PILO: Principal Component-based Implicit Regularization with Low-rank Optimization for Robust Transfer Learning

Shuaihe Liu , **Qiugang Zhan** , **Guisong Liu** and **Tai-Xiang Jiang***

The Complex Laboratory of New Finance and Economics, School of Computing and Artificial Intelligence, Southwestern University of Finance and Economics, Chengdu, P.R.China
Kash Institute of Electronics and Information Industry, Kash, P.R.China

Abstract

Adapting large, adversarially pre-trained models to specialized domains via transfer learning is a promising path toward building secure AI systems. However, a critical challenge arises when fine-tuning on limited downstream data: models often suffer from *catastrophic forgetting of robustness*, where the generalizable robust features from pre-training are erased, leading to severe robust overfitting. To address this, we first establish the underlying principle that adversarial vulnerability is not diffuse, but is concentrated in a low-dimensional subspace defined by the *principal components* of the model's weight matrices. We then introduce **P**rincipal component-based **I**mplicit regularization with **L**ow-rank **O**ptimization (**PILO**), a novel parameter-efficient framework that operationalizes this insight. PILO employs a principled decoupling of learning objectives: a *spectrally-guided* adversarial branch, initialized via Singular Value Decomposition (SVD), performs a surgical update on the identified vulnerable subspace to enhance robustness. This is complemented by a transient, standard low-rank branch that gently adapts the model to natural data, preserving clean accuracy. This dual-branch design acts as a form of implicit spectral regularization, anchoring the model to its robust pre-trained state. Extensive experiments show that PILO significantly outperforms state-of-the-art full-parameter and parameter-efficient methods in robust accuracy across multiple benchmarks, all while reducing trainable parameters by up to 71%. Our work thus establishes a new, more effective paradigm for robust transfer learning through principled and targeted parameter optimization.

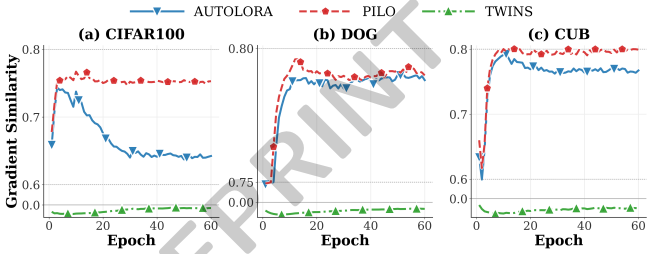


Figure 1: PILO's targeted update strategy substantially mitigates gradient conflict. The plots show the average cosine similarity between the gradients of natural and adversarial objectives across three datasets. While full-parameter methods like TWINS and AUTOLORA exhibit low or conflicting gradient alignment (measured on the full feature extractor), PILO maintains a consistently high gradient similarity. This is measured with respect to its trainable principal component matrices (A_{adv} and B_{adv}), validating that our approach leads to a more stable and cooperative optimization path.

