FID and inception score than StyleGAN images. Given their better generalization to the CIFAR-10 dataset, we will use samples from the DDPM model to improve robustness in the next section.

3.3 Improving adversarial robustness using proxy distributions

Now we demonstrate that integrating synthetic samples, i.e., samples from the proxy distribution can lead to state-of-the-art adversarial robustness. First we describe our experimental setup. Next we demonstrate the benefit of filtering poor quality synthetic samples. Finally, using this filtered set of samples, we significantly improve the adversarial robustness of deep neural networks.

Setup. We use 6M synthetic images sampled from the unconditional DDPM model which are later labeled using the LaNet model (Wang et al., 2019a). We combine real and synthetic images in a 1:1 ratio in each batch. We use the same number of training steps as in the baseline setup which doesn't use synthetic samples. Therefore our computational cost, despite using millions of synthetic samples, is only 2$\times$ of baseline and equal to earlier works that have incorporated extra samples (Carmon et al., 2019; Gowal et al., 2020).

3.3.1 Filtering poor quality images from the synthetic data

We explore different classifiers to label the unlabeled synthetic data generated from the DDPM model. In particular, we use BiT (Kolesnikov et al., 2020), SplitNet (Zhao et al., 2020), and LaNet (Wang et al., 2019a) where they achieve 98.5%, 98.7%, and 99.0% clean accuracy, respectively, on the CIFAR-10 dataset. We find that labels generated from different classifiers achieve slightly different downstream performance when used with adversarial training in the proposed approach (Table 2). We measure both clean accuracy (*Clean*) and robust accuracy with AutoAttack (*Auto*) when performing adversarial (adv.) training with both synthetic and original training images. We find that only up to 10% of synthetic images are labeled differently by these networks, which causes these differences. On manual inspection, we find that some of these images are of poor quality, i.e., images that aren't photorealistic or wrongly labeled and remain hard to classify, even for a human labeler.

Since filtering millions of images with a human in the loop is extremely costly, we use two deep neural networks, namely LaNet (Wang et al., 2019a) and SplitNet (Zhao et al., 2020), to solve this task. We avoid using labels from BiT as it requires transfer learning from ImageNet (Deng et al., 2009) dataset, whereas our goal is to avoid any dependency on extra real-world data. We discard an image when the predicted class of both networks doesn't match and it is classified with less than 90% confidence by both networks. While the former step flags images which are potentially hard to classify, the latter step ensures that we do not discard images where at least one network is highly confident in its prediction. We also try the 50% and 75% confidence threshold but find that 90% gives the best downstream results. In this process, we discarded 98.6K images from 6M synthetic images. We display some of these discarded images in Figure 5d in Appendix. Our experimental results show that the filtering step slightly increases the performance gains from synthetic images (Table 2). We use this filtered set of synthetic images in further experiments.

Table 2: **Filtering synthetic images.** Different labeling networks for synthetic data achieves slightly different downstream performance. Filtering poor quality images based on these labels further improves the performance.

Network	CIFAR-10 (*Clean*)	Adv. training *Clean*	*Auto*
BiT	98.5	84.0	54.3
SplitNet	98.7	84.3	54.7
LaNet	99.0	84.2	54.7
LaNet (filtered)	99.0	**84.4**	**54.8**

3.3.2 Integrating the filtered set of synthetic samples in adversarial training

State-of-the-art robust accuracy. We observe that incorporating the filtered set of synthetic images in adversarial training leads to state-of-the-art robust accuracy. In the l_∞ threat model and using a Wrn-34-10 network, it improves it to 59.1%, an improvement of 6.0% over previous work using the same network. Similarly, we observe improvement by up to 5.7% for l_2 attacks. Note that clean accuracy also improves simultaneously. In the category of not using any extra real-world data, we achieve first rank on RobustBench (Croce et al., 2020), a standardized benchmark for adversarial robustness, across both threat models where we outperform the previous start-of-the-art from Gowal et al. (2020). Note that in comparison to Gowal et al. (2020), which uses a Wrn-70-16 network with 266M parameters, we use a smaller Wrn-34-10 network with only 46M parameters.

Proxy distribution offsets increase in network parameters. We find that gains from using synthetic samples are equivalent to ones obtained by scaling network size by an order of magnitude

Table 3: **State-of-the-art adversarial robustness.** Experimental results with adversarial training on the CIFAR-10 dataset for both ℓ_∞ and ℓ_2 threat model. Using additional synthetic data brings a large gain in adversarial robustness across networks architecture and threat models. *Clean* and *Auto* refers to clean accuracy and robust accuracy measured with AutoAttack, respectively.

(a) ℓ_∞ threat model.

Method	Architecture	Parameters (M)	*Clean*	*Auto*
Zhang et al. (2019)	ResNet-18	11.2	82.0	48.7
Madry et al. (2018)	ResNet-50	23.5	87.0	49.0
Zhang et al. (2019)	Wrn-34-10	46.2	84.9	53.1
Rice et al. (2020)	Wrn-34-20	184.5	85.3	53.4
Gowal et al. (2020)	Wrn-70-16	266.8	85.3	57.2
Ours	ResNet-18	11.2	84.4	**54.8**
Ours	Wrn-34-10	46.2	85.8	**59.1**

(b) ℓ_2 threat model.

Method	Architecture	Parameters (M)	*Clean*	*Auto*
Rice et al. (2020)	ResNet-18	11.2	88.7	67.7
Madry et al. (2018)	ResNet-50	23.5	90.8	69.2
Wu et al. (2020)	Wrn-34-10	46.2	88.5	73.7
Gowal et al. (2020)	Wrn-70-16	266.8	90.9	74.5
Ours	ResNet-18	11.2	89.5	**73.4**
Ours	Wrn-34-10	46.2	90.3	**76.1**

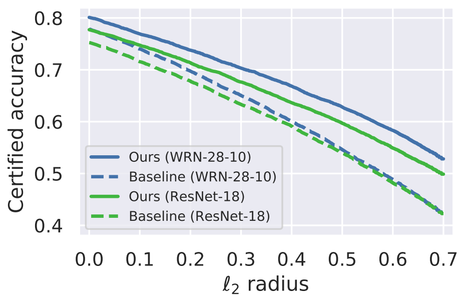


Figure 2: **Certified robustness.** Certified robust accuracy of baseline randomized smoothing technique, i.e., RST (Carmon et al., 2019) and our work with two different models.

