IMPROVING ROBUSTNESS WITH OPTIMAL TRANSPORT BASED ADVERSARIAL GENERALIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep nets have proven to be brittle against crafted adversarial examples. One of the main reasons is that the representations of the adversarial examples gradually become more divergent from those of the benign examples when feed-forwarding up to higher layers of deep nets. To remedy susceptibility to adversarial examples, it is natural to mitigate this divergence. In this paper, leveraging the richness and rigor of optimal transport (OT) theory, we propose an OT-based adversarial generalization technique that helps strengthen the classifier for tackling adversarial examples. The main idea of our proposed method is to examine a specific Wasserstein (WS) distance between the adversarial and benign joint distributions on an intermediate layer of a deep net, which can further be interpreted from a clustering view of OT as a generalization technique. More specifically, by minimizing the WS distance of interest, an adversarial example is pushed toward the cluster of benign examples sharing the same label on the latent space, which helps to strengthen the generalization ability of the classifier on the adversarial examples. Our comprehensive experiments with state-of-the-art adversarial training and defense on latent space approaches indicate the significant superiority of our method under specific attacks of various distortion sizes. The results demonstrate improvements in robust accuracy up to 5% against PGD attack on CIFAR-100 over the SOTA methods.

1 INTRODUCTION

Despite achieving great success, even state-of-the-art deep neural nets are susceptible to crafted perturbations (Szegedy et al., 2014; Goodfellow et al., 2015). To resolve this severe drawback, many defensive models have been developed (Madry et al., 2018; Zhang et al., 2019b; Xie et al., 2019; Qin et al., 2019). Recently, Athalye et al. (2018) undertook a comprehensive empirical study on a suite of defensive techniques, which identifies obfuscated gradients as the common reason why many defenses give a false sense of defending against gradient-based attacks. This research also reaffirms that adversarial training with Projected Gradient Descent (PGD) (Madry et al., 2018) is one of the most successful and widely-used defensive techniques that remains consistently resilient against attacks. Subsequently, another adversarial training approach, TRADES (Zhang et al., 2019b), has been demonstrated to outperform PGD in defending against attacks.

As indicated by Xie et al. (2019) and Bui et al. (2021), in deep neural nets, the representations on an intermediate layer of the clean data examples and their adversarial counterparts can become highly divergent, while their representations remain proximal in the data space. This observation has inspired a line of work that aims to reduce the divergence between the representations of the clean data examples and their adversarial examples on an intermediate layer of a deep net (Xie et al., 2019; Bui et al., 2020b; 2021). Adversarial Divergence Reduction (Bui et al., 2020a) encourages so-called local and global compactness to enhance adversarial robustness. Moreover, Bui et al. (2021) undertook a comprehensive study to understand the factors influencing the robust accuracy of adversarial defense on the latent space and proposed to adopt contrastive learning for improving adversarial robustness.

In this paper, we leverage optimal transport (OT) theory (Santambrogio, 2015; Villani, 2008), and propose an OT-based adversarial generalization technique that helps to strengthen the classifier for prediction on adversarial examples. Our proposed method is named *Adversarial **G**eneralization*

*with **O**ptimal **T**ransport* (GOT) which can be incorporated into any existing adversarial training approach (e.g., PGD and TRADES) to further improve them. The main idea of GOT is to examine a specific WS distance (Santambrogio, 2015; Villani, 2008) between the adversarial and benign joint distributions on an intermediate layer of a deep net. Using a clustering view of OT, we illustrate that by minimizing the WS distance of interest, an adversarial example is pushed toward the cluster of benign examples that share the same label as it in the latent space. Additionally, the classifier is encouraged to make its prediction for the adversarial example by imitating its predictions for the benign examples in the corresponding cluster. This allows us to make use of global information of the benign examples hence improving the generalization ability of the classifier on the adversarial examples and thereby enhancing adversarial robustness. Furthermore, to more vigorously push the adversarial examples, we propose a label matching variant of GOT that employs a filter to explicitly encourage the matching of the adversarial examples and their corresponding benign examples.

Our contributions in this paper can be summarized as follows:

- We propose an *OT-based adversarial generalization technique* that can help further improve existing adversarial training approaches. The underlying idea is to generalize the classifier to predict well on the adversarial examples by forcing the adversarial examples to move toward the clusters of benign examples with the same labels in the latent space. This is realized by minimizing a relevant WS distance, and interpreted using the clustering view of OT.
- We conduct comprehensive experiments to compare two variants of our proposed method (GOT-S and GOT-LM) with state-of-the-art adversarial training and defense on latent space approaches. The experimental results indicate that our proposed methods significantly outperform the baselines under attacks of various distortion sizes, hence demonstrating the merit of our OT-based adversarial generalization technique in improving adversarial robustness.

2 Related Work

2.1 Adversarial Training Defense

Adversarial training can be traced back to Goodfellow et al. (2015), wherein models were challenged by producing adversarial examples and incorporating them into the training data. The adversarial examples could be the worst-case examples (Goodfellow et al., 2015) or most divergent examples (Zhang et al., 2019b). The quality of the adversarial training defense crucially depends on the strength of the injected adversarial examples – e.g., training on non-iterative adversarial examples obtained from FGSM or Rand FGSM (a variant of FGSM where the initial point is randomized) is not robust to iterative attacks, for example PGD (Madry et al., 2018) or BIM (Kurakin et al., 2016).

Although many defense models were broken by Athalye et al. (2018), adversarial training with PGD (Madry et al., 2018) and TRADES (Zhang et al., 2019c) are among the few defenses that are resilient against attacks. Many defense models were developed based on adversarial examples from PGD or TRADES attacks and try to improve PGD and TRADES in terms of the robust accuracy and training time. Notable examples include Adversarial Logit Pairing (ALP) (Kannan et al., 2018), Feature Denoising (Xie et al., 2019), Defensive Quantization (Lin et al., 2019), Jacobian Regularization (Jakubovitz & Giryes, 2018), Stochastic Activation Pruning (Dhillon et al., 2018), Adversarial Training for Free (Shafahi et al., 2019), and Parameterized Rate-Distortion Stochastic Encoder (Hoang et al., 2020).

2.2 Defense with a Latent Space

The following works have made use of a latent space to realize adversarial defense (Jalal et al., 2017). DefenseGAN (Samangouei et al., 2018) and PixelDefense (Song et al., 2017) utilized a generator (i.e., a pretrained WS-GAN (Gulrajani et al., 2017) for DefenseGAN and a PixelCNN (Oord et al., 2016) for PixelDefense) in conjunction with a latent space to find a denoising version of an adversarial example on the data manifold. These approaches were found by Athalye et al. (2018) as being easy to attack, and it was impossible to defend the attacks on the CIFAR-10 dataset. Jalal et al. (2017) proposed an overpowered attack method to efficiently attack both DefenseGAN and

PixelDefense, and subsequently injected those adversarial examples to train the model. Though that work was proven to work well with simple datasets including MNIST and CelebA, no experiments were conducted on more complex datasets including, for example, CIFAR-10. Adversary Divergence Reduction Network (Bui et al., 2020a) encouraged local and global compactness to improve the robustness. Bui et al. (2021) made a comprehensive study to understand the factors influencing the robust accuracy of adversarial defense on the latent space, and proposed to adopt contrastive learning to improve adversarial robustness.

