Bridging Optimal Transport and Jacobian Regularization by Optimal Trajectory for Enhanced Adversarial Defense

Binh M. Le[1], Shahroz Tariq[2], and Simon S. Woo[1]*

[1] Dept. of Computer Science & Engineering, Sungkyunkwan University
{bmle,swoo}@g.skku.edu
[2] CSIRO's Data61
shahroz.tariq@data61.csiro.au

Abstract. Deep neural networks, particularly in vision tasks, are notably susceptible to adversarial perturbations. To overcome this challenge, developing a robust classifier is crucial. In light of the recent advancements in the robustness of classifiers, we delve deep into the intricacies of adversarial training and Jacobian regularization, two pivotal defenses. Our work is the first carefully analyzes and characterizes these two schools of approaches, both theoretically and empirically, to demonstrate how each approach impacts the robust learning of a classifier. Next, we propose our novel Optimal Transport with Jacobian regularization method, dubbed OTJR, bridging the input Jacobian regularization with the a output representation alignment by leveraging the optimal transport theory. In particular, we employ the Sliced Wasserstein distance that can efficiently push the adversarial samples' representations closer to those of clean samples, regardless of the number of classes within the dataset. The SW distance provides the adversarial samples' movement directions, which are much more informative and powerful for the Jacobian regularization. Our empirical evaluations set a new standard in the domain, with our method achieving commendable accuracies of 52.57% on CIFAR-10 and 28.36% on CIFAR-100 datasets under the AutoAttack. Further validating our model's practicality, we conducted real-world tests by subjecting internet-sourced images to online adversarial attacks. These demonstrations highlight our model's capability to counteract sophisticated adversarial perturbations, affirming its significance and applicability in real-world scenarios.

1 Introduction

Deep Neural Networks (DNNs) have established themselves as the de facto method for tackling challenging real-world machine learning problems. Their applications cover a broad range of domains, such as image classification, object detection, and recommendation systems. Nevertheless, recent research has revealed DNNs' severe vulnerability to adversarial examples [7,18], particularly

* Corresponding Author

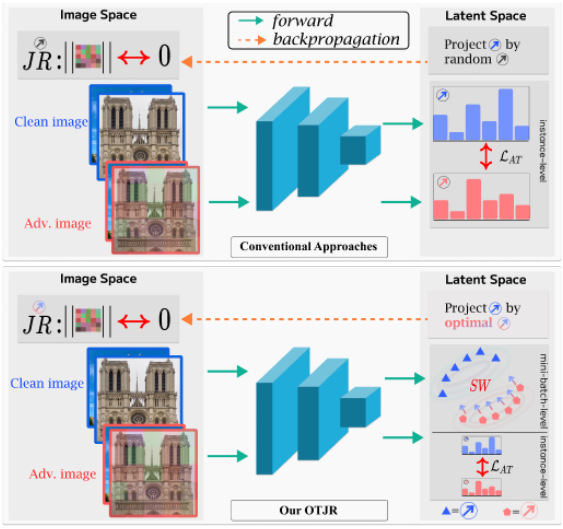

Fig. 1: Illustration of (top) two popular approaches to boost a model's robustness: Adversarial Training (AT) vs. Jacobian regularization (JR), and (bottom) our OTJR method. JR tries to silence the Jacobian matrix at the input end. The AT adjusts the distribution of perturbed samples at the output end. In conventional approach, JR backpropagates through random projections, whereas the AT via a loss function. Our proposed OTJR bridges AT and JR on framework by the optimal transport theory.

in computer vision tasks. Small imperceptible perturbations added to the image can easily deceive the neural networks into making incorrect predictions with high confidence. Moreover, this unanticipated phenomenon raises social concerns about DNNs' safety and trustworthiness, as they can be abused to attack many sophisticated and practical machine learning systems putting human lives into danger, such as in autonomous car [13] or medical systems [3, 22].

Meanwhile, there are numerous studies that devote their efforts to enhance the robustness of various models against adversarial examples. Among the existing defenses, adversarial training (AT) [18, 23] and Jacobian regularization (JR) [17, 19] are the two most predominant and popular defense approaches In AT, small perturbations are added to a clean image in its neighbor of L_p norm ball to generate adversarial samples. Thus, an adversarially trained model can force itself to focus more on the most relevant image's pixels. On the other hand, the second approach, Jacobian regularization, mitigates the effect of the perturbation to the model's decision boundary by suppressing its gradients. However, AT and Jacobian regularization have not been directly compared in both theoretical and empirical settings.

In this work, we embark on a dual-path exploration, offering both theoretical and empirical comparisons between AT and Jacobian regularization. Our objective is to deepen our comprehension of the adversarial robustness inherent to DNN models and subsequently enhance their defensive capability. While a myriad of prior research has spotlighted defense, they predominantly adopt either an empirical or theoretical lens, rarely both. To bridge this gap, we introduce an innovative approach, integrating both Jacobian regularization and AT. This fusion seeks to augment the adversarial robustness and defensive efficacy of a model, as elucidated in Fig. 1.

For AT, a plethora of studies have been proposed, presenting unique strategies to encourage the learning of robust classifiers. [20, 23, 31, 37, 38]. Notably, Sinkhorn Adversarial Training (SAT) [4] resonates with our methodology, particularly in its objective to bridge the distributional gap between clean and adversarial samples using optimal transport theory. However, the pillar of their algorithms mainly relies on the Sinkhorn algorithm [12] to utilize the space discretization property [33]. Therefore, their approach has several limitations in terms of handling high-dimensional data [24, 28]. Particularly, the Sinkhorn algorithm blurs the transport plan by adding an entropic penalty to ensure the optimization's convexity. The entropic penalty encourages the randomness of the transportation map. However, in high-dimensional spaces, such randomness reduces the deterministic movement plan of one sample, causing ambiguity. As a result, when training defense models on a large scale dataset, SAT results in a slow convergence rate, which is unjustifiable due to the introduction of additional training epochs with distinct learning schedules [27].

To address the outlined challenges, we present our pioneering method, Optimal Transport with Jacobian Regularization, denoted as OTJR, designed explicitly to bolster defenses against adversarial intrusions. We leverage the Sliced Wasserstein (SW) distance, which is more efficient for AT in high dimensional space with a faster convergence rate. In addition, the SW distance provides us with other advantages due to optimal latent trajectories of adversarial samples in the embedding space, which is critical for designing an effective defense. We further integrate the input-output Jacobian regularization by substituting its random projections with the optimal trajectories and constructing the optimal Jacobian regularization. Our main contributions are summarized as follows:

(1) Comprehensive Theoretical and Empirical Examination. Our research delves deep into the theoretical underpinnings of adversarial robustness, emphasizing the intricacies of AT and Jacobian Regularization. Distinctively, we pioneer a simultaneous theoretical and empirical analysis, offering an incisive, comparative exploration. This endeavor serves to elucidate the differential impacts that each methodology has on the design and efficacy of defensive DNNs.

(2) Innovative Utilization of the Sliced Wasserstein (SW) Distance. Charting new territory, we introduce the integration of the SW distance within our AT paradigm, denoted as OTJR. This innovation promises a marked acceleration in the training convergence, setting it apart from extant methodologies. Harnessing the prowess of the SW distance, we discern the optimal trajectories for adversarial samples within the latent space. Subsequently, we weave these optimal vectors into the framework of Jacobian Regularization, augmenting the resilience of DNNs by expanding their decision boundaries.