Method	*Clean*	*Certified*
Wong et al. (2018) (single)	68.3	53.9
Wong et al. (2018) (ensemble)	64.1	63.6
CROWN-IBP (Zhang et al., 2020)	71.5	54.0
Balunovic and Vechev (2019)	78.4	60.5
RST (Carmon et al., 2019)	77.9	58.6
RST_{500K} (Carmon et al., 2019)	80.7	63.8
`Ours` (ResNet-18)	**81.1**	**62.5**
`Ours` (WRN-28-10)	**83.5**	**65.3**

Table 4: **Detailed comparison of certified robustness.** Comparing both clean accuracy (*Clean*) and certified robust accuracy (*Certified*) of our work with earlier approaches at an ℓ_∞ perturbation of 2/255.

(Table 3). For example, a ResNet-18 network with synthetic data achieves higher robust accuracy (l_∞) than a WRN-34-20 trained without it, while having 16× fewer parameters than the latter. Similar trend holds for WRN-34-10 networks, when compared with a much larger WRN-70-16 network. This trend holds for both ℓ_∞ and ℓ_2 threat models (Table 3a, 3b).

Simultaneous improvement in clean accuracy. Due to the accuracy vs robustness trade-off in adversarial training, improvement in robust accuracy often comes at the cost of clean accuracy. However synthetic samples provide boost in both clean and robust accuracy, simultaneously. We observe improvement in clean accuracy by up to 2.4% and 1.8% across ℓ_∞ and ℓ_2 threat models, respectively.

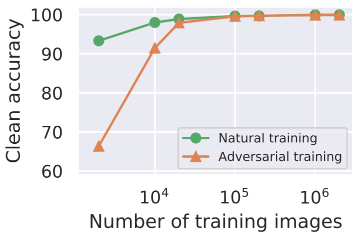
3.3.3 Improving certified robustness using the filtered set of synthetic samples

We provide results on certified adversarial robustness in Figure 2 and Table 4. We first compare the performance of our proposed approach with the baseline technique, i.e., RST (Carmon et al., 2019). We achieve significantly higher certified robust accuracy than the baseline approach at all ℓ_2 perturbations budgets for both ResNet-18 and WRN-28-10 network architectures. Additionally, the robustness of our approach decays at a smaller rate than the baseline. At ℓ_∞ perturbations of 2/255, equivalent to ℓ_2 perturbation of 111/255, our approach achieves 6.7% higher certified robust accuracy than RST. We also significantly outperform other certified robustness techniques which aren't based on randomized smoothing (Zhang et al., 2020; Wong et al., 2018; Balunovic and Vechev, 2019). Along with better certified robust accuracy, our approach also achieve better clean accuracy than previous approaches, simultaneously.

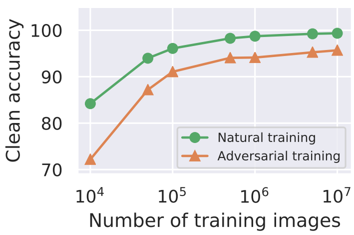
Synthetic images outperform real-world images. Using only synthetic samples, we also outperform RST when it uses an additional curated set of 500K real-world images (RST_{500K}). While the latter achieves 63.8% certified robust accuracy, we improve it to 65.3%. We also achieve 2.8% higher clean accuracy than RST_{500K}.

3.4 Investigating adversarial robustness with increasing number of training samples

Given the ability to sample an unlimited amount of synthetic images from a proxy distribution, now we investigate the performance of adversarial training with increasing number of training samples.

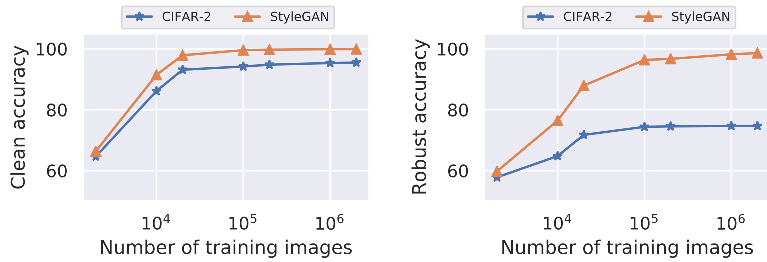


(a) CIFAR-2

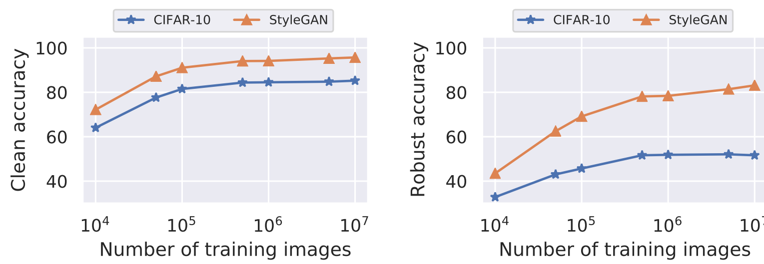


(b) CIFAR-10

Figure 3: **Reduction in accuracy vs robustness trade-off.** Accuracy vs robustness trade-off when training on an increasing amount of synthetic images from the StyleGAN model. It shows that the drop in clean accuracy with adversarial training decreases with increase in training samples.



(a) CIFAR-2



(b) CIFAR-10

Figure 4: **Sample complexity of adversarial training.** Clean and robust accuracy on the test set of synthetic samples when trained on an increasing number of synthetic samples from the StyleGAN model. It shows that performance of adversarial training continues to benefit from increase in number of training samples. We also measure generalization to the CIFAR-10 dataset, which also improves with number of training samples.