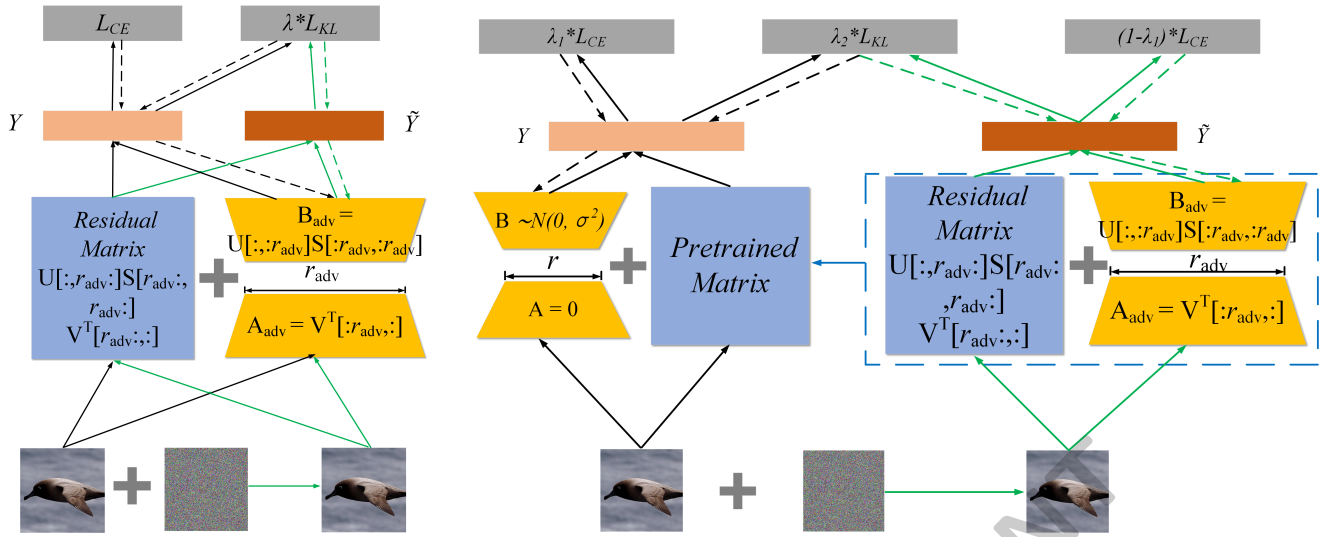
1 Introduction

The transfer learning paradigm [Yosinski *et al.*, 2014; Howard and Ruder, 2018; Devlin *et al.*, 2019; Raffel *et al.*, 2020]—comprising pre-training on large-scale datasets and subsequent task-specific fine-tuning—has become a cornerstone of modern machine learning.

With the rise of foundation models [Bommasani *et al.*, 2021], fine-tuning pre-trained feature extractors enables state-of-the-art performance across a multitude of downstream tasks. Empirical evidence from computer vision confirms that initializing weights from pre-trained models consistently surpasses random initialization [Kornblith *et al.*, 2019; He *et al.*, 2019]. However, this success is critically undermined by the models' vulnerability to adversarial perturbations, a flaw that poses significant risks in high-stakes applications like autonomous driving [Eykholt *et al.*, 2018; Kurakin *et al.*, 2018] and medical image analysis [Buch *et al.*, 2018]. This necessitates the development of robust fine-tuning methodologies to ensure model resilience in security-sensitive settings.

To this end, various robust fine-tuning techniques have emerged. Adversarial Training (AT), which incorporates PGD-generated examples [Madry *et al.*, 2017], is a foundational method but often impairs accuracy on clean data. Subsequent approaches like TRADES [Zhang *et al.*, 2019] achieved a better balance, while TWINS [Liu *et al.*, 2023] further enhanced performance using dual batch normalization architectures [Xie *et al.*, 2020]. Most recently, AUTOLORA [Xu *et al.*, 2023] introduced a dual-branch structure to miti-

*Corresponding author. taixiangjiang@gmail.com



(a) TRADES (Principal Component) (b) PILO

Figure 2: Architectural comparison of (a) a baseline TRADES fine-tuning on principal components and (b) our proposed PILO framework. The baseline uses a single, symmetric path for both natural and adversarial data, updating a single set of principal component matrices (A_{adv}, B_{adv}). In contrast, PILO employs a structurally asymmetric, dual-branch architecture that decouples the learning objectives. The natural branch uses a standard low-rank adapter to preserve clean accuracy, while the adversarial branch performs a targeted, surgical update on the SVD-initialized principal components to build robustness. This synergistic integration of structural separation with principal component optimization serves as a crucial mechanism for mitigating objective conflicts. (Yellow modules: trainable; blue: frozen. Solid/dashed arrows: forward/backward propagation for clean/adversarial data).

gate gradient conflicts. Despite these advances, these methods are all built on a *full-parameter fine-tuning* strategy. When applied to downstream tasks with limited data, this approach often leads to a critical failure mode we term **catastrophic forgetting of robustness**, where the generalizable robust features learned during pre-training are overwritten, resulting in severe robust overfitting [Kumar *et al.*, 2022; Rice *et al.*, 2020]