(3) Rigorous Evaluation against White- and Black-box Attacks. Our exhaustive experimental assessments underscore the superiority of our methodology. Pitted against renowned state-of-the-art defense strategies, our approach consistently emerges preeminent, underscoring the potency of enhancing Jacobian Regularization within the AT spectrum as a formidable defensive arsenal.

AT Framework	Training Objective
TRADES	$\mathcal{L}_{\mathcal{XE}}(\mathbb{S}(z), y) + \lambda \mathcal{L}_{\mathcal{XE}}(\mathbb{S}(\tilde{z}), \mathbb{S}(z))$
PGD-AT	$\mathcal{L}_{\mathcal{XE}}(\mathbb{S}(\tilde{z}), y)$
ALP	$\alpha \mathcal{L}_{\mathcal{XE}}(\mathbb{S}(z), y) + (1-\alpha)\mathcal{L}_{\mathcal{XE}}(\mathbb{S}(\tilde{z}), y) + \lambda \|\tilde{z} - z\|_2$

Table 1: Training objectives of the popular AT frameworks, where $\mathbb{S}$ denotes the softmax function.

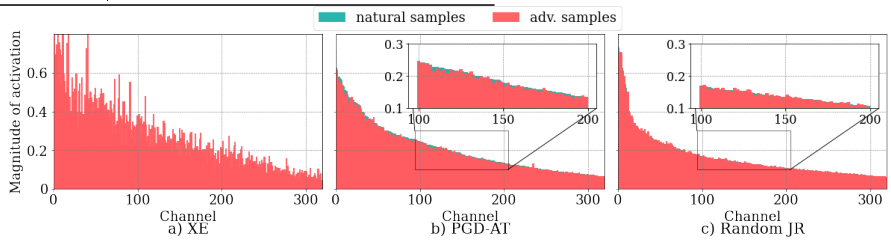

Fig. 2: The magnitude of activation at the penultimate layer for models trained with $\mathcal{XE}$ loss, PGD-AT adversarial training, and the input-output Jacobian regularization. The channels in the X-axis are sorted in descending order of the clean samples' magnitude.

2 Comparisons between AT and JR

2.1 Theoretical Preliminaries

Let a function f represent a deep neural network (DNN), which is parameterized by θ, and $x \in \mathbb{R}^I$ be a clean input image. Its corresponding output vector is $z = f(x) \in \mathbb{R}^C$, where z_c is proportional to the likelihood that x is in the c-th class. Also, let $\tilde{x} = x + \epsilon$ be an adversarial sample of x generated by adding a small perturbation vector $\epsilon \in \mathbb{R}^I$. Then, the Taylor expansion of the mapped feature of the adversarial sample with respect to ϵ is derived as follows:

$$\begin{aligned} \tilde{z} = f(x+\epsilon) = f(x) + J(x)\epsilon + O(\epsilon^2) \simeq z + J(x)\epsilon, \\ ||\tilde{z} - z||_q \simeq ||J(x)\epsilon||_q. \end{aligned} \tag{1}$$

Precisely, Eq. 1 is a basis to derive two primary schools of approaches, AT vs. Jacobian regularization, for mitigating adversarial perturbations and boosting the model robustness. In particular, each side of Eq. 1 targets the following two objectives, to improve the robustness:

1) Aligning adversarial representation (AT). Minimizing the left-hand side of Eq. 1 is to push the likelihood of an adversarial sample $\tilde{x}$ close to that of a clean sample x. For instance, the Kullback-Leibler divergence between two likelihoods is a popular AT framework such as TRADES [37]. More broadly, the likelihood differences can include the cross-entropy ($\mathcal{XE}$) loss of adversarial samples such as ALP [20], PGD-AT [23], or FreeAT [32]. These well-known AT frameworks are summarized in Table 1, to explain their core learning objective.

2) Regularizing input-output Jacobian matrix (JR). Regularizing the right-hand side of Eq. 1, which is independent of $\tilde{x}$, suppresses the spectrum of the input-output Jacobian matrix $J(x)$. Thus, the model becomes more stable with respect to input perturbation, as it was theoretically and empirically demonstrated in a line of recent research [8, 17, 19]. Particularly, by observing

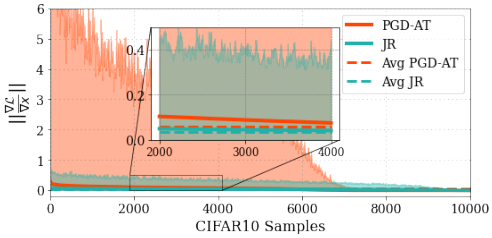

Fig. 3: Magnitude of $||\nabla_{\tilde{x}}\mathcal{L}_{\mathcal{XE}}||_1$ at the input layer for a model trained with PGD-AT and Jacobian regularization. Red and green-filled areas range from min. to max. values of each sample.

$||J(x)\epsilon||_q \leq ||J(x)||_F||\epsilon||$, one can instead minimize the square of the Frobenius norm of the Jacobian matrix, which can be estimated as follows [17]:

$$||J(x)||_F^2 = C\mathbb{E}_{\hat{v}\sim\mathcal{S}^{C-1}}\left[||\hat{v}\cdot J||^2\right], \tag{2}$$

where $\hat{v}$ is an uniform random vector drawn from a C-dimensional unit sphere $\mathcal{S}^{C-1}$. Using Monte Carlo method to approximate the integration of $\hat{v}$ over the unit sphere, Eq. 2 can be rewritten as:

$$||J(x)||_F^2 = \frac{1}{n_{\text{proj}}}\sum_{i=1}^{n_{\text{proj}}}\left[\frac{\partial(\hat{v}_i\cdot z)}{\partial x}\right]^2. \tag{3}$$

Considering a large number of samples in a mini-batch, n_{proj} is usually set to 1 for the efficient computation. Hence, the Jacobian regularization is:

$$||J(x)||_F^2 \simeq \left[\frac{\partial(\hat{v}\cdot z)}{\partial x}\right]^2, \hat{v}\sim\mathcal{S}^{C-1}. \tag{4}$$

Summary. As shown above, each approach takes different direction to tackle adversarial perturbations. While AT targets aligning the likelihoods at the output end, Jacobian regularization forces the norm of the Jacobian matrix at the input to zero, as also pictorially described in Fig. 1. However, their distinct effects on the robustness of a DNN have not been fully compared and analyzed.

2.2 Empirical Analysis

We also conduct an experimental analysis to ascertain and characterize the distinct effects of the two approaches (AT vs. Jacobian regularization) on a defense DNN when it is trained with each of the objectives. We use wide residual network WRN34 as the preliminary baseline architecture on CIFAR-10 dataset and apply two canonical frameworks: PGD-AT [23] and input-output Jacobian regularization [17] to enhance the model's robustness. Details of training settings are provided in the experiment section.