We train the network only on synthetic images and measure its performance on another held out set of synthetic images. We also measure how much the performance generalizes on the CIFAR-10 test set.

Setup. We robustly train a ResNet-18 network on 2K to 10M synthetic images from the StyleGAN model, as in both 10-class and 2-class setup. We opt for StyleGAN over DDPM model as sampling images from the former is much faster, thus we were able to generate up to 10M synthetic from it. Note that the cost of adversarial training increases almost linearly with the number of attack steps and training images. Thus to achieve manageable computational cost when training on millions of images, we opt for using only a 4-step PGD attack (PGD-4) in both training. Since robustness achieved with this considerably weak attack may not hold against a strong attack, such as AutoAttack, we opt for evaluating with the PGD-4 attack itself. We also perform natural training, i.e., training on unmodified images in some experiments. We test each network on a fixed set of 100K images from the StyleGAN and 10K images from the CIFAR-10 test set.

Accuracy vs robustness trade-off. We compare the clean accuracy achieved with both natural and adversarial training in Figure 3. Indeed with a very small number of samples, clean accuracy in adversarial training is traded to achieve robustness. This is evident from the gap between the clean accuracy of natural and adversarial training. However, with the increasing number of training samples, this gap keeps decreasing for both CIFAR-2 and CIFAR-10 datasets. Most interestingly, this trade-off almost vanishes when we use a sufficiently high number of training samples for the CIFAR-2 classification.

On sample complexity of adversarial training. We report both clean and robust accuracy with adversarial training in Figure 4. We find that both clean and robust accuracy continue to improve with the number of training samples. We also observe non-trivial generalization to test images from the CIFAR-10 dataset, which also improves with the number of training samples. Both of these results suggest that even with a small capacity network, such as ResNet-18, adversarial robustness can continue to benefit from an increase in the number of training samples.

4 Related work

4.1 Robust Machine learning

Adversarial examples. Adversarial examples, crafted using an adversarial attack, aim to evade the classifier at the test time (Biggio et al., 2013; Szegedy et al., 2014; Biggio and Roli, 2018). While we focus on ℓ_p perturbation based adversarial examples (Goodfellow et al., 2015), there exist multiple other threat models (Wong et al., 2019; Sehwag et al., 2019; Hosseini and Poovendran, 2018; Laidlaw and Feizi, 2019; Kang et al., 2019). Adversarial examples have been successful against almost all applications of machine learning (Xu et al., 2020, 2016; Rahman et al., 2019).

On use of l_p threat model. We consider ℓ_p perturbation based adversarial examples since it's the most widely studied threat model (Croce et al., 2020). Additionally, robustness achieved in this threat model has also led to further benefits in explainability (Chalasani et al., 2020), image recognition (Xie et al., 2020a; Salman et al., 2020b), transfer learning (Salman et al., 2020a; Utrera et al., 2021), self-supervised learning (Ho and Vasconcelos, 2020), and image synthesis (Santurkar et al., 2019).

Adversarial training. Adversarial training (Goodfellow et al., 2015; Madry et al., 2018) still remains the most effective defense against adversarial examples. The performance of the baseline technique (Madry et al., 2018) is further improved using adaptive loss functions (Zhang et al., 2019), larger networks (Gowal et al., 2020), pretraining (Hendrycks et al., 2019), smooth activation function (Xie et al., 2020b), and weight perturbations (Wu et al., 2020). The fundamental min-max robust optimization behind adversarial training has also been successful in robustifying other model architectures, such as decision trees (Chen et al., 2019) and graph neural networks (Feng et al., 2019), while also extending to other domains like reinforcement learning (Gleave et al., 2020) and natural language processing (Liu et al., 2020).

Progress on improving adversarial robustness. Improving adversarial robustness in deep neural networks remains a challenging problem and its progress has been slow, as tracked by the RobustBench (Croce et al., 2020). Following initial success of baseline adversarial training in Madry et al. (2018), robust accuracy has improved by only 12% in the setting where networks architecture is fixed to WRN-28-10 and no extra real-world images are used (from 44.0% in Madry et al. (2018) to 56.2% in Wu et al. (2020)). Additionally, there have been more than twenty failed attempts in improving it over the baseline (Croce et al., 2020). To put our work in context, we further improve robust accuracy on the CIFAR-10 dataset to 59.5%, without using additional techniques like smooth activations or weight perturbations.

Certified robustness. Here the goal is to certify the robustness of each examples (Wong et al., 2018; Cohen et al., 2019; Zhang et al., 2020). Certified robustness provides a lower bound on the adversarial robustness against all attacks in the threat model. We use randomized smoothing (Cohen et al., 2019) to achieve certified robustness, since it achieves much better performance over other methods in this domain (Cohen et al., 2019; Carmon et al., 2019). Further improvements in randomized smoothing includes integration with adversarial training (Salman et al., 2019), using sub-networks (Sehwag et al., 2020), and using sample-specific smoothing (Alfarra et al., 2020).

4.2 Intriguing properties of adversarial training

Sample complexity of adversarial training. Multiple earlier works provide theoretical results studying the effect of the number of training samples on adversarial robustness (Wei and Ma, 2019; Bhagoji et al., 2019; Schmidt et al., 2018b; Chen et al., 2020; Min et al., 2020). Chen et al. (2020) and Min et al. (2020) further suggest that more data may hurt generalization in adversarial training. We explore this direction empirically by adversarial training deep neural networks on an increasing number of training images.

Trade-off in adversarial training. There also exists an accuracy vs robustness trade-off where improvement in robust accuracy comes at the cost of clean accuracy in adversarial training (Raghunathan et al., 2020; Balaji et al., 2019; Javanmard et al., 2020). We empirically demonstrate that increase in the number of training samples can significantly reduce this trade-off.