3 OT-BASED ADVERSARIAL GENERALIZATION

3.1 REVISION OF WASSERSTEIN (WS) DISTANCE

We first revise the definition of a WS distance which serves the development of our proposed approach in a sequel. Let $\mathbb{P}$ and $\mathbb{Q}$ be two discrete distributions on the domain $\Omega \subseteq \mathbb{R}^d$ defined as

$$\mathbb{P} := \sum_{i=1}^m a_i \delta_{\mathbf{u}_i} \text{ and } \mathbb{Q} := \sum_{j=1}^n b_i \delta_{\mathbf{v}_j},$$

where $\delta_{\mathbf{x}}$ indicates a Dirac measure centered at $\mathbf{x}$, $\mathbf{a} = [a_i]_{i=1}^m \in \Delta_m$ and $\mathbf{b} = [b_j]_{j=1}^n \in \Delta_n$ are probability masses, and $\Delta_k := \left\{\boldsymbol{\pi} \in \mathbb{R}^k : \boldsymbol{\pi} \geq \mathbf{0} \text{ and } \|\boldsymbol{\pi}\|_1 = 1\right\}$ is the k-simplex. Consider a non-negative and continuous cost function or metric d on Ω. The WS distance (Santambrogio, 2015; Villani, 2008) between $\mathbb{P}$ and $\mathbb{Q}$ w.r.t. the metric d is defined as

$$\mathcal{W}_d\left(\mathbb{P}, \mathbb{Q}\right) := \min_{R \in \Gamma(\mathbb{P}, \mathbb{Q})} \sum_{i=1}^m \sum_{j=1}^n r_{ij} d\left(\mathbf{u}_i, \mathbf{v}_j\right), \tag{1}$$

where $\Gamma\left(\mathbb{P}, \mathbb{Q}\right)$ is defined as the set of transportation probability matrices

$$\left\{R \in \mathbb{R}_+^{m \times n} : \sum_{j=1}^n r_{ij} = a_i \, \forall i, \text{and} \sum_{i=1}^m r_{ij} = b_j \, \forall j\right\}.$$

3.2 OUR PROPOSED APPROACH

Let $\mathbb{P}^d$ be the data distribution and $\mathcal{D}$ be the joint data-label distribution of the benign pairs $(\mathbf{x}, y)$ with $y \in \{1, ..., M\}$ (i.e., M is the number of classes). Let g be the feature extractor and h be a classifier on top of the feature representations. We denote $f\left(\mathbf{x}\right) = h\left(g\left(\mathbf{x}\right)\right) \in \Delta_M$ where $\Delta_M := \left\{\boldsymbol{\pi} \in \mathbb{R}^M : \|\boldsymbol{\pi}\|_1 = 1\right\}$ is the M-simplex.

Given an adversary $\mathcal{A}$ (e.g., PGD or TRADES, definition in Appendix A.2), we denote $\mathbb{P}^a$ as the distribution of adversarial examples, the distribution including samples $\mathbf{x}^a = \mathcal{A}\left(\mathbf{x}\right)$ for $\mathbf{x} \sim \mathbb{P}^d$. Let $\mathbb{Q}^d, \mathbb{Q}^a$ be the distributions corresponding to $\mathbb{P}^d, \mathbb{P}^a$ on the latent space via the feature extractor g (i.e., $\mathbb{Q}^d = g_{\#}\mathbb{P}^d$ and $\mathbb{Q}^a = g_{\#}\mathbb{P}^a$). We denote $\mathbb{Q}_h^d$ as the joint distribution including pairs $\mathbf{u} = \left(g\left(\mathbf{x}\right), h\left(g\left(\mathbf{x}\right)\right)\right)$ with $\mathbf{x} \sim \mathbb{P}^d$ and $\mathbb{Q}_h^a$ as the joint distribution including pairs $\mathbf{u}^a = \left(g\left(\mathbf{x}^a\right), h\left(g\left(\mathbf{x}^a\right)\right)\right)$ with $\mathbf{x}^a \sim \mathbb{P}^a$.

We consider the WS distance $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ w.r.t. the cost metric d defined as

$$d\left(\mathbf{u}, \mathbf{u}^a\right) = \lambda d_x\left(\mathbf{z}, \mathbf{z}^a\right) + d_y\left(\mathbf{y}, \mathbf{y}^a\right),$$

where $\lambda > 0$ is a trade-off parameter, and $\mathbf{u} = (\mathbf{z}, \mathbf{y})$ and $\mathbf{u}^a = (\mathbf{z}^a, \mathbf{y}^a)$ for which $\mathbf{z} = g\left(\mathbf{x}\right)$, $\mathbf{y} = h\left(\mathbf{z}\right)$, $\mathbf{z}^a = g\left(\mathbf{x}^a\right)$, and $\mathbf{y}^a = h\left(\mathbf{z}^a\right)$ with $\mathbf{x} \sim \mathbb{P}^d$ and $\mathbf{x}^a \sim \mathbb{P}^a$. Here we note that d_x is a distance on the latent space (e.g., Euclidean distance or cosine distance), while d_y is a distance on the M-simplex Δ_M (e.g., the KL divergence or L1 distance).

The following proposition presents an inequality regarding the WS distance of interest, $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$.

Proposition 3.1. *We have the following inequality:* $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right) \geq \lambda \mathcal{W}_{d_x}\left(\mathbb{Q}^d, \mathbb{Q}^a\right)$.

The proof of Proposition 3.1 can be found in Appendix A.1. Inspired by Proposition 3.1, we propose minimizing $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ to boost the robustness of the classifier $f\left(\mathbf{x}\right) = h\left(g\left(\mathbf{x}\right)\right)$. The reason is

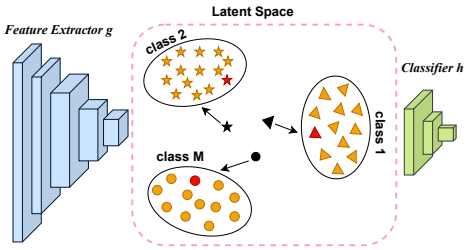

Figure 1: The red points represent the benign examples in focus on the latent space, while the black points represent their adversarial counterparts. By minimizing the *OT-based adversarial regularization term* $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$, the global information of all benign examples can be used to push an adversarial example to an appropriate cluster of the benign examples with the same label, and encourage the classifier h to reduce the mismatch in its predictions.

that by minimizing $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$, we also minimize $\mathcal{W}_{d_x}\left(\mathbb{Q}^d, \mathbb{Q}^a\right)$, which assists us in pushing the adversarial distribution toward the benign data distribution on the latent space for improving robustness. Moreover, by using the clustering view of the WS distance of interest as shown below, we observe that minimizing $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ encourages pushing of the adversarial examples $g\left(\mathbf{x}^a\right)$ to the cluster of the benign examples $g\left(\mathbf{x}\right)$ that shares the same label as $g\left(\mathbf{x}^a\right)$ (see Figure 1). In what follows, we present the clustering view of the WS distance of interest.

Clustering view of the WS distance of interest. To strengthen the robustness of the classifier $f\left(\mathbf{x}\right) = h\left(g\left(\mathbf{x}\right)\right)$, we propose to minimize the *OT-based adversarial regularization term* $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$. Note that we consider $\mathbb{Q}_h^d, \mathbb{Q}_h^a$ as the benign and adversarial empirical distributions on the latent space w.r.t. the training set of benign examples $\{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_N, y_N)\}$, that is

$$\mathbb{Q}_h^d = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{u}_i} \text{ and } \mathbb{Q}_h^a = \frac{1}{N}\sum_{i=1}^{N}\delta_{\mathbf{u}_i^a},$$

where $\mathbf{u}_i = \left(g\left(\mathbf{x}_i\right), h\left(g\left(\mathbf{x}_i\right)\right)\right)$ and $\mathbf{u}_i^a = \left(g\left(\mathbf{x}_i^a\right), h\left(g\left(\mathbf{x}_i^a\right)\right)\right)$.