Output and input sides. We measure the magnitude of channel-wise activation to characterize the connection between adversarial defense methods and the *penultimate layer's activation* [2]. Figure 2 provides the average magnitude of activations of clean vs. adversarial samples created by PGD-20 attacks [23]. As shown, not only AT [2], but also Jacobian regularization can effectively suppress the magnitude of the activation. Moreover, the Jacobian regularization

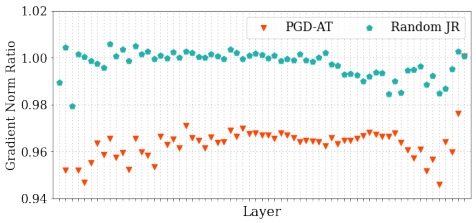

Fig. 4: Ratios of $\mathbb{E}(||\nabla_{\theta_i}\mathcal{L}(\tilde{x})||/||\nabla_{\theta_i}\mathcal{L}(x)||)$ w.r.t. the model's parameters θ_i on CIFAR-10. The lower the ratios are, the more emphasis the model puts on perturbations.

typically achieves the lower magnitude value of the activation compared to that of PGD-AT. This observation serves as a clear counter-example to earlier results from [2], where they claim that adversarial robustness can be generally achieved via channel-wise activation suppressing. As such, it is worth noting that while a more effective defense strategy can produce lower activation, the inverse is not always true. In addition, Fig. 3 represents the average gradient of $\mathcal{XE}$ loss with respect to the adversarial samples, at the *input layer*. This demonstrates that the model trained with Jacobian regularization suppresses input gradients more effectively than a typical AT framework, *i.e.*, PGD-AT. In other words, when a defensive model is abused to generate adversarial samples, pre-training with Jacobian regularization can reduce the severity of perturbations, hindering the adversary's target.

Layer-by-layer basis. We further provide Fig. 4 to depict the gradients of models trained with PGD-AT, and Jacobin regularization, respectively. Particularly, we compute the norm ratios of the loss gradient on the adversarial sample to the loss gradient on the clean samples for each layer of the model. As we can observe, the model trained with the Jacobian regularization produces higher ratio values, meaning it puts less emphasis on adversarial samples due to its agnostic defense mechanism. Meanwhile, most of the ratio values at the middle layers from Jacobian training vary around 1.This is explained by the regularization applied to its first derivatives. In summary, we can also empirically conclude that the Jacobian regularization tends to silence the gradient of the model from output to input layers. Therefore, it *agnostically* stabilizes the model under the changes of input samples, and produces low-magnitude adversarial perturbations, when the model is attacked. In contrast, by learning the meaningful pixels from input images, AT adjusts the model's parameters at every layer in such a way to reduce the impacts of adversarial perturbation on the model's outputs.

Our motivation. As elucidated by [17], the computational overhead introduced by training models with Jacobian regularization is marginal compared to standard training regimes. Therefore, a combination of AT and Jacobian regularization becomes an appealing approach for the adversarial robustness of a model. Furthermore, taking advantages from both approaches can effectively render a classifier to suppress the perturbation and adaptively learn crucial features from both clean and adversarial samples. However, merely adding both approaches together into the training loss is not the best option. Indeed it is insufficient, since the adversarial representations in the latent space can contain meaningful information for the Jacobian regularization, which we will discuss more in the next section.

Hence, in this work, we propose a novel optimization framework, OTJR, to leverage the movement direction information of adversarial samples in the latent space and optimize the Jacobian regularization. In this fashion, we can successfully establish a relationship and balance between silencing input's gradients and aligning output distributions, and significantly improve the overall model's robustness. Additionally, recent studies proposed approaches utilizing a surrogate model [35] or teacher-student framework [11] during training. Yet, while these methods improve the model's robustness, they also rely on previous training losses (such as TRADE or PGD). And, they introduce additional computation for the AT, which are so far well-known for their slow training speed and computational overhead. In our experiment, we show that our novel training loss can be compatible with these frameworks and further improve the model's robustness by a significant margin compared to prior losses.

3 Our Approach

Our approach explores the Sliced Wasserstein distance in order to push the adversarial distribution closer to the natural distribution with a faster convergence rate. Next, ***the Sliced Wasserstein distance results in optimal movement directions to sufficiently minimize the spectrum of input-output Jacobian matrix.***

3.1 Sliced Wasserstein Distance

The p-Wasserstein distance between two probability distributions μ and ν [34] in a general d-dimensional space Ω, to search for an optimal transportation cost between μ and ν, is defined as follows:

$$W_p(P_\mu, P_\nu) = \Big(\inf_{\pi \in \Pi(\mu,\nu)} \int_{\Omega \times \Omega} \psi(x,y)^p d\pi(x,y) \Big)^{1/p}, \tag{5}$$

where $\psi : \Omega \times \Omega \to \mathbb{R}^+$ is a transportation cost function, and $\Pi(\mu,\nu)$ is a collection of all possible transportation plans. The Sliced p-Wasserstein distance (SW_p), which is inspired by the W_p in one-dimensional, calculates the

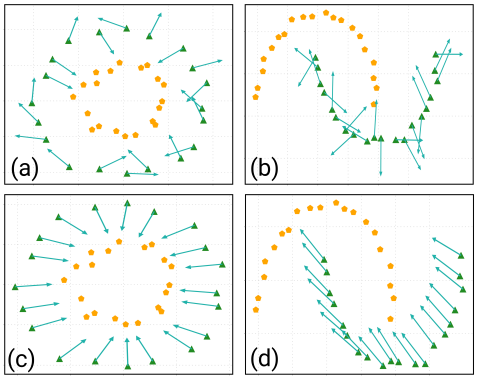

Fig. 5: Illustration of optimal latent trajectories. **Top row** ((a) & (b)): Random movement directions (green arrows), which is non-informative, uniformly sampled from two-dimensional unit sphere $\mathcal{S}^1$. **Bottom row** ((c) & (d)): The optimal trajectories from the SW distance between source distribution and target distribution obtained from Eq. 10.

p-Wasserstein distance by projecting μ and ν onto multiple one-dimensional marginal distributions using Radon transform [16]. The SW_p is defined as:

$$SW_p(\mu, \nu) = \int_{\mathcal{S}^{d-1}} W_p(\mathcal{R}_{(t,\hat{v})}\mu, \mathcal{R}_{(t,\hat{v})}\nu) d\hat{v}, \tag{6}$$

where $\mathcal{R}_{(t,\hat{v})}\mu$ is the Radon transform as follows:

$$\mathcal{R}_{(t,\hat{v})}\mu = \int_{\Omega} \mu(x)\sigma(t - \langle \hat{v}, x \rangle) dx, \forall \hat{v} \in \mathcal{S}^{d-1}, \forall t \in \mathbb{R}, \tag{7}$$

where $\langle \cdot, \cdot \rangle$ denote the Euclidean inner product, and σ is the Dirac delta function.

Next, let $\mathcal{B}$ denote the size of a mini-batch of samples, and C denote the number of classes. To calculate the transportation cost between adversarial samples' representations $\mu := \{\tilde{z}_i\}_{\mathcal{B}}$ and the corresponding original samples' representations $\nu := \{z_i\}_{\mathcal{B}}$, the integration of $\hat{v}$ over the unit sphere $\mathcal{S}^{C-1}$ is approximated via Monte Carlo method with K uniformly sampled random vector $\hat{v}_i \in \mathcal{S}^{C-1}$. In particular, $\mathcal{R}_{(t,\hat{v})}\mu$ and $\mathcal{R}_{(t,\hat{v})}\nu$ are sorted in ascending order using two permutation operators τ_1 and τ_2, respectively, and the approximation of Sliced *1*-Wasserstein is expressed as follows:

$$SW(\mu, \nu) \simeq \sum_{k=1}^{K} \psi \left(\tau_1 \circ \mathcal{R}_{(t,\hat{v}_k)}\mu, \tau_2 \circ \mathcal{R}_{(t,\hat{v}_k)}\nu \right). \tag{8}$$