Transfer of adversarial robustness. This line of work focuses on the transfer of adversarial robustness, i.e., correct classification even under adversarial perturbations, when testing the model on different data distributions (Shafahi et al., 2020; Sehwag et al., 2019). Note that this is different from just achieving correct classification on unmodified images across different distributions (Taori et al., 2020; Hendrycks and Dietterich, 2019). Here we provide a theoretical analysis of the transfer of adversarial robustness between data distributions.

Using extra curated real-world data. Prior works (Zhai et al., 2019; Carmon et al., 2019; Uesato et al., 2019; Najafi et al., 2019; Deng et al., 2020) have argued for using more training data in adversarial training and often resort to curating additional real-world samples. In contrast, we don't require additional real-world samples, as we model the proxy distribution from the limited training images available and sample additional synthetic images from this distribution.

4.3 Using generative models for proxy distributions

Generative models for proxy distributions. State-of-the-art generative models are capable of modeling the distribution of current large-scale image datasets. In particular, generative adversarial networks (GANs) have excelled at this task (Goodfellow et al., 2014; Karras et al., 2020; Gui et al., 2020). Though GANs generate images with high fidelity, they often lack high diversity (Ravuri and Vinyals, 2019). However, samples from recently proposed diffusion process based models achieve both high diversity and fidelity (Ho et al., 2020; Nichol and Dhariwal, 2021).

Evaluating quality of samples from a generative model. Fréchet Inception Distance (FID) (Heusel et al., 2017) and Inception Score (IS) (Heusel et al., 2017) are two key metrics to evaluate the quality of samples from generative models. While IS considers only generated images, FID considers both generative and real-world images. We find that better FID or Inception score may not translate to higher robustness with synthetic data. Another line of work evaluates generative models by training a classifier on its generated samples and testing it on real-world data (Ravuri and Vinyals, 2019; Semeniuta et al., 2018).

Using generative models to improve adversarial robustness. Earlier works have used generative models to learn training data manifold and the use it to map input samples to data manifold (Samangouei et al., 2018; Jalal et al., 2017; Xu et al., 2018). However, most of these techniques are broken against adaptive attacks (Athalye et al., 2018; Tramèr et al., 2020). We use generative models to sample additional training samples which lead to further improvement in adversarial robustness.

Additional applications of proxy distributions. While we use proxy distributions to improve robustness, synthetic samples from proxy distributions are also useful in privacy-preserving healthcare (Jordon et al., 2018), autonomous driving (Mayer et al., 2016), crowd-counting (Wang et al., 2019b), text recognition (Jaderberg et al., 2014; Ye et al., 2018), and natural language processing (Marzoev et al., 2020; Puri et al., 2020). Earlier works have also created dedicated synthetic datasets for some of these applications (Gaidon et al., 2016; Mayer et al., 2018).

Comparison with Rebuffi et al. (2021). A concurrent work by Rebuffi et al. (2021) also uses samples from generative models to improve adversarial robustness. While it broadly focuses on the effect of different data augmentations, with synthetic samples from a proxy distribution being one of them, our goal is to delve deeper into the integration of proxy distribution in adversarial training. This includes providing tight bounds on the transfer of adversarial robustness from the proxy distribution followed by empirical analysis of the transfer of robustness, accuracy vs robustness trade-off, and sample complexity of adversarial training. We also demonstrate an improvement in certified robust accuracy using proxy distributions. However, despite the differences, similar benefits of using generative models in two independent works further ascertain the importance of this research direction.

5 Discussion and Future work

Using synthetic data has been a compelling solution in many applications, such as healthcare (Jordon et al., 2018) and autonomous driving (Mayer et al., 2016), since it makes collecting a large amount of data feasible. In a similar spirit, we use synthetic data to make deep neural networks more robust to adversarial attacks. However synthetic data is sampled from a proxy distribution, i.e., a distribution only approximating the underlying data distribution of the training data. Thus the first key question is whether synthetic data will help at all in improving robustness. We study the transfer of robustness from proxy to original training data distribution and provide a tight upper bound on it. This result validates the intuition that a proxy distribution, which closely approximates the training data distribution, should be able to improve robustness.

When selecting a generative model for the proxy distribution, we argue that an inflection point exists in their progress, post which generative models sufficiently capture the modes of data, thus generating both photorealistic and diverse set of samples. On the CIFAR-10 dataset, we find that both StyleGAN and DDPM models are past this inflection point as samples from both improve the performance. However on ImageNet (Deng et al., 2009) dataset, we didn't observe any performance improvement with state-of-the-art BigGAN-deep (Brock et al., 2019) model. It suggests that we still remain far from the aforementioned inflection point on the ImageNet dataset.

On the CIFAR-10 dataset, we improve adversarial robust accuracy by up to 6.4% and certified robust accuracy by 6.7% using synthetic samples from the DDPM model. Since the main goal of the paper was to show the effectiveness of synthetic data, we didn't use additional techniques to boost robustness. Future work can incorporate these techniques, such as using weight perturbations, weight averaging, longer training schedules, and larger networks, along with synthetic data to further boost robustness. While a major push in recent works is to use larger networks to improve robustness (Gowal et al., 2020), we show that similar gains can be obtained by expanding training data to include synthetic samples. Motivated by these findings, we encourage the community to further innovate on the training data distribution itself to improve adversarial robustness.

References

Motasem Alfarra, Adel Bibi, Philip HS Torr, and Bernard Ghanem. Data dependent randomized smoothing. *arXiv preprint arXiv:2012.04351*, 2020.

Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018.

Yogesh Balaji, Tom Goldstein, and Judy Hoffman. Instance adaptive adversarial training: Improved accuracy tradeoffs in neural nets. *arXiv preprint arXiv:1910.08051*, 2019.

Mislav Balunovic and Martin Vechev. Adversarial training and provable defenses: Bridging the gap. In *International Conference on Learning Representations*, 2019.

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In , *32 Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 7496–7508, 2019.

Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331, 2018.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.

Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, pages 11190–11201, 2019.

Prasad Chalasani, Jiefeng Chen, Amrita Roy Chowdhury, Xi Wu, and Somesh Jha. Concise explanations of neural networks using adversarial training. In *International Conference on Machine Learning*, pages 1383–1391. PMLR, 2020.