The *OT-based adversarial regularization term* $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ is written as follows:

$$\min_{R\in\Gamma\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)} \sum_{i=1}^{N}\sum_{j=1}^{N} r_{ij} d\left(\mathbf{u}_i, \mathbf{u}_j^a\right). \tag{2}$$

We assume that each class of the source domain is formed by several clusters on the latent space. Let us denote $I_1^m, ..., I_{M_m}^m \subset \{1, ..., N_S\} (1 \leq m \leq M)$ as the mutually disjoint sets of indices in which $\{g\left(\mathbf{x}_i\right) : i \in I_k^m, 1 \leq k \leq M_m\}$ forms the k-th cluster of the class m. Since the classifier h is supervisorily trained on the source domain with labels, $\{h\left(g\left(\mathbf{x}_i\right)\right) : i \in I_k^m, 1 \leq k \leq M_m\}$ would be the consensus on predicting the source examples in this cluster with the label m. Given an adversarial example $\mathbf{u}_j^a$, we interpret r_{ij} as the probability to transport $\mathbf{u}_j^a$ to $\mathbf{u}_i$ with the cost

$$d\left(\mathbf{u}_i, \mathbf{u}_j^a\right) = \lambda d_x\left(g\left(\mathbf{x}_i\right), g\left(\mathbf{x}_j^a\right)\right) + d_y\left(h\left(g\left(\mathbf{x}_i\right)\right), h\left(g\left(\mathbf{x}_j^a\right)\right)\right).$$

To minimize $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$, the transportation probabilities $r_{ij},\ i = 1, ..., N$ must place the positive values on the source examples in the same class to minimize $\sum_{i=1}^{N} d_y\left(h\left(g\left(\mathbf{x}_i\right)\right), h\left(g\left(\mathbf{x}_j^a\right)\right)\right)$. Meanwhile, the transportation probabilities $r_{ij},\ i = 1, ..., N$ must place the positive values on the source examples in the same cluster to minimize $\sum_{i=1}^{N} d_x\left(g\left(\mathbf{x}_i\right), g\left(\mathbf{x}_j^a\right)\right)$. This leads $g\left(\mathbf{x}_j^a\right)$ to move to an appropriate cluster of $g\left(\mathbf{x}_i\right), i \in I$ with the same label and encourages a reduction of the mismatch in the predictions $h\left(g\left(\mathbf{x}_j^a\right)\right)$ and $h\left(g\left(\mathbf{x}_i\right)\right), i \in I$. In this way, we can leverage the global information of all benign examples when adapting the latent representations of the adversarial examples to the clusters of the benign examples with the same class. This certainly helps to reduce misclassification when predicting on the adversarial examples, and strengthens the generalization capability of the classifier on adversarial examples. This is visualized in Figure 1.

We now incorporate the *OT based generalization term* $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ to PGD-AT and TRADES to form GOT-PGD and GOT-TRADES respectively.

GOT-PGD. The optimization problem of our GOT-PGD is as follows:

$$\inf_{g,h}\left(\mathbb{E}_{\mathcal{D}}\left[\alpha \sup_{\mathbf{x}'\in B_\epsilon(\mathbf{x})} CE\left(h\left(g\left(\mathbf{x}'\right)\right), y\right) + CE\left(h\left(g\left(\mathbf{x}\right)\right), y\right)\right] + \beta\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)\right), \tag{3}$$

where $\beta > 0$ is the trade-off parameter.

GOT-TRADES. The optimization problem of our GOT-TRADES is as follows:

$$\inf_{g,h}\left(\mathbb{E}_{\mathcal{D}}\left[\alpha \sup_{\mathbf{x}'\in B_\epsilon(\mathbf{x})} D_{KL}\left(h\left(g\left(\mathbf{x}'\right)\right), h\left(g\left(\mathbf{x}\right)\right)\right) + CE\left(h\left(g\left(\mathbf{x}\right)\right), y\right)\right] + \beta \mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)\right). \quad (4)$$

Entropic regularization solution. To approximate $\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$, the entropic regularization algorithm can be applied according to Genevay et al. (2016) giving the entropic regularized WS as $\mathcal{R}^{WS} := \mathcal{W}_d^\theta\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$. Based on the Kantorovich dual problem in Genevay et al. (2016), the WS of interest can be defined as

$$\max_{\phi}\left\{-\theta \mathop{\mathbb{E}}_{\mathbb{P}^a}\left[\log\left(\mathop{\mathbb{E}}_{\mathbb{P}^d}\left[\exp\left\{\frac{-d\left(\mathbf{u}, \mathbf{u}^a\right) + \phi(\mathbf{z})}{\theta}\right\}\right]\right)\right] + \mathop{\mathbb{E}}_{\mathbb{P}^d}[\phi(\mathbf{z})]\right\}, \quad (5)$$

where $\mathbf{z} = g\left(\mathbf{x}\right)$ with $\mathbf{x} \sim \mathbb{P}^d$ and $\mathbf{z}^a = g\left(\mathbf{x}^a\right)$ with $\mathbf{x}_a \sim \mathbb{P}^a$, and $\mathbf{u} = \left(\mathbf{z}, h\left(\mathbf{z}\right)\right)$ and $\mathbf{u}^a = \left(\mathbf{z}^a, h\left(\mathbf{z}^a\right)\right)$. In addition, $\theta > 0$ is the entropic regularization parameter (Peyré et al., 2019).

The function ϕ here acting on the latent space is the optimal dual variable, also known as the Kantorovich potential function. In the implementation, we formulate ϕ using a neural network. We refer to the variant with $\mathcal{R}^{WS} := \mathcal{W}_d^\theta\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ defined as in (Equation 5) as GOT-PGD-S and GOT-TRADES-S (i.e., the standard versions) respectively depending on the adversary $\mathcal{A}$ in use.

Label matching variants. Moreover, to tightly match $\mathbf{u}^a = \left(\mathbf{z}^a, h\left(\mathbf{z}^a\right)\right) = \left(g\left(\mathbf{x}^a\right), h\left(g\left(\mathbf{x}^a\right)\right)\right)$ with $\mathbf{x}^a \sim \mathbb{P}^a$ and $\mathbf{u} = \left(g\left(\mathbf{x}\right), h\left(g\left(\mathbf{x}\right)\right)\right)$ with $\mathbf{x} \sim \mathbb{P}^d$ with the same labels y^a (i.e., the label of $\mathbf{x}_a$) and y (i.e., the label of $\mathbf{x}$), we propose using the indicator $\mathbb{I}\left(y^a, y\right)$ which returns 1 if $y^a = y$ and 0 otherwise to focus on minimizing $d\left(\mathbf{u}, \mathbf{u}^a\right)$ with $y^a = y$. This variant is named GOT-PGD-LM and GOT-TRADES-LM (i.e., the label matching versions) respectively. The entropic regularized WS $\mathcal{R}^{WS} := \mathcal{W}_d^\theta\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right)$ for the label matching variant is defined as

$$\max_{\phi}\left\{-\theta \mathop{\mathbb{E}}_{\mathbb{P}^a}\left[\log\left(\mathop{\mathbb{E}}_{\mathbb{P}^d}\left[\mathbb{I}\left(y^a, y\right)\exp\left\{\frac{-d\left(\mathbf{u}, \mathbf{u}^a\right) + \phi(\mathbf{z})}{\theta}\right\}\right]\right)\right] + \mathop{\mathbb{E}}_{\mathbb{P}^d}[\phi(\mathbf{z})]\right\}. \quad (6)$$

The OT loss, defined as the approximation of the WS as above, is then obtained as the maximum value through optimising the ϕ layer. The model we would like to derive is $f = h(g(x))$, where $h(\cdot)$ is the classifier acted on top of latent representations induced by g.