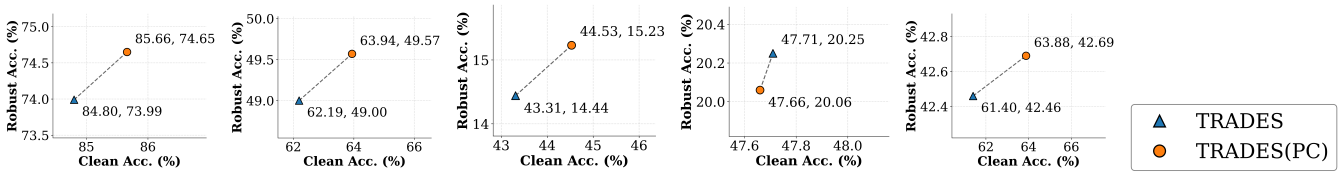
In this work, we argue that this failure stems from a flawed monolithic assumption. We posit a new principle: **adversarial vulnerability is not diffuse but is disproportionately concentrated in the principal components of pre-trained weight matrices.** To validate this, we conducted a granular spectral ablation study with identical rank budgets. As shown in **Table 1**, fine-tuning strictly the top-r components (Top-r) consistently outperforms the middle and bottom counterparts and even surpasses full-parameter fine-tuning. This suggests that lower spectral components likely harbor task-irrelevant noise that exacerbates robust overfitting. These high-magnitude spectral components represent the most sensitive directions of the feature space, which we hypothesize are the primary targets of adversarial attacks. This principle offers a direct solution to the gradient instability plaguing prior methods; as shown in **Figure 1**, isolating optimization to the principal components yields a remarkably high and stable gradient similarity between the competing objectives, in stark contrast to the conflicts observed in full-parameter approaches.

Building on this insight, we introduce Principal component-based Implicit regularization with Low-rank Optimization (**PILO**), a novel framework that operationalizes our principle through a **structurally asymmetric** dual-branch architecture (as shown in Figure 2). PILO decouples the learning objectives:

1. An **adversarial branch** performs a targeted update exclusively on the principal components, initialized via SVD [Meng *et al.*, 2024]. This acts as a powerful form of implicit regularization, anchoring the model to its robust pre-trained state and preserving its core robustness, an idea motivated by prior work on robust transfer [Shafahi *et al.*, 2019].
2. A complementary **natural branch** employs a standard low-rank adapter to gently fine-tune for clean accuracy, which is discarded during inference to maximize efficiency.

As illustrated in Figure 3, fine-tuning only the principal components of a ResNet18 model with TRADES already surpasses a fully fine-tuned baseline across five datasets, confirming the fundamental benefit of our targeted strategy. Our key contributions are as follows:

- We address the problem of robust overfitting by introducing a parameter-efficient fine-tuning strategy focused solely on the principal components of pre-trained weights. This approach imposes strong implicit spectral regularization, improving generalization on limited downstream datasets.
- We advance the resolution of gradient conflicts through a novel dual-branch architecture that strategically separates optimization paths: a spectrally-guided branch for adversarial objectives and a complementary low-rank



(a) CIFAR-10 (b) CIFAR-100 (c) DOG-120 (d) CUB-200 (e) CALTECH-256

Figure 3: Comparison of robust and clean accuracy between full parameter fine-tuning and principal component fine-tuning using TRADES on five datasets

module for natural objectives.

- Extensive experiments demonstrate that PILO significantly improves model robustness on downstream datasets compared to state-of-the-art methods. In contrast to prior full-parameter approaches, our method achieves this with only about 29–40 % of the trainable parameters.

Datasets	Evaluation	Top	Middle	Bottom	Full
CUB	PGD-10	**21.59%**	17.97%	16.90%	21.54%
DOG	PGD-10	**17.04%**	16.33%	16.06%	16.35%

Table 1: Spectral Ablation Analysis with identical rank r. We report the PGD-10 accuracy (%) on CUB and DOG datasets. Top spectral components consistently outperform Middle and Bottom components, and even surpass Full fine-tuning, indicating that robustness is concentrated in the principal components.

2 Related Work

2.1 Parameter-Efficient Fine-tuning

Parameter-efficient fine-tuning (PEFT) adapts large pretrained models to downstream tasks by updating only a small subset of parameters, significantly reducing storage and computational overhead. Representative strategies include adapter-based methods [Houlsby *et al.*, 2019; Pfeiffer *et al.*, 2020; Rebuffi *et al.*, 2017], which insert lightweight trainable modules into transformer layers, and prompt-based techniques [Li and Liang, 2021; Lester *et al.*, 2021; Liu *et al.*, 2021] that prepend learnable vectors to input sequences. Among these, Low-Rank Adaptation (LoRA) [Hu *et al.*, 2022] has gained prominence by freezing the pretrained backbone and injecting rank-decomposition matrices, ensuring zero additional inference latency. However, LoRA's standard random initialization ignores the structural information of original weights, potentially hindering optimization. PISSA [Meng *et al.*, 2024] addresses this by initializing adapters via principal component decomposition. This spectrally-guided approach better preserves the pretrained capabilities at initialization, leading to faster convergence and superior performance.