Hongge Chen, Huan Zhang, Duane Boning, and Cho-Jui Hsieh. Robust decision trees against adversarial examples. In *International Conference on Machine Learning*, pages 1122–1131. PMLR, 2019.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *International Conference on Machine Learning*, pages 1670–1680. PMLR, 2020.

Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. Certified adversarial robustness via randomized smoothing. In *International Conference on Machine Learning*, pages 1310–1320. PMLR, 2019.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216. PMLR, 2020.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

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.

Zhun Deng, Linjun Zhang, Amirata Ghorbani, and James Zou. Improving adversarial robustness via unlabeled out-of-domain data. *arXiv preprint arXiv:2006.08476*, 2020.

Fuli Feng, Xiangnan He, Jie Tang, and Tat-Seng Chua. Graph adversarial training: Dynamically regularizing based on graph structure. *IEEE Transactions on Knowledge and Data Engineering*, 2019.

Adrien Gaidon, Qiao Wang, Yohann Cabon, and Eleonora Vig. Virtual worlds as proxy for multi-object tracking analysis. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4340–4349, 2016.

Adam Gleave, Michael Dennis, Cody Wild, Neel Kant, Sergey Levine, and Stuart Russell. Adversarial policies: Attacking deep reinforcement learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In , *Advances in Neural Information Processing Systems*, volume 27, 2014.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In , *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *arXiv preprint arXiv:2001.06937*, 2020.

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

Dan Hendrycks and Thomas G. Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, 2019.

Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019.

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 *30 Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 6626–6637, 2017.

Chih-Hui Ho and Nuno Vasconcelos. Contrastive learning with adversarial examples. In , *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *arXiv preprint arXiv:2006.11239*, 2020.

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

Max Jaderberg, Karen Simonyan, Andrea Vedaldi, and Andrew Zisserman. Synthetic data and artificial neural networks for natural scene text recognition. *arXiv preprint arXiv:1406.2227*, 2014.

Ajil Jalal, Andrew Ilyas, Constantinos Daskalakis, and Alexandros G Dimakis. The robust manifold defense: Adversarial training using generative models. *arXiv preprint arXiv:1712.09196*, 2017.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, pages 2034–2078. PMLR, 2020.

James Jordon, Jinsung Yoon, and Mihaela Van Der Schaar. Pate-gan: Generating synthetic data with differential privacy guarantees. In *International Conference on Learning Representations*, 2018.

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

Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data. *arXiv preprint arXiv:2006.06676*, 2020.

Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *ECCV*, 2020.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. In *32 Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 10408–10418, 2019.

Xiaodong Liu, Hao Cheng, Pengcheng He, Weizhu Chen, Yu Wang, Hoifung Poon, and Jianfeng Gao. Adversarial training for large neural language models. *arXiv preprint arXiv:2004.08994*, 2020.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*, 2018.

Alana Marzoev, Samuel Madden, M Frans Kaashoek, Michael Cafarella, and Jacob Andreas. Unnatural language processing: Bridging the gap between synthetic and natural language data. *arXiv preprint arXiv:2004.13645*, 2020.

Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4040–4048, 2016.

Nikolaus Mayer, Eddy Ilg, Philipp Fischer, Caner Hazirbas, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. What makes good synthetic training data for learning disparity and optical flow estimation? *International Journal of Computer Vision*, 126(9):942–960, 2018.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization. *arXiv preprint arXiv:2002.11080*, 2020.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

Amir Najafi, Shin-ichi Maeda, Masanori Koyama, and Takeru Miyato. Robustness to adversarial perturbations in learning from incomplete data. In , *32 Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5542–5552, 2019.

Preetum Nakkiran, Behnam Neyshabur, and Hanie Sedghi. The bootstrap framework: Generalization through the lens of online optimization. In *International Conference on Learning Representations*, 2021.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. *arXiv preprint arXiv:2102.09672*, 2021.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

Raul Puri, Ryan Spring, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. Training question answering models from synthetic data. *arXiv preprint arXiv:2002.09599*, 2020.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C. Duchi, and Percy Liang. Understanding and mitigating the tradeoff between robustness and accuracy. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7909–7919, 2020.

Mohammad Saidur Rahman, Mohsen Imani, Nate Mathews, and Matthew Wright. Mockingbird: Defending against deep-learning-based website fingerprinting attacks with adversarial traces. *arXiv preprint arXiv:1902.06626*, 2019.

Suman Ravuri and Oriol Vinyals. Classification accuracy score for conditional generative models. In , *Advances in Neural Information Processing Systems*, volume 32, 2019.

Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021.

Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pages 8093–8104. PMLR, 2020.

Tim Salimans, Ian J. Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training gans. In *29 Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 2226–2234, 2016.

Hadi Salman, Jerry Li, Ilya P. Razenshteyn, Pengchuan Zhang, Huan Zhang, Sébastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *32 Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 11289–11300, 2019.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? In , *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020a.

Hadi Salman, Andrew Ilyas, Logan Engstrom, Sai Vemprala, Aleksander Madry, and Ashish Kapoor. Unadversarial examples: Designing objects for robust vision. *arXiv preprint arXiv:2012.12235*, 2020b.

Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018.

Shibani Santurkar, Andrew Ilyas, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Image synthesis with a single (robust) classifier. In , *32 Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 1260–1271, 2019.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In , *Advances in Neural Information Processing Systems*, volume 31, 2018a.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. In , *31 Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 5019–5031, 2018b.

Vikash Sehwag, Arjun Nitin Bhagoji, Liwei Song, Chawin Sitawarin, Daniel Cullina, Mung Chiang, and Prateek Mittal. Analyzing the robustness of open-world machine learning. In *Proceedings of the 12th ACM Workshop on Artificial Intelligence and Security*, pages 105–116, 2019.

Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. In *33 Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Stanislau Semeniuta, Aliaksei Severyn, and Sylvain Gelly. On accurate evaluation of gans for language generation. *arXiv preprint arXiv:1806.04936*, 2018.

Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David W. Jacobs, and Tom Goldstein. Adversarially robust transfer learning. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In , *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. In , *Advances in Neural Information Processing Systems*, volume 33, pages 18583–18599, 2020.

Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *33 Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

Jonathan Uesato, Jean-Baptiste Alayrac, Po-Sen Huang, Alhussein Fawzi, Robert Stanforth, and Pushmeet Kohli. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Francisco Utrera, Evan Kravitz, N. Benjamin Erichson, Rajiv Khanna, and Michael W. Mahoney. Adversarially-trained deep nets transfer better. In *International Conference on Learning Representations*, 2021.

Linnan Wang, Saining Xie, Teng Li, Rodrigo Fonseca, and Yuandong Tian. Sample-efficient neural architecture search by learning action space. *arXiv preprint arXiv:1906.06832*, 2019a.

Qi Wang, Junyu Gao, Wei Lin, and Yuan Yuan. Learning from synthetic data for crowd counting in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8198–8207, 2019b.

Colin Wei and Tengyu Ma. Improved sample complexities for deep neural networks and robust classification via an all-layer margin. In *International Conference on Learning Representations*, 2019.

Eric Wong, Frank R. Schmidt, Jan Hendrik Metzen, and J. Zico Kolter. Scaling provable adversarial defenses. In *31 Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 8410–8419, 2018.

Eric Wong, Frank Schmidt, and Zico Kolter. Wasserstein adversarial examples via projected sinkhorn iterations. In *International Conference on Machine Learning*, pages 6808–6817. PMLR, 2019.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020.

Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 819–828, 2020a.

Cihang Xie, Mingxing Tan, Boqing Gong, Alan Yuille, and Quoc V Le. Smooth adversarial training. *arXiv preprint arXiv:2006.14536*, 2020b.

Han Xu, Yao Ma, Hao-Chen Liu, Debayan Deb, Hui Liu, Ji-Liang Tang, and Anil K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

Weilin Xu, Yanjun Qi, and David Evans. Automatically evading classifiers. In *Proceedings of the 2016 network and distributed systems symposium*, volume 10, 2016.

Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *25th Annual Network and Distributed System Security Symposium, NDSS 2018, San Diego, California, USA, February 18-21, 2018*, 2018.

Guixin Ye, Zhanyong Tang, Dingyi Fang, Zhanxing Zhu, Yansong Feng, Pengfei Xu, Xiaojiang Chen, and Zheng Wang. Yet another text captcha solver: A generative adversarial network based approach. In *Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security*, pages 332–348, 2018.

Sergey Zagoruyko and Nikos Komodakis. Wide Residual Networks. In *British Machine Vision Conference*, 2016.

Runtian Zhai, Tianle Cai, Di He, Chen Dan, Kun He, John Hopcroft, and Liwei Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pages 7472–7482, 2019.

Huan Zhang, Hongge Chen, Chaowei Xiao, Sven Gowal, Robert Stanforth, Bo Li, Duane S. Boning, and Cho-Jui Hsieh. Towards stable and efficient training of verifiably robust neural networks. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*, 2020.

Shuai Zhao, Liguang Zhou, Wenxiao Wang, Deng Cai, Tin Lun Lam, and Yangsheng Xu. Towards Better Accuracy-efficiency Trade-offs: Divide and Co-training. *arXiv preprint arXiv:2011.14660*, 2020.

Stephan Zheng, Yang Song, Thomas Leung, and Ian Goodfellow. Improving the robustness of deep neural networks via stability training. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 4480–4488, 2016.

A Appendix

A.1 Proofs

Theorem 1. *Let D and $\tilde{D}$ be two distributions supported on $X \times Y$. Then for any classifier $h : X \to Y$*

$$|\mathsf{Rob}_d(h, \tilde{D}) - \mathsf{Rob}_d(h, D)| \leq \mathsf{CWD}_d(D, \tilde{D}).$$

Proof. We first provide the sketch the steps of the proof informally and then formalize the steps after that. Consider D' to be the distribution that is the outcome of the following process: First sample (x, y) from D, then find the closest x' such that $h(x') \neq y$ and output (x', y). By definition, the conditional Wasserestein distance between D and D' is equal to $\mathsf{Rob}_d(h, D)$. Now consider a similar distribution $\tilde{D}'$ corresponding to $\tilde{D}$, so that we have $\mathsf{Rob}_d(h, \tilde{D}) = \mathsf{CWD}(\tilde{D}, \tilde{D}')$. By triangle inequality for Wasserestein distance we have,

$$\mathsf{Rob}_d(h, \tilde{D}) = \mathsf{CWD}(\tilde{D}, \tilde{D}') \leq \mathsf{CWD}(\tilde{D}, D) + \mathsf{CWD}(D, \tilde{D}'). \tag{1}$$

Also, by the way the distributions D' and $\tilde{D}'$ are defined we have

$$\mathsf{CWD}(D, D') \leq \mathsf{CWD}(D, \tilde{D}'). \tag{2}$$

Roughly, the reason behind this is that all examples (x', y) sampled from $\tilde{D}'$ could be seen as an adversarial example for all elements of D with the label y. And we know that D' consists of optimal adversarial examples for D, therefore, the optimal transport between D and D' should be smaller than the optimal transport between D and $\tilde{D}'$. Now combining Inequality 1 and 2 we have

$$\mathsf{Rob}_d(h, \tilde{D}) = \mathsf{CWD}(\tilde{D}, \tilde{D}') \leq \mathsf{CWD}(\tilde{D}, D) + \mathsf{CWD}(D, D') = \mathsf{CWD}(\tilde{D}, D) + \mathsf{Rob}_d(h, D). \tag{3}$$

With a similar argument, because of symmetry of D and $\tilde{D}$, we can also prove

$$\mathsf{CWD}(\tilde{D}, \tilde{D}') = \mathsf{CWD}(\tilde{D}, D) + \mathsf{Rob}_d(h, \tilde{D}). \tag{4}$$