4 Experiments

In this section our aim is to compare the clean and robust accuracy of our proposed model with SOTA defense models like PGD-AT (Madry et al., 2018) and TRADES (Zhang et al., 2019a) as well as benchmark work from Croce et al. (2020), against two adversarial attacks, PGD (Madry et al., 2018) and Auto-Attack (Croce & Hein, 2020). The model will also be investigated with regard to different distortion sizes of the defense as well as other ablation studies.

In our experiments, we use MNIST (Lecun et al., 1998), CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100 (Krizhevsky et al., 2010) and SVHN (Netzer et al., 2011) data sets. We also apply different model structures according to these datasets including CNN (Carlini & Wagner, 2017), ResNet-18 (He et al., 2016) and WideResNet (WRN)-34-10 (Zagoruyko & Komodakis, 2016).

For our proposed method, we generated four variants w.r.t. the two defense models, and whether label matching has been used (i.e., S means the standard version without the label matching, and LM means using the label matching): (i) **GOT-PGD-S**, (ii) **GOT-TRADES-S**, (iii) **GOT-PGD-LM**, and (iv) **GOT-TRADES-LM**. Our detailed experimental settings for different datasets and models can be found in Appendix A.3.

4.1 Performance Evaluation Results

MNIST with CNN model. The experimental results on the MNIST dataset are shown in Table 1. From this table, it is clear that the GOT-AT method can boost both the clean accuracy and the robust

accuracy. The best case for defense is generated from the GOT-TRADES-LM method for which the robust accuracy against PGD attack and Auto-Attack see an increase of 2.4% and 3.1%, respectively, over the PGD-AT baseline scenarios. The clean data testing accuracy for GOT-TRADES-LM also shows an increase compared to the baseline, indicating the improvement in robustness has not hurt the clean accuracy. The improvements with the inclusion of the label matching technique are also obvious from Table 1. Both GOT-TRADES-LM and GOT-PGD-LM have provided 1% higher robuts accuracy under PGD attack and up to 2% under Auto-Attack over the standard ones. The clean accuracy, however, does not enjoy an advantage with the label matching method, and slightly decreases in value.

Table 1: Clean and robust accuracy (in %) comparison on the MNIST dataset against different attacks with distortion $\epsilon = 0.3$ using the Standard CNN architecture.

	Nat	**PGD**	**AA**
PGD-AT	98.99	95.63	91.07
TRADES	98.8	96.55	89.4
MART (Wang et al., 2019)	98.66	94.77	90.06
ARN (Bui et al., 2020a)	99.36	96.96	-
GOT-PGD-S	**99.37**	96.94	92.73
GOT-TRADES-S	99.34	96.82	92.26
GOT-PGD-LM	99.22	97.2	93.91
GOT-TRADES-LM	99.26	**97.79**	**94.15**

Table 2: Clean and robust accuracy (in %) comparison with model ResNet-18 on various datasets with attack distortion $\varepsilon = 8/225$.

Method	CIFAR-10			CIFAR-100			SVHN		
	Nat	**PGD**	**AA**	**Nat**	**PGD**	**AA**	**Nat**	**PGD**	**AA**
PGD-AT	80.76	50.07	48.44	61.95	30.78	24.4	87.3	51.81	40.1
TRADES	78.98	55.59	52.53	62.04	32.61	27.3	87.64	54.14	44.0
GOT-PGD-S	**83.79**	55.69	50.0	62.43	**32.78**	25.8	85.92	52.82	41.7
GOT-TRADES-S	81.79	62.18	58.41	**63.21**	31.88	27.6	92.39	55.65	**44.8**
GOT-PGD-LM	83.65	54.97	51.02	61.83	32.43	**32.2**	84.68	53.1	42.2
GOT-TRADES-LM	81.86	**62.77**	**59.1**	61.48	32.23	30.1	**95.93**	**57.91**	44.5

CIFAR-10, CIFAR-100 and SVHN with ResNet-18 model. Table 2 compares the robustness results for different defense models on CIFAR-10, CIFAR-100 and SVHN datasets using the ResNet-18 model structure. For CIFAR-10, GOT methods show much improved robustness, with the robust accuracy enhanced by over 7% under a PGD attack, and 6% under Auto-Attack. The testing accuracy also increased by 2% compared to the best case in the baseline, showing that the classification accuracy has also been maintained under the defensive model. GOT-PGD has not shown such good results, but can still beat PGD-AT by over 2% in defending against PGD attacks for GOT-PGD-S and 3.5% for GOT-PGD-LM. The employment of the label matching mechanism is also shown to be effective in the CIFAR-10 case. In addition, the label matching variants GOT-PGD-LM and GOT-TRADES-LM further improve the standard variants GOT-PGD-S and GOT-TRADES-S. Similar performance appears also on SVHN, where the robust accuracy sees an improvement of 3% for a PGD attack and 2% for Auto-Attack. The label matching methods are more effective under PGD attacks. In the case of CIFAR-100, GOT-PGD-S has seen an improvement of 2% over the baseline under a PGD attack and GOT-PGD-LM sees a 8% increase for Auto-Attack. The label matching technique can help improve the accuracy compared with the standard one, especially under Auto-Attack.

CIFAR-10 and CIFAR-100 with WRN-34-10 model. In addition, model structure WRN-34-10 has also been applied for further experiments on CIFAR-10 and CIFAR-100. The results for WRN have been compared to the baseline presented by benchmarks in Croce et al. (2020). According to the results in Table 3, CIFAR-10 dataset experiments show results with the applied GOT methods

Table 3: Clean and robust accuracy (in %) comparison with model WRN-34-10 on various datasets with attack distortion $\varepsilon = 8/225$. Benchmarks sourcing from Croce et al. (2020).

Method	CIFAR-10			CIFAR-100		
	Nat	**PGD**	**AA**	**Nat**	**PGD**	**AA**
Proxy (Sehwag et al., 2021)	85.85	59.09	-	-	-	-
LGBAT (Cui et al., 2020)	**88.22**	52.86	-	62.55	30.20	-
AWP (Wu et al., 2020)	85.36	59.09	-	60.38	28.86	-
ATES (Sitawarin et al., 2020)	86.84	50.72	-	62.82	24.57	-
GOT-PGD-S	84.62	56.41	52.2	63.8	34.22	33.0
GOT-TRADES-S	82.70	62.75	58.5	**64.56**	34.02	31.0
GOT-PGD-LM	83.84	57.75	52.6	63.24	**34.57**	**33.9**
GOT-TRADES-LM	83.25	**63.25**	**59.3**	64.13	34.27	32.5

of 4% higher than the benchmark results under the PGD attack. However, the clean accuracy of the models including GOT have not seen better performance comparably. In the results regarding CIFAR-100, there has been an enhancement by 5% under a PGD attack for robust accuracy and 4% for clean accuracy. The label matching technique can still show its advantages over the standard ones for both datasets in WRN.

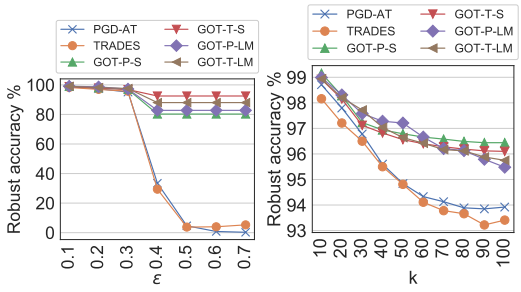

Figure 2: Robustness comparison on the MNIST dataset against PGD attack with $\eta = 0.01$, while varying $\epsilon \in [0.1, 0.7]$ k=40 (left), $k \in [10, 100]$ ϵ = 0.3 (right).