2.2 Adversarial Attacks

The field of adversarial attacks originated with the Fast Gradient Sign Method (FGSM) [Goodfellow *et al.*, 2014]. Subsequently, Projected Gradient Descent (PGD) [Madry *et al.*, 2017] became the standard for multi-step optimization, while C&W [Carlini and Wagner, 2017] introduced high-quality optimization-based attacks. To ensure reliable evaluation, AutoAttack (AA) [Croce and Hein, 2020b] was developed as a parameter-free ensemble incorporating diverse strategies like FAB [Croce and Hein, 2020a] and Square Attack [Andriushchenko *et al.*, 2020]. In black-box scenarios where gradients are unavailable, methods rely on zeroth-order optimization like ZOO [Chen *et al.*, 2017] or leverage the transferability of adversarial examples [Papernot *et al.*, 2016].

2.3 Robust Fine-Tuning for Downstream Tasks

Following foundational work on the transferability of robust features [Salman *et al.*, 2020; Shafahi *et al.*, 2019; Hendrycks *et al.*, 2019], research into robust fine-tuning has grown substantially. While it was shown that robustly pretrained models can improve clean accuracy on downstream tasks [Salman *et al.*, 2020], directly transferring robustness requires specialized fine-tuning. Conventional adversarial training is the most straightforward approach, but it often incurs high computational costs and degrades performance on natural examples. This trade-off catalyzed more sophisticated methodologies.

Subsequent research has formulated robust fine-tuning as a multi-objective optimization problem. TRADES [Zhang *et al.*, 2019] pioneered a regularization-based approach to explicitly balance standard and adversarial performance. TWINS [Liu *et al.*, 2023] improved upon this by addressing the distribution shift between natural and adversarial examples with a dual batch normalization (BN) architecture. More recently, AUTOLORA [Xu *et al.*, 2023] introduced a dual-branch architecture to decouple the competing objectives and stabilize optimization. Despite these architectural advances, these methods still rely on full-parameter updates or generic PEFT approaches, which can lead to the catastrophic forgetting of robustness on limited data.

In contrast to these methods, our work is founded on a more fundamental principle. Instead of only decoupling the optimization objectives architecturally, PILO performs a *structural* decoupling at the parameter level. By executing a surgical, spectrally-guided update on the principal components, i.e., the hypothesized locus of vulnerability, while using a distinct low-rank adapter for natural data, PILO introduces a principled, asymmetric fine-tuning strategy that previous methods lack.

3 Methodology

Building on the challenges identified in the introduction, we propose a parameter-efficient fine-tuning strategy that selectively updates principal components while employing a dual-branch optimization approach. The following sections detail each component of our strategy.

3.1 The Spectrally-Guided Adversarial Branch

For the adversarial objective, we design a branch that directly targets the hypothesized locus of vulnerability: the principal components of the pre-trained weights. This design is motivated by the principle that adversarial perturbations primarily exploit the most sensitive directions in the weight space, which correspond to the largest singular values. By initializing this branch using singular value decomposition (SVD), we perform a surgical update on these critical directions, providing a more effective and direct path to enhancing adversarial robustness.

The trainable matrices $\mathbf{B}_{\text{adv}} \in \mathbb{R}^{m \times r_{\text{adv}}}$ and $\mathbf{A}_{\text{adv}} \in \mathbb{R}^{r_{\text{adv}} \times n}$ are initialized directly from the SVD of the pre-trained weight matrix $\mathbf{W}_{\text{pre}} = \mathbf{U}\mathbf{S}\mathbf{V}^\top$:

$$\mathbf{B}_{\text{adv}} = \mathbf{U}[:, : r_{\text{adv}}]\, \mathbf{S}[: r_{\text{adv}}, : r_{\text{adv}}] \in \mathbb{R}^{m \times r_{\text{adv}}}, \tag{1}$$

$$\mathbf{A}_{\text{adv}} = \mathbf{V}^T[: r_{\text{adv}}, :] \in \mathbb{R}^{r_{\text{adv}} \times n}. \tag{2}$$