3.2 Optimal Latent Trajectory

Using Eq. 8, we can straightforwardly compute the trajectory of each $\tilde{z}_i$ in the latent space $\mathbb{R}^C$ in order to minimize the $SW(\mu, \nu)$. We refer to the directions of these movements as *optimal latent trajectories*, because the SW distance produces the lowest transportation cost between the source and target distributions. Particularly, let the trajectories of $\{\tilde{z}_i\}_{\mathcal{B}}$ in each single projection $\hat{v}_k$ be:

$$m_k = (\tau_1^{-1} \circ \tau_2 \circ \mathcal{R}_{(t,\hat{v}_k)}\nu - \mathcal{R}_{(t,\hat{v}_k)}\mu) \otimes \hat{v}_k, \tag{9}$$

where $\otimes$ denotes the outer product. Then, the overall optimal trajectory direction of each $\tilde{z}_i$ is expressed as follows:

$$\sigma_i = \frac{\sum_{k=1}^{K} m_{k,i}}{|| \sum_{k=1}^{K} m_{k,i} ||_2}. \tag{10}$$

A demonstrative example is provided in Fig. 5 to illustrate the concept of optimal trajectories. This example highlights the trajectory direction of an adversarial sample, denoted as $\tilde{x}$, within the latent space. In particular, this direction is significant, as it represents the most sensitive axis of perturbation for x when exposed to adversarial interference. To address this, our approach diverges from the traditional method of random projection, as previously proposed by Hoffman et al. (2019) [17]. Instead, we propose to regularize the Jacobian matrix of x specifically along this identified optimal trajectory. Utilizing Eq. 4, we

Table 2: Classification accuracy (%) under *white-box*, *black-box* attacks and *AutoAttack*. Different defense methods trained on CIFAR-10 and CIFAR-100 datasets using WRN34 in 100 epochs, except that SAT^{400} was trained in 400 epochs.

Dataset	Defense	*Clean*	PGD^{20}	PGD^{100}	L_2-PGD	MIM	FGSM	CW	FAB	Square	SimBa	*AutoAtt*
CIFAR-10	TRADES	*84.71*$_{.15}$	54.23$_{.07}$	53.91$_{.11}$	61.04$_{.21}$	54.13$_{.11}$	60.46$_{.15}$	53.17$_{.22}$	53.65$_{.39}$	62.25$_{.26}$	70.66$_{.52}$	52.06$_{.15}$
	ALP	*86.63*$_{.23}$	46.99$_{.25}$	46.48$_{.25}$	55.69$_{.33}$	46.88$_{.23}$	58.14$_{.21}$	47.50$_{.30}$	**55.61**$_{.27}$	59.26$_{.28}$	68.28$_{.31}$	46.28$_{.25}$
	PGD-AT	*86.54*$_{.20}$	46.67$_{.11}$	46.23$_{.12}$	55.93$_{.10}$	46.66$_{.09}$	56.83$_{.23}$	47.56$_{.15}$	48.39$_{.22}$	58.92$_{.12}$	68.66$_{.14}$	45.77$_{.07}$
	SAT	*83.19*$_{.33}$	53.52$_{.15}$	53.23$_{.23}$	60.37$_{.06}$	52.46$_{1.34}$	60.36$_{.13}$	52.15$_{.20}$	52.54$_{.36}$	61.50$_{.19}$	69.70$_{.29}$	50.73$_{.20}$
	Random JR	*84.99*$_{.14}$	22.67$_{.15}$	21.89$_{.14}$	60.98$_{.19}$	22.49$_{.18}$	32.99$_{.21}$	22.00$_{.06}$	21.74$_{.11}$	45.86$_{.20}$	71.20$_{.31}$	20.54$_{.14}$
	SW	*84.26*$_{.80}$	54.51$_{.33}$	54.20$_{.37}$	61.08$_{.12}$	54.46$_{.27}$	61.32$_{.12}$	53.78$_{.05}$	54.90$_{.29}$	62.95$_{.43}$	70.74$_{.65}$	51.97$_{.06}$
	OTJR (***ours***)	*84.01*$_{.53}$	**55.38**$_{.29}$	**55.08**$_{.36}$	**63.87**$_{.09}$	**55.31**$_{.29}$	**61.03**$_{.18}$	**54.09**$_{.12}$	54.17$_{.07}$	**63.11**$_{.21}$	**72.04**$_{.68}$	**52.57**$_{.12}$
CIFAR-100	TRADES	57.46$_{.16}$	30.42$_{.05}$	30.36$_{.09}$	35.85$_{.10}$	30.37$_{.08}$	33.04$_{.16}$	27.97$_{.13}$	27.93$_{.15}$	33.70$_{.06}$	44.28$_{.14}$	27.15$_{.09}$
	ALP	*60.61*$_{.05}$	26.23$_{.18}$	25.87$_{.17}$	33.75$_{.09}$	26.14$_{.07}$	31.40$_{.08}$	25.78$_{.22}$	25.69$_{.15}$	33.09$_{.15}$	43.25$_{.22}$	24.57$_{.22}$
	PGD-AT	*59.77*$_{.21}$	24.05$_{.03}$	23.74$_{.03}$	31.41$_{.16}$	24.01$_{.06}$	29.21$_{.18}$	24.67$_{.04}$	24.47$_{.06}$	31.47$_{.11}$	41.15$_{.21}$	23.28$_{.01}$
	SAT^{400}	*53.61*$_{.52}$	26.63$_{.14}$	26.42$_{.16}$	32.34$_{.25}$	26.57$_{.19}$	31.04$_{.21}$	25.22$_{.06}$	26.63$_{.11}$	31.32$_{.34}$	39.64$_{.90}$	24.32$_{.05}$
	Random JR	*66.58*$_{.17}$	9.41$_{.44}$	8.87$_{.42}$	37.79$_{.31}$	9.27$_{.49}$	16.38$_{.33}$	10.27$_{.45}$	9.26$_{.23}$	23.53$_{.15}$	48.86$_{.55}$	8.10$_{.57}$
	SW	*57.69*$_{.28}$	26.01$_{.16}$	25.82$_{.24}$	31.37$_{.17}$	25.97$_{.18}$	30.78$_{.34}$	25.48$_{.23}$	26.11$_{.07}$	31.34$_{.26}$	40.68$_{.43}$	24.35$_{.29}$
	OTJR (***ours***)	*58.20*$_{.13}$	**32.11**$_{.21}$	**32.01**$_{.18}$	**43.13**$_{.12}$	**32.07**$_{.19}$	**34.26**$_{.30}$	**29.71**$_{.06}$	**29.24**$_{.08}$	**36.27**$_{.05}$	**49.92**$_{.23}$	**28.36**$_{.10}$

are able to reformulate the input-output Jacobian regularization, incorporating these strategically derived projections for each sample. The formula for this new regularization approach is presented below:

$$||J(x_i|\sigma_i)||_F^2 \simeq \left[\frac{\partial(\sigma_i \cdot z_i)}{\partial x_i}\right]^2. \tag{11}$$

Then, our overall loss function for a batch of samples $\{(x_i, y_i)\}_{\mathcal{B}}$ is expressed in the following way:

$$\mathcal{L} = \sum_{i=1}^{\mathcal{B}} \left(\mathcal{L}_{\text{AT}}(\tilde{x_i}, y_i) + \lambda_J ||J(x_i|\sigma_i)||_F^2\right) + \lambda_{SW} SW(\mu, \nu), \tag{12}$$

where $\mathcal{L}_{\text{AT}}$, unless stated otherwise, is cross-entropy loss of adversarial samples. In practice, sampling K uniform vectors $\hat{v}_k$ can be performed simultaneously thanks to deep learning libraries. Then, the calculation of random projections and optimal trajectory steps can be vectorized and performed simultaneously.