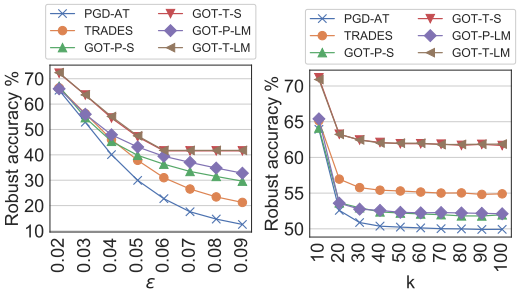

Figure 3: Robustness comparison on the CIFAR-10 dataset against PGD attack at η = 0.003, while varying $\epsilon \in$ [0.02, 0.1], k=20 (left), $k \in$ [10, 100],ϵ = 8/255 (right).

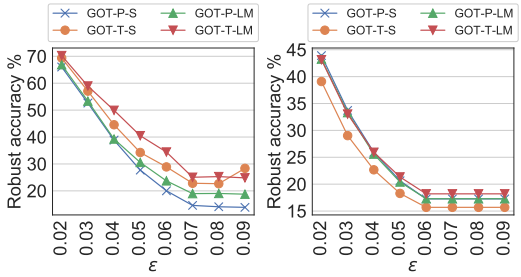

Figure 4: Robustness accuracy for SVHN (left) and CIFAR-100 (right) against PGD attack with varying $\epsilon \in$ [0.02, 0.1], k=20 in model ResNet-18.

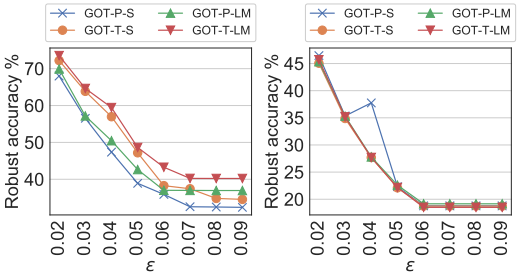

Figure 5: Robustness accuracy for CIFAR-10 (left) and CIFAR-100 (right) against PGD attack with varying $\epsilon \in$ [0.02, 0.1], k=20 in model WRN-34-10.

Impact of Distortion Size and Number of Steps. Figures 2 and 3 (and Appendix A.4) show the impact of distortion size on robust accuracy trend under the PGD attacks, using MNIST and CIFAR-10 datasets. It can be observed that the GOT-AT method performs much more stably as distortion size ϵ increases. For the MNIST dataset, the robust accuracy is maintained at approximately 92.5% for GOT-TRADES-S and 80.3% for GOT-PGD-S. The robustness again to different numbers of steps for PGD attacks does not vary that much as there has been a higher level of approximately 95.3% for GOT-PGD-LM and 96.5% for GOT-PGD-S than the baselines.

Table 4: The influence of the latent layer position to the clean and robust accuracy (in %) on the MNIST dataset against PGD attack with $k = 40, \epsilon = 0.3, \eta = 0.01$.

Methods	Layer	**Nat**	**PGD**	**AA**
GOT-PGD-S	Penultimate	99.37	96.94	92.73
GOT-TRADES-S	Layer	99.34	96.82	92.26
GOT-PGD-S	Middle	99.22	95.32	86
GOT-TRADES-S	Layer	99.24	94.38	87.8

Looking at the changing distortion size for CIFAR-10, the GOT-AT method can provide a robust accuracy of over 41.5% with increasing perturbation size from attacks. This also shows that our proposed method is less sensitive to the change of perturbation size from attacks on CIFAR-10. Moreover, robustness regarding the distortion size can be kept with the label matching mechanism employed. The changes of the number of steps from PGD attack influences the robustness. The GOT-TRADES methods can keep the accuracy at 61.9% and GOT-PGD gains also relatively high accuracy with the increasing number of steps.

From Figures 4 and 5, we can see CIFAR-100 and SVHN's results under PGD attacks with various distortion sizes in ResNet18 and WRN-34-10, respectively. The detailed data for the distortion size discussion can also be found in Appendix A.4. Generally, with the increasing distortion sizes, the trend of robust model performance will decrease, sharply at the beginning and will slow down to converge later. For CIFAR-10, the WRN structured model will have higher value with smaller distortion and will decrease more with increasing sizes. For CIFAR-100, the trend does not show much influence and the WRN models are generally better than the ResNet models after the trends flatten.

Label Distribution Clustering Visualization. In order to better present the effectiveness of our GOT methods, the label distributions from the penultimate layers have been visualised in Figure 6 for both baseline model TRADES (left) and the proposed model GOT-TRADES-LM (right). It is clear from the figures that the proposed model can push the adversarial examples to an appropriate cluster of the benign examples sharing same original classes through the proceeding of the model. The samples with the same original classes are better clustered for the proposed model which will result in more enhanced robustness.

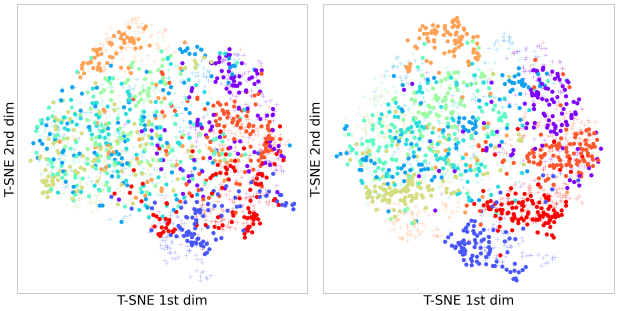

Figure 6: Label distribution from penultimate layers for CIFAR-10 against PGD-20, based on baseline model TRADES (left), and proposed model GOT-TRADES-LM (right).

4.2 Ablation Study

Latent Space Position. We conduct an ablation study to investigate the influence of the latent space position in which we apply our OT-based adversarial regularization technique. We experiment on MNIST with the simple CNN architecture. We consider two places for the latent layer: i) a middle layer and ii) the penultimate layer. Note that if we employ the penultimate layer as the latent layer, the classifier h on the top of the latent layer is a linear classifier. Middle layer refers to the outputs of the last max-pool layer before the fully connected layer. The latent space position does not influence the clean accuracy much, while the robust accuracy seems to be improved if we establish the latent layer more closely to the output layer. This makes sense because the closer the latent layer to the output layer, the more the adversarial latent representations tend to diverge from the benign latent representations.

Parameter Sensitivity. We investigate the sensitivity of our proposed method w.r.t. the parameter β (i.e., the weight of the OT-based adversarial regularization term). The searching of β is conducted

on MNIST against PGD attacks with $k = 40, \epsilon = 0.3, \eta = 0.01$, $\beta \in [0.01, 100]$. As shown in Figure 7, the robust accuracy trends peaks at $\beta = 10$, but the performance has not seen significant influence with the changing β. Accordingly, in experiments for MNIST, β has been set as 10.

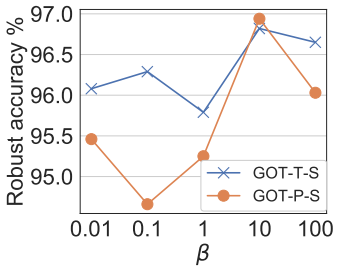

Figure 7: Robustness comparison on the MNIST dataset regarding changing OT loss trade off.

Table 5: Clean and robust accuracy (in %) comparison for different choices of d_x and d_y on the MNIST dataset against PGD attack with $k = 40, \epsilon = 0.3, \eta = 0.01$.

Methods	d_x/d_y	**Nat**	**PGD**	**AA**
GOT-TRADES-S	COS/KL	**99.34**	**96.82**	**92.26**
	COS/L1	98.86	96.5	92.2
	L2/KL	98.83	96.31	91.8
	L2/L1	98.04	93.44	89.1
GOT-PGD-S	COS/KL	**99.37**	**96.94**	**92.73**
	COS/L1	99.33	93.75	84.9
	L2/KL	99.12	95.22	87.9
	L2/L1	98.4	91.89	78.5

Influence of Cost Metrics d_x ***and*** d_y**.** We investigate the influence of the metrics d_x and d_y used to define the metric d to the robust accuracy. Basically, we consider some options for d_x including cosine distance (COS) and L2 distance, and some options for d_y including KL divergence and L1 distance. As shown in Table 5, the combination of the cosine distance for d_x and KL divergence for d_y is the best choice. Therefore, we apply this combination in the main experiments.