Here, r_{adv} is the rank of the branch, determining the number of principal components to fine-tune. The forward propagation for an adversarial example $\mathbf{x}^*$ is then formulated as:

$$\mathbf{Y} = (\mathbf{W}_{\text{res}} + \mathbf{B}_{\text{adv}}\mathbf{A}_{\text{adv}})\mathbf{x}^*, \tag{3}$$

where $\mathbf{W}_{\text{res}} = \mathbf{U}[:, r_{\text{adv}} :]\mathbf{S}[r_{\text{adv}} :, r_{\text{adv}} :]\mathbf{V}^\top[:, r_{\text{adv}} :]$ is the residual matrix containing the less significant, frozen components of the original weight. This decomposition constrains all parameter updates to the principal component subspace, acting as a form of **implicit spectral regularization**. This anchors the fine-tuned model to the robust representations of the pre-trained model, preventing catastrophic forgetting of robustness.

The adversarial input $\mathbf{x}^*$ is generated using a standard attack method:

$$\mathbf{x}^* = \mathbf{x} + \arg\max_{\|\boldsymbol{\delta}\|_p < \epsilon} \ell\left(F(\mathbf{x} + \boldsymbol{\delta}), p_d(y|\mathbf{x})\right), \tag{4}$$

where $F(\cdot)$ is the model, $\boldsymbol{\delta}$ is the adversarial perturbation constrained by the L_p-norm, and ϵ denotes the adversarial budget.

3.2 The Auxiliary Natural Branch for Adaptation

To maintain accuracy on clean data, we employ a second, lightweight branch for the natural objective. This branch uses a standard low-rank adaptation approach, initialized for gentle, gradual learning. The matrices are defined as:

$$\mathbf{B} \sim \mathcal{N}(0, \sigma^2) \in \mathbb{R}^{m \times r}, \tag{5}$$

$$\mathbf{A} = \mathbf{0} \in \mathbb{R}^{r \times n}, \tag{6}$$

where r is the rank of the natural branch. Initializing matrix $\mathbf{B}$ with small random values and matrix $\mathbf{A}$ to zero ensures that the model's performance on natural data starts exactly from the pre-trained model's state and is only subtly adjusted during fine-tuning. The forward propagation for a clean example $\mathbf{x}$ is:

$$\mathbf{Y} = (\mathbf{W}_{\text{pre}} + \mathbf{B}\mathbf{A})\mathbf{x}, \tag{7}$$

where the full pre-trained weight $\mathbf{W}_{\text{pre}}$ remains frozen. This design allows for effective adaptation to the downstream task while mitigating the risk of overfitting.

Crucially, this natural branch is **auxiliary** and is used exclusively during the training phase to provide a stable target for the adversarial branch. At inference time, matrices $\mathbf{A}$ and $\mathbf{B}$ are completely removed. This design choice not only enhances computational efficiency but also improves generalization by resulting in a more streamlined final model, a finding confirmed by our ablation studies in the appendix.

3.3 The Proposed PILO

The complete PILO framework, illustrated in Figure 2, integrates the two branches into a comprehensive multi-objective loss function:

$$\begin{aligned}\mathcal{L} = \sum_{(\mathbf{x},y)\in\mathcal{D}} \Big[& \lambda_1 \cdot \ell_{\text{CE}}(h_{\boldsymbol{\theta}_{\text{nat}}}(\mathbf{x}), y) \\ & + (1-\lambda_1) \cdot \ell_{\text{CE}}(h_{\boldsymbol{\theta}_{\text{adv}}}(\mathbf{x}^*), y) \\ & + \lambda_2 \cdot \ell_{\text{KL}}(h_{\boldsymbol{\theta}_{\text{adv}}}(\mathbf{x}^*), h_{\boldsymbol{\theta}_{\text{nat}}}(\mathbf{x})) \Big],\end{aligned}$$

where $\boldsymbol{\theta}_{\text{nat}} = \{\mathbf{W}_{\text{pre}} + \mathbf{B}\mathbf{A}, \mathbf{W}_{\text{fc}}\}$ and $\boldsymbol{\theta}_{\text{adv}} = \{\mathbf{W}_{\text{res}} + \mathbf{B}_{\text{adv}}\mathbf{A}_{\text{adv}}, \mathbf{W}_{\text{fc}}\}$, and $\mathbf{W}_{\text{fc}}$ indicates the classifier parameters in the last fully connected layer. The classifier parameters $\mathbf{W}_{\text{fc}}$ are updated during training, while the pre-trained weights within $\boldsymbol{\theta}_{\text{nat}}$ and $\boldsymbol{\theta}_{\text{adv}}$ remain frozen.