4 Experimental Results

4.1 Experiment Settings

We employ WideResNet34-10 [36] as our backbone architecture on two datasets CIFAR-10 and CIFAR-100 [21]. The model are trained for 100 epochs with the momentum SGD [29] optimizer, whereas its initial learning rate is set to 0.1 and decayed by 10 at 75^{th} and 90^{th} epoch, respectively. Adversarial samples in the training phase are generated by L_∞-PGD [23] in 10 iterations with the maximal perturbation $\epsilon = 8/255$ and the perturbation step size $\eta = 2/255$. For a fair comparison with different approaches [27], we use the above settings throughout our experiments without early stopping or modifying models' architecture, and report their performances on the last epoch. For our OTJR based defense models, we

Defense	CIFAR-10-WEB			CIFAR-100-WEB		
	$k = 100$	$k = 200$	$k = 500$	$k = 100$	$k = 200$	$k = 500$
TRADES	$78.1_{3.0}$	$75.6_{2.1}$	$46.9_{1.4}$	$84.7_{3.4}$	$85.5_{2.2}$	$84.9_{1.5}$
ALP	$78.6_{2.9}$	$77.1_{2.0}$	$49.1_{1.4}$	$85.1_{4.6}$	$85.6_{2.3}$	$85.3_{1.7}$
PGD-AT	$79.2_{2.6}$	$79.8_{2.1}$	$51.4_{1.6}$	$84.9_{3.6}$	$86.2_{2.9}$	$86.3_{0.9}$
SAT	$77.7_{5.4}$	$72.8_{3.2}$	$48.6_{2.1}$	$86.9_{2.6}$	$87.2_{1.8}$	$87.2_{0.6}$
Random JR	$82.3_{3.9}$	$82.5_{2.8}$	$63.7_{1.3}$	$90.5_{2.9}$	$90.1_{1.7}$	$90.0_{1.4}$
SW	$77.6_{4.2}$	$72.0_{2.3}$	$46.0_{1.6}$	$\mathbf{83.9}_{4.0}$	$85.5_{2.6}$	$85.5_{1.2}$
OTJR	$\mathbf{74.0}_{4.3}$	$\mathbf{71.3}_{3.1}$	$\mathbf{46.0}_{1.9}$	$84.9_{3.8}$	$\mathbf{84.6}_{2.5}$	$\mathbf{84.1}_{1.4}$

Table 3: Online fool rate (%) [25] of various defense models with our downloaded CIFAR-10-WEB and CIFAR-100-WEB datasets.

use the following hyper-parameter settings: $\{K = 32, \lambda_J = 0.002, \lambda_{SW} = 64\}$ and $\{K = 128, \lambda_J = 0.001, \lambda_{SW} = 64\}$ for CIFAR-10 and CIFAR-100, respectively. Overall, we compare our method with four different recent SOTA AT frameworks: TRADES [37], ALP [20], PGD-AT [23], and SAT [4].

4.2 Popular Adversarial Attacks

Follow [37], we assess defense methods against a wide range *white-box* attacks (20 iterations)[3]: FGSM [18], PGD [23], MIM [14], CW_∞ [6], DeepFool [26], and FAB [9]; and *black-box* attacks (1000 iterations): Square [1] and Simba [15]. We use the same parameters as [30] for our experiments: for the L_∞ threat model, the values of epsilon and step size are $8/255$ and $2/255$ for CIFAR-10 and CIFAR-100, respectively. For the L_2 threat model, the values of epsilon and step size are $128/255$ and $15/255$ for all datasets. Additionally, we include AutoAttack [10] which is a reliable adversarial evaluation framework and an ensemble of four distinct attacks: APGD-CE, APGD-DLR [10], FAB [9], and Square [1]. The results of this experiment are presented in Table 2, where the best results are highlighted in bold. Evidently, our proposed OTJR method demonstrates its superior performance across different attack paradigms. The improvement is considerable on AutoAttack, by more than 0.51% and 1.21% on average compared to other methods on CIFAR-10 and CIFAR-100, respectively. In addition, we include two primary components of our proposed OTJR, *i.e.,* SW and random JR in Table 2. While the random JR, as expected, is highly vulnerable to most of the *white-box* attacks due to its adversarial-agnostic defense strategy, the SW approach is on par with other AT methods.

4.3 Online Adversarial Attacks

In order to validate the applicability of our defense model in real-world systems, we employ the *stochastic virtual* method [25]. This method is designed as an online attack algorithm with a transiency property; that is, an attacker makes an **irrevocable decision** regarding whether to initiate an attack or not.

For our experiment, we curate a dataset by downloading 1,000 images and categorizing them into 10 distinct classes, mirroring the structure of the CIFAR-10 dataset. We refer to this new subset as CIFAR-10-WEB. In a similar vein, we

[3] https://github.com/Harry24k/adversarial-attacks-pytorch.git

Defense	AWP*		LBGAT*		UDR†	
	Clean	AutoAtt	*Clean*	AutoAtt	*Clean*	AutoAtt
TRADES	*60.17*	28.80	*60.43*	29.34	68.04	47.87
OTJR	***60.55***	**29.79**	***62.15***	**29.64**	**68.31**	**49.34**

Table 4: Classification accuracy (%) from defense losses integrated AWP, LBGAT, and UDR, respectively. (∗: validating pre-trained model, †: re-runing model.).

Defense	*Clean*	MIM	CW	FAB	Square	AutoAtt	Avg.
TRADES	$35.91_{.10}$	$11.82_{.08}$	$8.92_{.13}$	$10.88_{.24}$	$15.86_{.04}$	$8.28_{.09}$	$11.15_{3.00}$
ALP	$39.69_{.37}$	$8.13_{.06}$	$7.66_{.08}$	$9.92_{.16}$	$15.98_{.34}$	$6.48_{.10}$	$9.63_{3.75}$
PGD	$33.81_{.22}$	$11.49_{.33}$	$\mathbf{10.14}_{.24}$	$11.29_{.32}$	$16.20_{.43}$	$\mathbf{8.98}_{.16}$	$11.60_{2.77}$
SAT[400]	*Not converge*						-
Random JR	$47.65_{.19}$	$0.29_{.02}$	$0.44_{.04}$	$3.79_{.13}$	$5.87_{.03}$	$0.19_{.01}$	$2.12_{2.59}$
SW	$37.05_{.09}$	$12.43_{.21}$	$9.35_{.19}$	$11.19_{.14}$	$16.77_{.31}$	$8.72_{.14}$	$11.19_{3.20}$
OTJR	$37.97_{.11}$	$\mathbf{12.57}_{.08}$	$9.55_{.04}$	$\mathbf{11.30}_{.09}$	$\mathbf{17.33}_{.12}$	$8.91_{.08}$	$\mathbf{11.97}_{3.34}$

Table 5: Classification accuracy (%) of PreActResNet18 with Tiny-IMAGENET dataset, under *white-* and *black-box* attacks.

assemble a CIFAR-100-WEB dataset comprising 5,000 images on the Internet, distributed across 100 classes analogous to the CIFAR-100 dataset. To ensure the reliability of our findings, we replicate the experiment ten times. The outcomes of these trials are detailed in Table 3. Notably, our novel system, OTJR, persistently showcases the most minimal fooling rate in comparison to other methods in the *k-secretary* settings [25], underscoring its robustness for real-world applications.