Combining inequalities 3 and 4 finishes the proof. To formalize the proof steps mentioned above, let $J^*_y = \inf_{J \in \mathcal{J}(D|y, \tilde{D}|y)}$ be the optimal transport between the conditional distributions $D \mid y$ and $\tilde{D} \mid y$. We have

$$\begin{aligned}
Rob_d(h, D, d) &= \mathop{\mathbb{E}}_{(.,y) \leftarrow D}\left[\mathop{\mathbb{E}}_{x \leftarrow D|y}\big[\inf_{h(x') \neq y} d(x', x)\big]\right] \\
&= \mathop{\mathbb{E}}_{(.,y) \leftarrow D}\left[\mathop{\mathbb{E}}_{(x,x'') \leftarrow J^*_y}\big[\inf_{h(x') \neq y} d(x', x)\big]\right] \\
&\leq \mathop{\mathbb{E}}_{(.,y) \leftarrow D}\left[\mathop{\mathbb{E}}_{(x,x'') \leftarrow J^*_y}\left[\inf_{h(x') \neq y} d(x'', x') + d(x'', x)\right]\right] \\
&= \mathop{\mathbb{E}}_{(.,y) \leftarrow D}\left[\mathop{\mathbb{E}}_{(x,x'') \leftarrow J^*_y}\left[\inf_{h(x') \neq y} d(x'', x')\right]\right] + \mathop{\mathbb{E}}_{(.,y) \leftarrow D}\left[\mathop{\mathbb{E}}_{(x,x'') \leftarrow J^*_y}[d(x'', x)]\right] \\
&= \mathop{\mathbb{E}}_{(x'',y) \leftarrow D'}\big[\inf_{h(x') \neq y} d(x'', x')\big] + cw_d(D, \tilde{D}) \\
&= Rob_d(\tilde{D}, h) + \mathsf{CWD}_d(D, \tilde{D}).
\end{aligned}$$

$\square$

Corollary 2. *Let D and $\tilde{D}$ be two labeled distributions supported on $X \times Y$ with identical label distributions and let $\bar{D} = p \cdot D + (1-p) \cdot \tilde{D}$ be the weighted mixture of D and $\tilde{D}$. Then for any classifier $h : X \to Y$*

$$|\mathsf{Rob}_d(h, \bar{D}) - \mathsf{Rob}_d(h, D)| \leq (1-p) \cdot \mathsf{CWD}_d(D, \tilde{D}).$$

Proof. We just need to show that $\mathsf{CWD}_d(D, \bar{D}) \leq (1-p) \cdot \mathsf{CWD}_d(D, \tilde{D}))$. Note that since the label distributions are equal, we have

$$\bar{D} \mid y \equiv p \cdot D \mid y + (1-p) \cdot \tilde{D} \mid y$$

. Now let J_y be the optimal transport between $D \mid y$ and $\tilde{D} \mid y$. Now construct a joint distribution $J'_y \equiv (1-p) \cdot J + p \cdot (x, x)_{x \leftarrow D|y}$. Notice that J'_y is a joint distribution with marginals equal to D and $\bar{D}$. Therefor J'_y is a transport between D and $\bar{D}$ and we can calculate its cost. We have

$$\mathop{\mathbb{E}}_{(x,x')\leftarrow J'_y}[d(x,x')] = (1-p)\cdot \mathop{\mathbb{E}}_{(x,x')\leftarrow J_y}[d(x,x')] + \mathop{\mathbb{E}}_{x\leftarrow D|y}[d(x,x)] = (1-p)\cdot \mathop{\mathbb{E}}_{(x,x')\leftarrow J_y}[d(x,x')].$$

Therefore, we have

$$\mathsf{CWD}(D,\bar{D}) \le \mathop{\mathbb{E}}_{(\cdot,y)\leftarrow D}[\mathop{\mathbb{E}}_{(x,x')\leftarrow J'_y}[d(x,x')]] = (1-p)\cdot \mathop{\mathbb{E}}_{(\cdot,y)\leftarrow D}[\mathop{\mathbb{E}}_{(x,x')\leftarrow J_y}[d(x,x')]] = (1-p)\cdot \mathsf{CWD}(D,\tilde{D}).$$

□

Theorem 3 (Tightness of Theorem 1)**.** *For any distribution D supported on $X \times Y$, any classifier h, any homogeneous distance d and any $\epsilon \le \mathsf{Rob}_d(h,D)$, there is a labeled distribution $\tilde{D}$ such that*

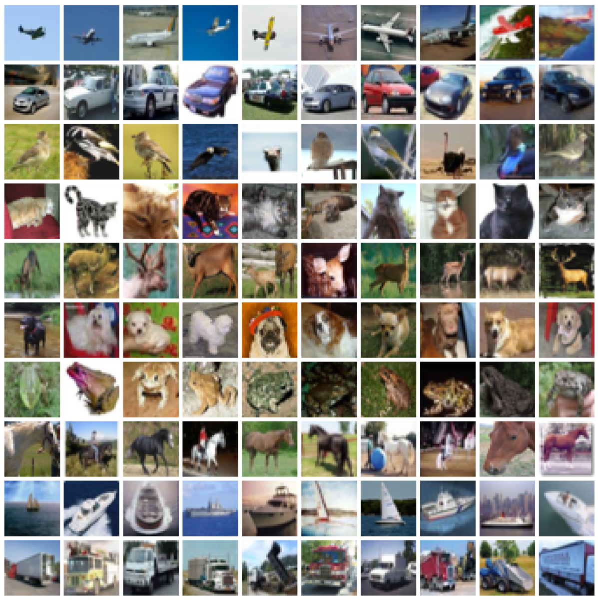
$$\mathsf{Rob}_d(h,D) - \mathsf{Rob}_d(h,\tilde{D}) = \mathsf{CWD}(D,\tilde{D}) = \epsilon.$$

Proof. for $\alpha \in [0,1]$ let $\tilde{D}_\alpha$ be the distribution of the following process: First sample (x,y) from D, then find the closest x' such that $h(x') \ne y$ and output $(x+\alpha(x'-x),y)$. By definition, the conditional Wasserestein distance between D and $\tilde{D}_1$ is equal to $\mathsf{Rob}_d(h,D)$. We also have $\mathsf{CWD}(D,\tilde{D}_\alpha) = \alpha\cdot\mathsf{CWD}(D,\tilde{D}_1)$.