This loss function has three components: (1) a standard cross-entropy loss on clean samples $(\mathbf{x}, y)$ to preserve accuracy; (2) a cross-entropy loss on adversarial samples $(\mathbf{x}^*, y)$ to build robustness; and (3) a KL divergence term that regularizes the adversarial branch. This term encourages the model's output for an *adversarial input* $\mathbf{x}^*$ to match the natural branch's output for the corresponding *clean input* $\mathbf{x}$, promoting robust generalization. The coefficients λ_1 and λ_2 are dynamically scheduled based on validation accuracy

$$\begin{aligned}\lambda_1^{(e)} &= 1 - \text{Acc}(\boldsymbol{\theta}^{(e)}; D_{\text{val}}), \\ \lambda_2^{(e)} &= \gamma \cdot \text{Acc}(\boldsymbol{\theta}^{(e)}; D_{\text{val}}),\end{aligned} \tag{8}$$

where e is the epoch, $\boldsymbol{\theta}^{(e)}$ represents the combined model parameters, and γ is a scaling hyperparameter. $\text{Acc}(\cdot)$ denotes the clean accuracy, and D_{val} denotes the validation set.

A primary advantage of PILO is its exceptional parameter efficiency. In contrast to traditional methods that fine-tune all $m \times n$ parameters of a weight matrix, PILO concentrates updates into four small, low-rank matrices ($\mathbf{A}_{\text{adv}}, \mathbf{B}_{\text{adv}}, \mathbf{A}, \mathbf{B}$). This reduces the number of trainable parameters for each adapted layer to only $r \times (m+n) + r_{\text{adv}} \times (m+n)$, a substantial reduction. This efficient parameterization offers multiple benefits: it significantly mitigates the risk of overfitting on limited data, enables better generalization, and reduces computational overhead during training. Furthermore, the removal of the auxiliary natural branch at inference results in a more streamlined final model, making PILO particularly

Dataset	Evaluation Setting	Methods				Improvement (Δ vs.)		
		TRADES	TWINS	AUTOLORA	**PILO**	TRADES	TWINS	AUTOLORA
Caltech-256	Clean	61.40%	67.49%	63.74%	**65.77%**	+4.37%	-1.72%	+2.03%
	PGD-10	42.46%	42.92%	43.93%	**44.55%**	+2.09%	+1.63%	+0.62%
	AA	35.73%	37.14%	38.07%	**39.08%**	+3.35%	+1.94%	+1.01%
CUB-200	Clean	47.71%	51.11%	49.62%	**52.00%**	+4.29%	+0.89%	+2.38%
	PGD-10	20.25%	20.28%	21.54%	**21.59%**	+1.34%	+1.31%	+0.05%
	AA	14.77%	15.38%	15.10%	**15.81%**	+1.04%	+0.43%	+0.71%
DOG-120	Clean	43.31%	48.73%	48.02%	**49.02%**	+5.71%	+0.29%	+1.00%
	PGD-10	14.44%	15.12%	16.35%	**17.04%**	+2.60%	+1.92%	+0.69%
	AA	8.99%	9.35%	10.29%	**10.77%**	+1.78%	+1.42%	+0.48%
CIFAR-10	Clean	84.80%	86.26%	87.54%	**87.87%**	+3.07%	+1.61%	+0.33%
	PGD-10	73.99%	74.08%	75.71%	**75.84%**	+1.85%	+1.76%	+0.13%
	AA	73.03%	73.21%	74.59%	**74.80%**	+1.77%	+1.59%	+0.21%
CIFAR-100	Clean	62.19%	67.21%	66.43%	**66.35%**	+4.16%	-0.86%	-0.08%
	PGD-10	49.00%	49.91%	50.80%	**51.37%**	+2.37%	+1.46%	+0.57%
	AA	46.59%	47.27%	48.17%	**48.88%**	+2.29%	+1.61%	+0.71%

Table 2: Performance comparison of ResNet18 fine-tuned using PILO versus three baseline approaches across five downstream datasets. Ours results for each row are highlighted in bold.

well-suited for deployment on resource-constrained devices while achieving a superior balance between performance and robustness.