4.4 Compatibility with Different Frameworks and Datasets

Frameworks with surrogate models and relaxed perturbations. AWP [35] and LBGAT [11] are two SOTA benchmarks boosting adversarial robustness using surrogate models during training, albeit rather computationally expensive. Furthermore, they utilized existing AT losses, such as TRADES. We integrate our proposed optimization loss into AWP and LBGAT, respectively. Additionally, we also include the results of UDR [5] that used for creating relaxed perturbed noises upon its entire distribution. We selected TRADES as the best existing loss function that was deployed with these frameworks and experimented with CIFAR-100-WRN34 for comparison. For consistent comparison, we use TRADES loss as our $\mathcal{L}_{\text{AT}}$ in Eq. 12. As shown in Table 4, our OTJR is compatible with the three frameworks and surpasses the baseline performance by a significant margin.

Large scale dataset. We conducted further experiments on Tiny-IMAGENET with PreactResnet18, comparing our method against competitive AT methods under most challenging attacks, including MIM, CW, FAB, Square, and AutoAttack, in Table 5. All defensive methods are trained in 150 epochs with same settings as in Sec. 4.1. We note that SAT unable to converge within 400 epochs. As observed, our method not only sustains competitive performance on clean sample but also displays superior robustness against adversarial perturbations, while the second best robust model, PGD, has to sacrifice substantially its clean accuracy compared to ours (33.81% *vs.* 37.97%). In addition, this experiment illustrates that our proposed approach does not hinder the convergence process, unlike certain alternative optimum transport methods such as SAT, when applied to datasets of significant size.

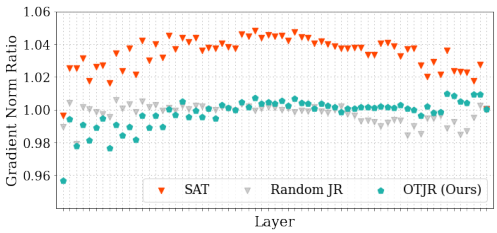

Fig. 6: Average gradient norm ratios between adversarial and clean samples. The ratios of $\mathbb{E}(||\nabla_{\theta_i}\mathcal{L}(\tilde{x})||/||\nabla_{\theta_i}\mathcal{L}(x)||)$ with respect to the model's parameters θ_i on CIFAR-10 dataset.

Method	$\mathbb{E}(\|\|\nabla_x\mathcal{L}\|\|_1)$					
	Clean	PGD^1	PGD^5	PGD^{10}	PGD^{15}	PGD^{20}
$\mathcal{XE}$	$854{\cdot}e^{-4}$	$5492{\cdot}e^{-4}$	$4894{\cdot}e^{-4}$	$4852{\cdot}e^{-4}$	$4975{\cdot}e^{-4}$	$4900{\cdot}e^{-4}$
Random JR	$105{\cdot}e^{-4}$	$149{\cdot}e^{-4}$	$272{\cdot}e^{-4}$	$335{\cdot}e^{-4}$	$360{\cdot}e^{-4}$	$370{\cdot}e^{-4}$
PGD	$158{\cdot}e^{-4}$	$232{\cdot}e^{-4}$	$420{\cdot}e^{-4}$	$520{\cdot}e^{-4}$	$568{\cdot}e^{-4}$	$586{\cdot}e^{-4}$
ALP	$118{\cdot}e^{-4}$	$161{\cdot}e^{-4}$	$249{\cdot}e^{-4}$	$298{\cdot}e^{-4}$	$322{\cdot}e^{-4}$	$332{\cdot}e^{-4}$
TRADES	$45{\cdot}e^{-4}$	$53{\cdot}e^{-4}$	$73{\cdot}e^{-4}$	$84{\cdot}e^{-4}$	$89{\cdot}e^{-4}$	$90{\cdot}e^{-4}$
SAT	$48{\cdot}e^{-4}$	$54{\cdot}e^{-4}$	$65{\cdot}e^{-4}$	$75{\cdot}e^{-4}$	$81{\cdot}e^{-4}$	$84{\cdot}e^{-4}$
OTJR (*Ours*)	$\mathbf{43}{\cdot}e^{-4}$	$\mathbf{48}{\cdot}e^{-4}$	$\mathbf{61}{\cdot}e^{-4}$	$\mathbf{70}{\cdot}e^{-4}$	$\mathbf{73}{\cdot}e^{-4}$	$\mathbf{75}{\cdot}e^{-4}$

Table 6: Average derivative of $\mathcal{XE}$ loss *w.r.t* input x at the pixel level after various PGD iterations. The lower the derivative is, the less perturbed the adversarial samples are when the model is abused. Best results are in **bold**.

Table 7: Comparison between ours and SAT+*Random JR* with WRN34.

Dataset	Defense	*Clean*	PGD^{20}	PGD^{100}	L_2-PGD	MIM	FGSM	CW	FAB	Square	SimBa	*AutoAtt*
CIFAR-10	*SAT+JR*	$83.75_{.48}$	$54.15_{.17}$	$53.87_{.07}$	$62.37_{.33}$	$54.12_{.16}$	$60.25_{.22}$	$52.22_{.31}$	$52.63_{.53}$	$61.72_{.39}$	$71.03_{.65}$	$51.15_{.37}$
	OTJR (***ours***)	$84.01_{.53}$	$\mathbf{55.38}_{.29}$	$\mathbf{55.08}_{.36}$	$\mathbf{63.87}_{.09}$	$\mathbf{55.31}_{.29}$	$\mathbf{61.03}_{.18}$	$\mathbf{54.09}_{.12}$	$54.17_{.07}$	$\mathbf{63.11}_{.21}$	$\mathbf{72.04}_{.68}$	$\mathbf{52.57}_{.12}$
CIFAR-100	$SAT+JR^{400}$	$52.13_{1.55}$	$26.18_{.46}$	$25.99_{.40}$	$32.64_{.90}$	$26.11_{.43}$	$30.14_{.84}$	$25.10_{.47}$	$25.19_{.85}$	$30.39_{1.02}$	$39.71_{1.14}$	$23.79_{.42}$
	OTJR (***ours***)	$58.20_{.13}$	$\mathbf{32.11}_{.21}$	$\mathbf{32.01}_{.18}$	$\mathbf{43.13}_{.12}$	$\mathbf{32.07}_{.19}$	$\mathbf{34.26}_{.30}$	$\mathbf{29.71}_{.06}$	$\mathbf{29.24}_{.08}$	$\mathbf{36.27}_{.05}$	$\mathbf{49.92}_{.23}$	$\mathbf{28.36}_{.10}$

Table 8: Accuracy (%) of WRN34 model trained with JR having random projections, and informative projections respectively.