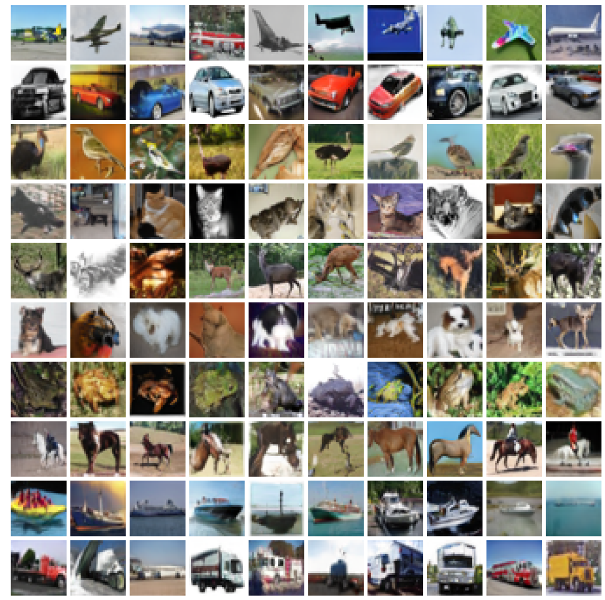
Observe that for any classifier we have $\mathsf{Rob}_d(h,\tilde{D}_\alpha) \le (1-\alpha)\mathsf{Rob}_d(h,D)$ because if (x',y) is an adversarial example for (x,y), then x' is also an adversarial example for $(x+\alpha(x'-x),y)$ with distance $(1-\alpha)d(x,x')$. On the other hand we have $\mathsf{Rob}_d(h,\tilde{D}_\alpha) \ge (1-\alpha)\mathsf{Rob}_d(h,D)$ because any adversarial example for $(x+\alpha(x'-x),y)$ with distance r is also an adversarial example for x with distance at most $r+\alpha d(x'-x)$ and since x' is the optimal adversarial example for x then r must be at least $\alpha(x'-x)$. Therefore, we have $\mathsf{Rob}_d(h,\tilde{D}_\alpha) = (1-\alpha)\mathsf{Rob}_d(h,D)$. Putting everything together and setting $\alpha = \epsilon/\mathsf{Rob}_d(h,D)$we have

$$\mathsf{Rob}_d(h,D) - \mathsf{Rob}_d(h,\tilde{D}_\alpha) = \alpha\mathsf{Rob}_d(h,D) = \alpha\mathsf{CWD}(D,\tilde{D}_1) = \mathsf{CWD}(D,\tilde{D}_\alpha) = \epsilon.$$

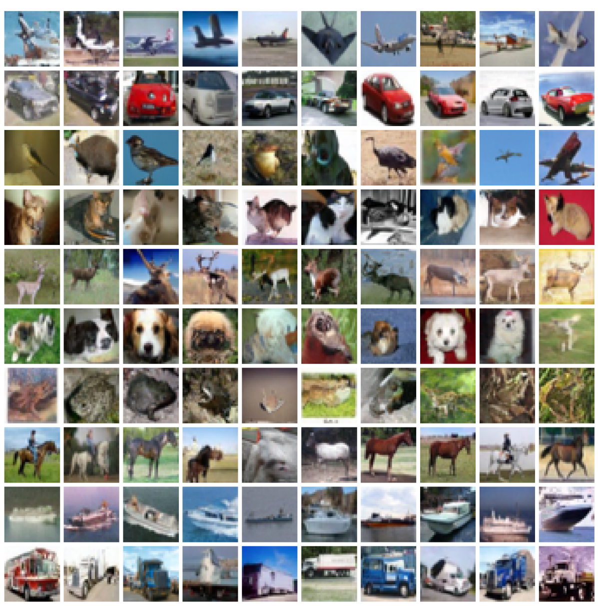
□



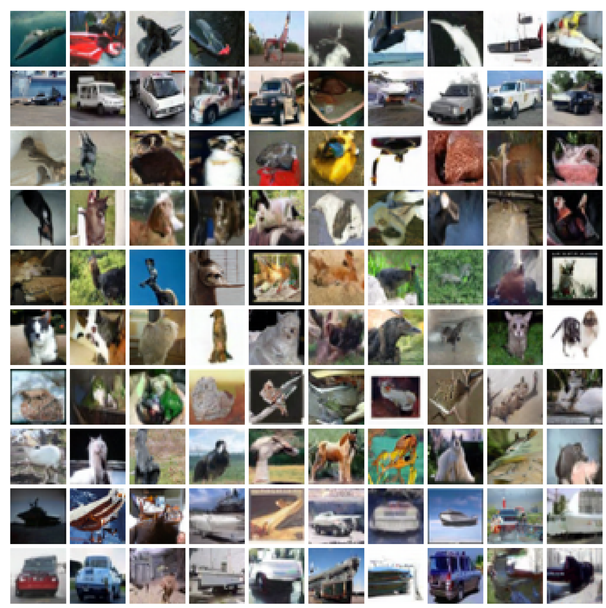
(a) CIFAR-10



(b) StyleGAN



(c) DDPM



(d) Discarded images from DDPM model.

Figure 5: **Visualizing images from different sets.** Randomly selected images form the CIFAR-10 dataset and synthetic images from the StyleGAN (Karras et al., 2020) and DDPM (Ho et al., 2020) model. In figure (d) we show some of the discarded images from the images generated by the DDPM model. Rows in each figure correspond to following classes: Airplane, automobile, bird, cat, deer, dog, frog, horse, ship, and truck.