4 Experiments

In this section, we conduct a series of robustness benchmarks across multiple downstream tasks to evaluate the efficacy of our proposed PILO method against state-of-the-art approaches.

4.1 Experiment Settings

Downstream Datasets We evaluate PILO on five distinct datasets. **CIFAR-10** and **CIFAR-100** [Krizhevsky *et al.*, 2009] are low-resolution image datasets, with CIFAR-10 comprising 50,000 training and 10,000 test images across 10 classes, and CIFAR-100 containing the same number of images distributed among 100 classes. For high-resolution benchmarks, we use **Caltech-256** [Griffin *et al.*, 2007], ensuring our data partitioning is consistent with that of AUTOLORA for fair comparison. For fine-grained classification, we use **CUB-200** [Wah *et al.*, 2011] (5,994 training and 5,794 test images across 200 bird species) and **Stanford Dogs** (DOG-120)[Khosla *et al.*, 2011] (12,000 training and 8,580 test images of 120 dog breeds).

Pre-trained Models We use adversarially pre-trained ResNet architectures (**ResNet-18** and **ResNet-50**) [He *et al.*, 2016] as our primary backbones. These models were pre-trained on the ImageNet dataset [Deng *et al.*, 2009] with input images standardized to 224×224 pixels. Following the AUTOLORA configuration, the pre-trained models use a default adversarial budget of $\epsilon_{\text{pt}} = 4/255$.

Training Configurations To ensure a fair comparison, we follow the training procedure of AUTOLORA, training all models for 60 epochs using SGD with a weight decay of 1×10^{-4}. For PILO, we set the rank of the **auxiliary** branch to $r = 8$. For the adversarial branch in ResNet models, we set the rank r_{adv} of the last two layers to be 4 times that of the first two layers to account for differing layer shapes. For Transformer-based models (ViT, Swin), we maintain a uniform r_{adv} across all layers. We hold out 5% of the training data as a validation set. Adversarial examples for training are generated using PGD-10 with $\epsilon = 8/255$ and a step size of $2/255$, using a batch size of 128. We adopt AUTOLORA's automated hyperparameter scheduling mechanism for λ_1 and λ_2, with $\gamma = 6$ by default. Detailed learning rates and specific r_{adv} values are provided in the appendix.

Evaluation We evaluate final model robustness against two standard adversarial attacks: Projected Gradient Descent (**PGD-10**) and **AutoAttack (AA)** [Croce and Hein, 2020b]. All attacks are constrained by the ℓ_∞-norm with an adversarial budget of $\epsilon = 8/255$. The PGD settings used for evaluation are identical to those used during training.

Baselines Our primary baselines for comparison are **TRADES** [Zhang *et al.*, 2019], **TWINS** [Liu *et al.*, 2023], and **AUTOLORA** [Xu *et al.*, 2023]. To ensure a fair and direct comparison, we adopt the official hyperparameter configurations specified in their original papers, including learning rates, schedulers, weight decay, and any method-specific parameters (e.g., the β scalar in TWINS).

4.2 Experimental Results

As presented in Tables 2 and 3, our comprehensive benchmarks demonstrate that PILO consistently establishes a new state-of-the-art for robust fine-tuning. Across two different model architectures (ResNet18 and ResNet50) and five diverse downstream datasets, PILO significantly outperforms all three established baselines, i.e., TRADES, TWINS, and