5 Conclusion

Deep nets are brittle against crafted adversarial examples. In this paper, by leveraging optimal transport (OT) theory, we propose an OT-based adversarial generalization technique that strengthens a classifier to improve adversarial robustness. The underlying idea of our proposed method is to investigate a specific WS distance between the adversarial and benign joint distributions on an intermediate layer of a deep net. More specifically, by minimizing the WS distance of interest, an adversarial example is pushed toward the cluster of benign examples sharing the same label as it on the latent space. Additionally, the classifier is encouraged to mitigate the mismatch for its prediction on the adversarial example and its predictions on the benign examples in the corresponding cluster, which helps to strengthen the generalization ability of the classifier on the adversarial examples. Comprehensive experiments with state-of-the-art adversarial training and defense on latent space approaches indicate the significant superiority of our proposed method under specific and various distortion size attacks.

References

Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. In *European Conference on Computer Vision*, pp. 484–501. Springer, 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*, pp. 274–283, 2018.

Anh Bui, Trung Le, He Zhao, Paul Montague, Olivier deVel, Tamas Abraham, and Dinh Phung. Improving adversarial robustness by enforcing local and global compactness. *arXiv preprint arXiv:2007.05123*, 2020a.

Anh Bui, Trung Le, He Zhao, Paul Montague, Olivier Y. DeVel, Tamas Abraham, and Dinh Q. Phung. Improving adversarial robustness by enforcing local and global compactness. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVII*, volume 12372 of *Lecture Notes in Computer Science*, pp. 209–223. Springer, 2020b.

Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. Understanding and achieving efficient robustness with adversarial contrastive learning. *CoRR*, abs/2101.10027, 2021.

N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pp. 39–57. IEEE, 2017.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. *arXiv preprint arXiv:2003.01690*, 2020.

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

Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. *arXiv preprint arXiv:2011.11164*, 2020.

G S. Dhillon, K. Azizzadenesheli, Z C. Lipton, J. Bernstein, J. Kossaifi, A. Khanna, and A. Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

A. Genevay, M. Cuturi, G. Peyré, and F. Bach. Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems 29*, pp. 3440–3448. Curran Associates, Inc., 2016. URL http://papers.nips.cc/paper/6566-stochastic-optimization-for-large-scale-optimal-transport.pdf.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In Yoshua Bengio and Yann LeCun (eds.), *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6572.

I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A C. Courville. Improved training of wasserstein gans. In *Advances in neural information processing systems*, pp. 5767–5777, 2017.

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2016.

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*, pp. 770–778, 2016.

Quan Hoang, Trung Le, and Dinh Phung. Parameterized rate-distortion stochastic encoder. In Hal DaumÃ© III and Aarti Singh (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 4293–4303. PMLR, 13–18 Jul 2020.

D. Jakubovitz and R. Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European Conference on Computer Vision*, pp. 514–529, 2018.

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.

H. Kannan, A. Kurakin, and I. Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). *URL http://www. cs. toronto. edu/kriz/cifar. html*, 5:4, 2010.

Alex Krizhevsky et al. Learning multiple layers of features from tiny images. 2009.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016.

Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

J. Lin, C. Gan, and S. Han. Defensive quantization: When efficiency meets robustness. *arXiv preprint arXiv:1904.08444*, 2019.

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

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

A. vd Oord, N. Kalchbrenner, and K. Kavukcuoglu. Pixel recurrent neural networks. *arXiv preprint arXiv:1601.06759*, 2016.

G. Peyré, M. Cuturi, et al. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607, 2019.

Chongli Qin, James Martens, Sven Gowal, Dilip Krishnan, Krishnamurthy Dvijotham, Alhussein Fawzi, Soham De, Robert Stanforth, and Pushmeet Kohli. Adversarial robustness through local linearization. In *Advances in Neural Information Processing Systems*, pp. 13824–13833, 2019.

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

F. Santambrogio. Optimal transport for applied mathematicians. *Birkäuser, NY*, pp. 99–102, 2015.

Vikash Sehwag, Saeed Mahloujifar, Tinashe Handina, Sihui Dai, Chong Xiang, Mung Chiang, and Prateek Mittal. Improving adversarial robustness using proxy distributions. *arXiv preprint arXiv:2104.09425*, 2021.

A. Shafahi, M. Najibi, M A. Ghiasi, Z. Xu, J. Dickerson, C. Studer, L S. Davis, G. Taylor, and T. Goldstein. Adversarial training for free! In *Advances in Neural Information Processing Systems*, pp. 3353–3364, 2019.

Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Improving adversarial robustness through progressive hardening. *arXiv preprint arXiv:2003.09347*, 2020.

Y. Song, T. Kim, S. Nowozin, S. Ermon, and N. Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. *arXiv preprint arXiv:1710.10766*, 2017.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In Yoshua Bengio and Yann LeCun (eds.), *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL `http://arxiv.org/abs/1312.6199`.

C. Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008. ISBN 9783540710509.

Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *arXiv preprint arXiv:2004.05884*, 2020.

C. Xie, Y. Wu, L v d. Maaten, A L Yuille, and K. He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 501–509, 2019.

S. Zagoruyko and N. Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 7472–7482, 2019a.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P. Xing, Laurent El Ghaoui, and Michael I. Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019b.

Y. Zhang, Y. Liu, M. Long, and M. I. Jordan. Bridging theory and algorithm for domain adaptation. *CoRR*, abs/1904.05801, 2019c. URL `http://arxiv.org/abs/1904.05801`.

A Appendix

A.1 Proofs

We first prove an important lemma. Given a pair of distribution $\mathbb{Q}_A$ and deterministic classifer h_A, we define $\mathbb{Q}_{h_A}^A$ as a distribution including sample pair $(\mathbf{z}, h_A(\mathbf{z}))$ by first sampling $\mathbf{z} \sim \mathbb{P}_A$ and then computing $h_A(\mathbf{z})$. Similarly, we can define $\mathbb{Q}_{h_B}^B$ for another pair of distribution $\mathbb{Q}_B$ and deterministic classifer h_B. We define a distance d between $\mathbf{u}_1 = (\mathbf{z}_1, h_A(\mathbf{z}_1))$ and $\mathbf{u}_2 = (\mathbf{z}_2, h_B(\mathbf{z}_2))$ as

$$d(\mathbf{u}_1, \mathbf{u}_2) = \lambda d_z(\mathbf{z}_1, \mathbf{z}_2) + d_y(h_A(\mathbf{z}_1), h_B(\mathbf{z}_2))$$

where d_z and d_y are two distances on the space $\mathcal{Z}$ of $\mathbb{Q}_A, \mathbb{Q}_B$ and the simplex Δ_M respectively.