Dataset	Defense	*Clean*	PGD^{20}	PGD^{100}	L_2-PGD	MIM	FGSM	CW	FAB	Square	SimBa	*AutoAtt*
CIFAR-10	*Random JR*	$84.99_{.14}$	$22.67_{.15}$	$21.89_{.14}$	$60.98_{.19}$	$22.49_{.18}$	$32.99_{.21}$	$22.00_{.06}$	$21.74_{.11}$	$45.86_{.20}$	$71.20_{.31}$	$20.54_{.14}$
	Optimal JR	$84.47_{.15}$	$\mathbf{24.81}_{.56}$	$\mathbf{23.97}_{.48}$	$\mathbf{61.67}_{.50}$	$\mathbf{24.62}_{.49}$	$\mathbf{34.22}_{.50}$	$\mathbf{23.91}_{.70}$	$\mathbf{24.74}_{.49}$	$\mathbf{47.18}_{.54}$	$\mathbf{73.59}_{.15}$	$\mathbf{22.84}_{.64}$
CIFAR-100	*Random JR*	$66.58_{.17}$	$9.41_{.44}$	$8.87_{.42}$	$37.79_{.31}$	$9.27_{.49}$	$16.38_{.33}$	$10.27_{.45}$	$9.26_{.23}$	$23.53_{.15}$	$48.86_{.55}$	$8.10_{.57}$
	Optimal JR	$65.16_{.27}$	$\mathbf{10.90}_{.29}$	$\mathbf{10.32}_{.35}$	$\mathbf{38.73}_{.20}$	$\mathbf{10.78}_{.32}$	$\mathbf{17.30}_{.29}$	$\mathbf{11.29}_{.43}$	$\mathbf{10.43}_{.34}$	$\mathbf{24.23}_{.41}$	$\mathbf{50.31}_{.13}$	$\mathbf{9.14}_{.34}$

4.5 Ablation Studies and Discussions

Loss's derivative. In continuation with our preliminary analysis, we highlight the disparities in defense model gradients across layers between our OTJR and SAT. Throughout intermediate layers in an attacked model, both frameworks provide stable ratios between perturbed and clean sample's gradients as shown in Fig. 6. It is worth noting that the gradients are derived on unobserved samples in the test set. In the forward path, our OTJR adeptly equilibrates gradients of adversarial and clean samples, with the majority of layers presenting ratio values approximating 1. Moreover, in the backward path, since the victim model's gradients are deployed to generate more perturbations, our OTJR model achieves better robustness when it can produce smaller gradients *w.r.t.* its inputs.

In addition, in Table 6, we present the average magnitude of cross-entropy loss's derivative *w.r.t.* the input images from CIFAR-10 dataset with different PGD white-box attack iterations on WRN34. Notably, as the number of attack iterations increases, the perturbation noise induced by the output loss derivatives intensifies. However, our proposed framework consistently exhibits the low-

est magnitude across all iterations. This characteristic underscores the superior robustness performance of OTJR.

SW distance vs. Sinkhorn divergence. Bouniot *et al.* propose the SAT algorithm that deploys Sinkhorn divergence to push the distributions of clean and perturbed samples towards each other [4]. While SAT can achieve comparable results to previous research, its limitations become pronounced in high-dimensional space, i.e., datasets with a large number of classes exhibit slower convergence, as demonstrated in other research [24, 28]. Our empirical results indicate that **SAT struggles to converge within 100 epochs** for the CIFAR-100 dataset. Meanwhile, we observed that hyper-parameter settings highly affect the adversarial training [27], and the performance improvement by SAT can be partly achieved with additional training epochs. Nevertheless, we still include results from SAT trained in 400 epochs on CIFAR-100 in [4].

Our OTJR vs. Naive Combination of SAT & random JR. To discern the impact of Jacobian regularization and distinguish our method from the naive combination of SAT and JR, we report their robustness under wide range of *white-box* PGD attack in Table 7. The experiments are conducted with CIFAR-10 and CIFAR-100 with WRN34. Our optimal approach attains slightly better robustness on small dataset (CIFAR-10). On the large dataset (CIFAR-100), however, our OTJR achieves significant improvement. This phenomenon is explained by the fact that regularizing the input-output Jacobian matrix increases the difficulty of the SAT algorithm's convergence, which results in a slower convergence. Therefore, naively combining AT and random Jacobian regularization can restrain the overall optimization process.

Optimal vs. random projections. To verify the efficiency of the optimal Jacobian regularization, WRN34 is trained on CIFAR-10 using $\mathcal{XE}$ loss with clean samples and the regularization term using Eq. 4 and Eq. 11 respectively, as follow: $\mathcal{L} = \sum_{i=1}^{\mathcal{B}} \left(\mathcal{L}_{\mathcal{XE}}(x_i, y_i) + \lambda_J ||J(x_i)||_F^2\right)$, where λ_J is a hyper-parameter to balance the regularization term and $\mathcal{XE}$ loss that is set to 0.02 for this experiment. As we can observe from Table 8, the Jacobian regularized model trained with SW-supported projections consistently achieves higher robustness, compared to the random projections. As shown, our proposed optimal regularization consistently achieves up to 2.5% improvement in accuracy under *AuttoAttack* compared to the random one.

In addition, we highlight the advantages of optimal Jacobian regularization on decision boundaries in Fig.7. Models trained without this regularizer are notably susceptible to perturbations. However, integrating the Jacobian regularizer augments robustness by broadening the decision boundaries, evidenced by an **enlarged black circle**. Our optimal Jacobian regularizer further extends the decision boundaries, amplifying model resilience. The rationale behind this enhancement lies in the informative directions showcased in Fig. 5, guiding the model to achieve optimal projections within the input-output Jacobian regularization framework.

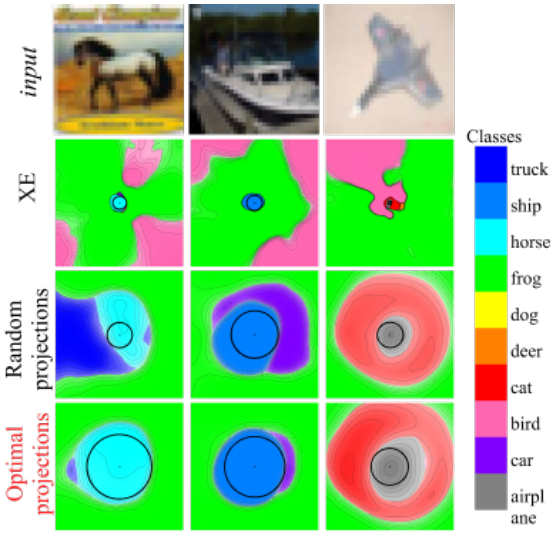

Fig. 7: Cross sections of decision boundaries in the input space. 1st row: Sample input images. 2nd row: Model trained without regularization. 3rd row: Model train with Jacobian regularization having random projection. 4th row: Model trained with Jacobian regularization having informative projections. Our optimal JR benefits from the informative directions obtained by SW distance, and thus helps the model to regularize these sensitive direction of clean samples and produce larger decision cells.

5 Conclusion

In this study, we explore the interplay between Adversarial Training (AT) and Jacobian regularization, especially in bolstering the robustness of DNNs against adversarial forays. We show that the AT pays more attention to meaningful input samples' pixels, whereas the Jacobian regularizer agnostically silences the DNN's gradients under any perturbation from its output to input layers. Based on these characterizations, we effectively augment the AT framework by integrating input-output Jacobian matrix in order to more effectively improve the DNN's robustness. Using the optimal transport theory, our work is the first to jointly minimize the difference between the distributions of original and adversarial samples with much faster convergence. Also, the proposed SW distance produces the optimal projections for the Jacobian regularization, which can further increase the decision boundaries of a sample under perturbations, and achieves much higher performance through optimizing the best of both worlds. Our rigorous empirical evaluations, pitted against four state-of-the-art defense mechanisms across both controlled and real-world datasets, underscore the supremacy of our method.

Acknowledgements. This work was partly supported by Institute for Information & communication Technology Planning & evaluation (IITP) grants funded by the Korean government MSIT: (RS-2022-II221199, RS-2024-00337703, RS-2022-II220688, RS-2019-II190421, RS-2023-00230337, RS-2024-00356293, RS-2022-II221045, RS-2021-II212068, No.RS-2023-00231200, RS-2024-00437849). Lastly, this work was supported by Korea Internet & Security Agency (KISA) grant funded by the Korea government (PIPC) (No.RS-2023-00231200, Development of personal video information privacy protection technology capable of AI learning in an autonomous driving environment).

References

1. Andriushchenko, M., Croce, F., Flammarion, N., Hein, M.: Square attack: a query-efficient black-box adversarial attack via random search. In: European Conference on Computer Vision (ECCV). pp. 484–501. Springer (2020) 10
2. Bai, Y., Zeng, Y., Jiang, Y., Xia, S.T., Ma, X., Wang, Y.: Improving adversarial robustness via channel-wise activation suppressing. International Conference on Learning Representations (ICLR) (2021) 5, 6
3. Bortsova, G., Gonzalez-Gonzalo, C., Wetstein, S., Dubost, F., Katramados, I., Hogeweg, L., Liefers, B., Ginneken, B., Pluim, J., Veta, M., et al.: Adversarial attack vulnerability of medical image analysis systems: Unexplored factors. Medical Image Analysis p. 102141 (2021) 2
4. Bouniot, Q., Audigier, R., Loesch, A.: Optimal transport as a defense against adversarial attacks. In: 2020 25th International Conference on Pattern Recognition (ICPR). pp. 5044–5051. IEEE (2021) 3, 10, 13
5. Bui, T.A., Le, T., Tran, Q., Zhao, H., Phung, D.: A unified wasserstein distributional robustness framework for adversarial training. arXiv preprint arXiv:2202.13437 (2022) 11
6. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. IEEE (2017) 10
7. Christian, S., Wojciech, Z., Ilya, S., Joan, B., Dumitru, E., Ian, G., Rob, F.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013) 1
8. Co, K.T., Rego, D.M., Lupu, E.C.: Jacobian regularization for mitigating universal adversarial perturbations. 30th International Conference on Artificial Neural Networks (ICANN) (2021) 4
9. Croce, F., Hein, M.: Minimally distorted adversarial examples with a fast adaptive boundary attack. In: International Conference on Machine Learning (ICML). pp. 2196–2205. PMLR (2020) 10
10. Croce, F., Hein, M.: Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In: International conference on machine learning. pp. 2206–2216. PMLR (2020) 10
11. Cui, J., Liu, S., Wang, L., Jia, J.: Learnable boundary guided adversarial training. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 15721–15730 (2021) 7, 11
12. Cuturi, M.: Sinkhorn distances: Lightspeed computation of optimal transport. Advances in neural information processing systems (NeurIPS) **26**, 2292–2300 (2013) 3
13. Deng, Y., Zheng, X., Zhang, T., Chen, C., Lou, G., Kim, M.: An analysis of adversarial attacks and defenses on autonomous driving models. In: 2020 IEEE International Conference on Pervasive Computing and Communications (PerCom). pp. 1–10. IEEE (2020) 2
14. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR). pp. 9185–9193 (2018) 10
15. Guo, C., Gardner, J., You, Y., Wilson, A.G., Weinberger, K.: Simple black-box adversarial attacks. In: International Conference on Machine Learning (ICML). pp. 2484–2493. PMLR (2019) 10
16. Helgason, S.: Integral geometry and Radon transforms. Springer Science & Business Media (2010) 8

17. Hoffman, J., Roberts, D.A., Yaida, S.: Robust learning with jacobian regularization. arXiv preprint arXiv:1908.02729 (2019) 2, 4, 5, 6, 8
18. Ian, G., Jonathon, S., Christian, S.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014) 1, 2, 10
19. Jakubovitz, D., Giryes, R.: Improving dnn robustness to adversarial attacks using jacobian regularization. In: Proceedings of the European Conference on Computer Vision (ECCV). pp. 514–529 (2018) 2, 4
20. Kannan, H., Kurakin, A., Goodfellow, I.: Adversarial logit pairing. arXiv preprint arXiv:1803.06373 (2018) 3, 4, 10
21. Krizhevsky, A., Hinton, G., et al.: Learning multiple layers of features from tiny images. Tech. rep., University of Toronto, Toronto (2009) 9
22. Ma, X., Niu, Y., Gu, L., Wang, Y., Zhao, Y., Bailey, J., Lu, F.: Understanding adversarial attacks on deep learning based medical image analysis systems. Pattern Recognition **110**, 107332 (2021) 2
23. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. International Conference on Learning Representations (ICLR) (2018) 2, 3, 4, 5, 9, 10
24. Meng, C., Ke, Y., Zhang, J., Zhang, M., Zhong, W., Ma, P.: Large-scale optimal transport map estimation using projection pursuit. Advances in Neural Information Processing Systems (NeurIPS) (2019) 3, 13
25. Mladenovic, A., Bose, J., Hamilton, W.L., Lacoste-Julien, S., Vincent, P., Gidel, G., et al.: Online adversarial attacks. In: International Conference on Learning Representations (2021) 10, 11
26. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016) 10
27. Pang, T., Yang, X., Dong, Y., Su, H., Zhu, J.: Bag of tricks for adversarial training. arXiv preprint arXiv:2010.00467 (2020) 3, 9, 13
28. Petrovich, M., Liang, C., Sato, R., Liu, Y., Tsai, Y.H.H., Zhu, L., Yang, Y., Salakhutdinov, R., Yamada, M.: Feature robust optimal transport for high-dimensional data. arXiv preprint arXiv:2005.12123 (2020) 3, 13
29. Qian, N.: On the momentum term in gradient descent learning algorithms. Neural networks **12**(1), 145–151 (1999) 9
30. Rice, L., Wong, E., Kolter, Z.: Overfitting in adversarially robust deep learning. In: International Conference on Machine Learning. pp. 8093–8104. PMLR (2020) 10
31. Rony, J., Hafemann, L.G., Oliveira, L.S., Ayed, I.B., Sabourin, R., Granger, E.: Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4322–4330 (2019) 3
32. Shafahi, A., Najibi, M., Ghiasi, A., Xu, Z., Dickerson, J., Studer, C., Davis, L.S., Taylor, G., Goldstein, T.: Adversarial training for free! arXiv preprint arXiv:1904.12843 (2019) 4
33. Vialard, F.X.: An elementary introduction to entropic regularization and proximal methods for numerical optimal transport (2019) 3
34. Villani, C.: Optimal transport: old and new, vol. 338. Springer Science & Business Media (2008) 7
35. Wu, D., Xia, S.T., Wang, Y.: Adversarial weight perturbation helps robust generalization. Advances in Neural Information Processing Systems **33**, 2958–2969 (2020) 7, 11

36. Zagoruyko, S., Komodakis, N.: Wide residual networks. arXiv preprint arXiv:1605.07146 (2016) 9
37. Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., Jordan, M.: Theoretically principled trade-off between robustness and accuracy. In: International Conference on Machine Learning (ICML). pp. 7472–7482. PMLR (2019) 3, 4, 10
38. Zhang, J., Zhu, J., Niu, G., Han, B., Sugiyama, M., Kankanhalli, M.: Geometry-aware instance-reweighted adversarial training. International Conference on Learning Representations (ICLR) (2020) 3