Lemma A.1. *The WS distance of interest can be expressed as:*

$$\begin{aligned}\mathcal{W}_d\left(\mathbb{Q}_{h_A}^A, \mathbb{Q}_{h_B}^B\right) &= \min_{L: L_{\#}\mathbb{Q}_A = \mathbb{Q}_B} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}_A}\left[\lambda d_z(\mathbf{z}, L(\mathbf{z})) + d_y(h_A(\mathbf{z}), h_B(L(\mathbf{z})))\right] \\ &= \min_{K: K_{\#}\mathbb{Q}_B = \mathbb{Q}_A} \mathbb{E}_{\mathbf{z}\sim\mathbb{P}_B}\left[\lambda d_z(\mathbf{z}, K(\mathbf{z})) + d_\dagger(h_B(\mathbf{z}), h_A(K(\mathbf{z})))\right].\end{aligned}$$

Proof. Observe first that for any $U_A \subset Z \times \Delta_M$, we have $\mathbb{Q}_{h_A}^A(U_A) = \mathbb{Q}_A(V_A)$ where $V_A \coloneqq \{\mathbf{z} \in \mathcal{Z} \mid (\mathbf{z}, h_A(\mathbf{z})) \in U_A\}$. Similarly, we have for any $U_B \subset Z \times \Delta_M$ that $\mathbb{Q}_{h_B}^B(U_B) = \mathbb{Q}_B(V_B)$ where $V_B \coloneqq \{\mathbf{z} \in \mathcal{Z} \mid (\mathbf{z}, h_B(\mathbf{z})) \in U_B\}$.

Let $H : \operatorname{supp}\left(\mathbb{Q}_{h_A}^A\right) \to \operatorname{supp}\left(\mathbb{Q}_{h_B}^B\right)$ (i.e., supp indicates the support of a distribution) be such that $H_{\#}\mathbb{Q}_{h_A}^A = \mathbb{Q}_{h_B}^B$. We can express H as

$$H(\mathbf{z}, h_A(\mathbf{z})) := (H_1(\mathbf{z}, h_A(\mathbf{z})), H_2(\mathbf{z}, h_A(\mathbf{z}))),$$

with $H_1(\mathbf{z}, h_A(\mathbf{z})) \in \mathcal{Z}$ and $H_2(\mathbf{z}, h_A(\mathbf{z})) \in \Delta_M$. Define $L(\mathbf{z}) := H_1(\mathbf{z}, h_A(\mathbf{z}))$. We claim that $L_{\#}\mathbb{Q}_A = \mathbb{Q}_B$. Indeed, let $V_B \subset \mathcal{Z}$ be any measurable set and take $U_B \coloneqq V_B \times \triangle_M$. Then by using the observation above and the fact $H_{\#}\mathbb{Q}_{h_A}^A = \mathbb{Q}_{h_B}^B$, we obtain

$$\mathbb{Q}_B(V_B) = \mathbb{Q}_{h_B}^B(U_B) = \mathbb{Q}_{h_A}^A\left(H^{-1}(U_B)\right) = \mathbb{Q}_{h_A}^A\left(L^{-1}(V_B) \times \triangle_M\right) = \mathbb{Q}_A\left(L^{-1}(V_B)\right).$$

Thus the claim is proved.

It also follows from $H_{\#}\mathbb{Q}_{h_A}^A = \mathbb{Q}_{h_B}^B$ and the claim that $H_2(\mathbf{z}, h_A(\mathbf{z})) = h_B(L(\mathbf{z}))$, which gives

$$d((\mathbf{z}, h_A(\mathbf{z})), H(\mathbf{z}, h_A(\mathbf{z}))) = \lambda d_z(\mathbf{z}, L(\mathbf{z})) + d_y(h_A(\mathbf{z}), h_B(L(\mathbf{z}))). \tag{7}$$

Therefore, we deduce that

$$\mathcal{W}_d\left(\mathbb{Q}_{h_A}^A, \mathbb{Q}_{h_B}^B\right) \geq \min_{L: L_{\#}\mathbb{Q}_A = \mathbb{Q}_B} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}_A}\left[\lambda d_z(\mathbf{z}, L(\mathbf{z})) + d_y(h_A(\mathbf{z}), h_B(L(\mathbf{z})))\right]$$

In order to prove the reverse inequality, let us consider any map L satisfying $L_{\#}\mathbb{Q}_A = \mathbb{Q}_B$. Define $H(\mathbf{z}, h_A(\mathbf{z})) \coloneqq (L(\mathbf{z}), h_B(L(\mathbf{z})))$. Then (7) holds and $H_{\#}\mathbb{Q}_{h_A}^A = \mathbb{Q}_{h_B}^B$. To verify the latter, let $U_B \subset Z \times \triangle_M$ be any measurable set and take $V_B \coloneqq \{\mathbf{z} \in \mathcal{Z} \mid (\mathbf{z}, h_B(\mathbf{z})) \in U_B\}$. Then as

$$H^{-1}(U_B) = \{(\mathbf{z}, h_A(\mathbf{z})) \mid L(\mathbf{z}) \in V_B\} = \left\{(\mathbf{z}, h_A(\mathbf{z})) \mid \mathbf{z} \in L^{-1}(V_B)\right\},$$

we have

$$\mathbb{Q}_{h_A}^A \left(H^{-1}\left(U_B\right)\right) = \mathbb{Q}_A \left(L^{-1}\left(V_B\right)\right) = \mathbb{Q}_B\left(V_B\right) = \mathbb{Q}_{h_B}^B\left(U_B\right).$$

Thus it follows that

$$\mathcal{W}_d\left(\mathbb{Q}_{h_A}^A, \mathbb{Q}_{h_B}^B\right) \leq \min_{L: L_{\#}\mathbb{Q}_A = \mathbb{Q}_B} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}_A}\left[\lambda d_z\left(\mathbf{z}, L\left(\mathbf{z}\right)\right) + d_y\left(h_A\left(\mathbf{z}\right), h_B\left(L\left(\mathbf{z}\right)\right)\right)\right].$$

By combining the above two inequalities, we obtain the equality

$$\mathcal{W}_d\left(\mathbb{Q}_{h_A}^A, \mathbb{Q}_{h_B}^B\right) \leq \min_{L: L_{\#}\mathbb{Q}_A = \mathbb{Q}_B} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}_A}\left[\lambda d_z\left(\mathbf{z}, L\left(\mathbf{z}\right)\right) + d_y\left(h_A\left(\mathbf{z}\right), h_B\left(L\left(\mathbf{z}\right)\right)\right)\right].$$

Symmetrically, we achieve the other equality. □

Proof of Proposition 3.1. By applying Lemma A.1 for $\mathbb{Q}^d = \mathbb{Q}_A$, $\mathbb{Q}^a = \mathbb{Q}_B$, and $h = h_A = h_B$, we reach

$$\begin{aligned}\mathcal{W}_d\left(\mathbb{Q}_h^d, \mathbb{Q}_h^a\right) &= \min_{L: L_{\#}\mathbb{Q}^d = \mathbb{Q}^a} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}^d}\left[\lambda d_z\left(\mathbf{z}, L\left(\mathbf{z}\right)\right) + d_y\left(h\left(\mathbf{z}\right), h\left(L\left(\mathbf{z}\right)\right)\right)\right] \\ &\geq \min_{L: L_{\#}\mathbb{Q}^d = \mathbb{Q}^a} \mathbb{E}_{\mathbf{z}\sim\mathbb{Q}^d}\left[\lambda d_z\left(\mathbf{z}, L\left(\mathbf{z}\right)\right)\right] = \lambda \mathcal{W}_d\left(\mathbb{Q}^d, \mathbb{Q}^a\right).\end{aligned}$$

□

A.2 State-of-the-Art Adversarial Training Methods

Let f be the classifier that needs to be strengthened and $\mathcal{D}$ be the joint distribution of data-label pairs $(\mathbf{x}, y)$. The underlying idea of adversarial training (AT) is to seek the most challenging data instances and incorporate them into the training process to strengthen the classifier f.

PGD-AT (Madry et al., 2018) seeks the *worst-case* examples and uses them to improve model robustness:

$$\inf_f \mathbb{E}[\alpha \sup_{\mathbf{x}'\in B_\epsilon(\mathbf{x})} CE\left(f\left(\mathbf{x}'\right), y\right) + CE\left(f\left(\mathbf{x}\right), y\right)], \tag{8}$$

where $B_\epsilon\left(\mathbf{x}\right) = \{\mathbf{x}' : \|\mathbf{x} - \mathbf{x}'\| \leq \epsilon\}$, $\alpha > 0$ is the trade-off parameter, and CE is the cross-entropy loss.

TRADES (Zhang et al., 2019a) seeks the *most divergent* examples and uses them to improve model robustness:

$$\inf_f \mathbb{E}[\alpha \sup_{\mathbf{x}'} D_{KL}\left(f\left(\mathbf{x}'\right), f\left(\mathbf{x}\right)\right) + CE\left(f\left(\mathbf{x}\right), y\right)], \tag{9}$$

where $\mathbf{x}' \in B_\epsilon\left(\mathbf{x}\right)$ and D_{KL} is the usual Kullback-Leibler (KL) divergence.

A.3 Experimental Settings

All the data sets were normalized to $[0, 1]$. We apply padding of 4 pixels at all borders before random cropping and random horizontal flips as used in Zhang et al. (2019c). We train our proposed method and baselines in 100 epochs. The adversarial samples generalisation is based on the input testing data and the predicted labels instead of ground truth labels. Following provides more detailed settings of each data set.

MNIST We use the standard CNN architecture (Carlini & Wagner, 2017) for the MNIST experiments with Adam optimizer on learning rate 1×10^{-3} (adjusted to 1×10^{-6} in label matching case). In the adversarial training settings, We use $\{k = 40, \epsilon = 0.3, \eta = 0.01\}$, where k is the number of iteration steps, ϵ is the distortion bound and η is the step size of the adversaries. For our proposed variants, the parameter of adversarial loss trade off α in Equation 3 is set to 1 and in 4 is set as 6. In the training for GOT-PGD-S and GOT-PGD-LM, only the adversarial loss was used for adversarial training, which provides better robust accuracy.

CIFAR-10, CIFAR-100 and SVHN For the these three datasets, We have applied both the ResNet-18 (He et al., 2016) and WRN-34-10 (Zagoruyko & Komodakis, 2016) in our experiments. Stochastic Gradient Decent (SGD) optimiser with momentum on learning rate 1×10^{-4} and weight decay 2×10^{-4} is used. For adversarial training in these datasets, the settings be as $\{k = 20, \epsilon = 8/255, \eta = 2/255\}$. For both CIFAR-10 and CIFAR-100, the parameter of adversarial loss trade off α is set to 1 for PGD and 6 for TRADES cases. Also, only the adversarial loss was used in case of GOT-PGD. While for SVHN, the adversarial loss trade off α in TRADES this is set as 3.0. In the training for GOT-PGD cases, 2 adversarial loss and 0.5 clean loss have been included to maintain both clean and robust accuracy.

We use different SOTA attacks to evaluate the defense methods including: (i) **PGD attack** (Madry et al., 2018) with l_∞ distortion metric and full testing set of $10,000$ test samples. For the PGD attack, the random restart is 1 and 40 iterations for MNIST while 20 iterations for other datasets have been chosen. (ii) **Auto-Attack (AA)** (Croce & Hein, 2020) which is an ensemble based attack. The standard version of the attack has been used, which provides an ensemble of four different attacks including APGD-CE, APGD-DLR, FAB (Croce & Hein, 2020) and the Square Attack (Andriushchenko et al., 2020). In Auto-attacks the version is 'standard' and the adversarial samples generalisation is based on the input testing data and the predicted labels instead of ground truth labels. The distortion metric we use in our experiments is l_∞ and only $1,000$ test samples have been used to measure in CIFAR-10(WRN), CIFAR-100 and SVHN datasets.

The OT loss trade off β is searched in $\{0.01, 0.1, 1, 10, 100\}$ in ablation study for MNIST while set as 10 for MNIST and 10 in all other experiments scenarios. In Equations 5 and 6, θ is consistently set to 0.1, λ is set to 1.0. Regarding the cost metric d of the WS distance, as pointed out by our ablation study, we choose the cosine distance for d_x and the KL divergence for d_y. The ϕ, Kantorovich potential function, is a linear real value layer with the inputs as the features extracted from the penultimate layer of the original classification model.

A.4 Distortion Size Influence Detailed Data

Table 6: Clean and robust accuracy (in %) comparison with model Standard CNN on MNIST with different attack distortions.

	MNIST						
ϵ	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**
PGD-AT	98.77	97.36	95.15	33.47	4.65	0.8	0.26
GOT-PGD-S	99.1	98.03	96.94	80.26	80.26	80.26	80.26
GOT-PGD-LM	99.12	98.66	97.2	82.76	82.76	82.76	82.76
TRADES	98.23	96.95	95.71	29.4	3.77	3.98	5.28
GOT-TRADES-S	99	97.86	96.27	84.61	84.61	84.61	84.61
GOT-TRADES-LM	99.05	98.17	97.56	88.01	88.01	88.01	88.01

Table 7: Clean and robust accuracy (in %) comparison with model WRN-34-10 on various datasets with different attack distortions.

	CIFAR-10							
ϵ	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
GOT-PGD-S	68.00	56.54	47.36	38.90	35.90	32.55	32.48	32.39
GOT-PGD-LM	69.90	57.18	50.45	42.63	37.01	36.98	36.95	36.95
GOT-TRADES-S	72.20	63.85	56.97	47.18	38.27	37.43	34.81	34.55
GOT-TRADES-LM	73.54	64.63	59.45	48.63	43.27	40.24	40.22	40.22
	CIFAR-100							
ϵ	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
GOT-PGD-S	46.46	35.34	37.75	22.31	18.79	18.79	18.79	18.79
GOT-PGD-LM	45.33	35.31	27.8	22.6	19.16	19.16	19.16	19.16
GOT-TRADES-S	45.06	34.88	27.69	22.1	18.74	18.74	18.74	18.74
GOT-TRADES-LM	45.7	35.18	27.7	22.17	18.51	18.51	18.51	18.51

Table 8: Clean and robust accuracy (in %) comparison with model ResNet-18 on various datasets with different attack distortions.

	CIFAR-10							
ϵ	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
PGD-AT	65.49	52.87	40.08	29.94	22.73	17.56	14.68	12.57
GOT-PGD-S	67.02	54.65	45.32	39.81	36.32	33.46	31.15	28.64
GOT-PGD-LM	66.03	56	47.91	43.1	39.44	37.04	34.79	32.77
TRADES	66.41	56.25	46	37.84	30.99	26.53	23.42	21.25
GOT-TRADES-S	72.19	63.59	54.53	47.1	41.56	41.56	41.57	41.56
GOT-TRADES-LM	72.38	63.62	55.1	47.48	41.76	41.76	41.76	41.76
	SVHN							
ϵ	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
GOT-PGD-S	66.03	52.64	38.85	27.72	20.03	14.60	14.09	13.88
GOT-PGD-LM	66.88	53.35	39.26	30.57	23.72	18.98	19.04	18.77
GOT-TRADES-S	69.36	57.06	44.56	34.26	28.90	22.83	22.66	28.36
GOT-TRADES-LM	70.26	58.96	49.93	40.50	34.43	25.06	25.30	24.83
	CIFAR-100							
ϵ	**0.02**	**0.03**	**0.04**	**0.05**	**0.06**	**0.07**	**0.08**	**0.09**
GOT-PGD-S	43.91	33.72	25.93	20.63	17.26	17.26	17.26	17.26
GOT-PGD-LM	43.26	33.31	25.59	20.4	17.29	17.29	17.29	17.29
GOT-TRADES-S	39.09	29.03	22.67	18.28	15.69	15.69	15.69	15.69
GOT-TRADES-LM	43.14	33.03	25.93	21.34	18.2	18.2	18.2	18.2