Dataset	Evaluation Setting	Methods				Improvement (Δ vs.)		
		TRADES	TWINS	AUTOLORA	**PILO**	TRADES	TWINS	AUTOLORA
Caltech-256	Clean	66.78%	69.41%	69.63%	**72.42%**	+5.64%	+3.01%	+2.79%
	PGD-10	47.76%	47.79%	50.02%	**50.41%**	+2.65%	+2.62%	+0.39%
	AA	42.17%	42.27%	43.36%	**44.41%**	+2.24%	+2.14%	+1.05%
CUB-200	Clean	57.65%	64.36%	60.99%	**62.22%**	+4.57%	-2.14%	+1.23%
	PGD-10	28.27%	28.60%	30.57%	**30.57%**	+2.30%	+1.97%	+0.00%
	AA	21.97%	21.30%	23.56%	**24.23%**	+2.26%	+2.93%	+0.67%
DOG-120	Clean	62.68%	63.64%	61.40%	**61.61%**	-1.07%	-2.03%	+0.21%
	PGD-10	24.87%	24.98%	26.14%	**26.26%**	+1.39%	+1.28%	+0.12%
	AA	12.73%	14.41%	17.55%	**17.62%**	+4.89%	+3.21%	+0.07%
CIFAR-10	Clean	89.13%	90.97%	90.71%	**91.52%**	+2.39%	+0.55%	+0.81%
	PGD-10	78.10%	78.27%	79.16%	**79.77%**	+1.67%	+1.50%	+0.61%
	AA	76.58%	77.36%	77.51%	**77.93%**	+1.35%	+0.57%	+0.42%
CIFAR-100	Clean	68.48%	70.07%	71.11%	**72.19%**	+3.71%	+2.12%	+1.08%
	PGD-10	52.93%	53.25%	54.05%	**55.77%**	+2.84%	+2.52%	+1.72%
	AA	50.54%	50.86%	50.91%	**52.15%**	+1.61%	+1.29%	+1.24%

Table 3: Performance comparison of ResNet50 fine-tuned using PILO versus three baseline approaches across five downstream datasets. Ours results for each row are highlighted in bold.

AUTOLORA, in adversarial robustness metrics under the attack of PGD-10 and AutoAttack. Crucially, these gains in robustness are achieved while maintaining or often improving accuracy on clean data, showcasing the effectiveness of our decoupled optimization strategy.

The results on the ResNet18 architecture, shown in Table 2, provide strong evidence for our central hypothesis. On challenging, fine-grained datasets with limited samples like **CUB-200** and **DOG-120**, where full-parameter models are highly susceptible to overfitting, PILO shows marked improvements in both robust and clean accuracy over AUTOLORA. This demonstrates PILO's ability to prevent the *catastrophic forgetting of robustness*; by performing a surgical update only on the principal components, PILO preserves the essential robust features learned during pre-training, while the auxiliary branch carefully adapts to the new data distribution. The consistent gains across all datasets validate that our method's implicit spectral regularization is more effective at generalizing on limited data than the full-parameter approaches of the baselines.

As exhibited in Table 3, the performance advantages of PILO are even more pronounced on the deeper ResNet50 architecture. For instance, on **Caltech-256**, PILO improves clean accuracy by a significant +2.79% over the best baseline, while also improving robustness. On **CIFAR-100**, it achieves a remarkable +1.72% gain in PGD-10 robustness and +1.24% in AA robustness over AUTOLORA. This trend suggests that as model capacity increases, the risk of disturbing the delicately learned features of a robust pre-trained model via full fine-tuning also increases. PILO's principled decomposition becomes even more critical in this high-capacity regime. It effectively isolates and preserves the most critical feature directions (the principal components) from a much larger and more complex parameter space, leading to superior scalability and performance.

The collective results empirically validate the core design principles of PILO. The consistent performance improvements are not coincidental but are a direct result of our targeted, spectrally-guided optimization strategy. By identifying and exclusively fine-tuning the parameter subspace most relevant to adversarial vulnerability, PILO achieves a more efficient and effective optimization than prior methods. These findings firmly establish PILO as a novel and state-of-the-art method for robust transfer learning.

4.3 Analysis and Ablation Studies

Impact of Adversarial Rank (r_{adv}) Table 4 presents our ablation on the adversarial rank r_{adv}, providing direct empirical support for our core hypothesis. The results reveal a clear, non-monotonic relationship between rank and performance, highlighting a crucial trade-off. A low rank (e.g., 18) provides insufficient capacity to capture the full complexity of the vulnerable subspace, limiting performance. Conversely, a high rank (e.g., 22) begins to degrade performance, which we attribute to robust overfitting; by including too many minor components, the model starts fitting to noise in the limited downstream data, mirroring the failure mode of full-parameter fine-tuning.

The optimal performance achieved at a moderate rank (e.g., $r_{\text{adv}} = 19$ for CUB-200, $r_{\text{adv}} = 20$ for CIFAR-100) represents the "sweet spot" that best isolates the essential vulnerable subspace. As shown in Table 6, these optimal ranks correspond to using only 30-40% of the original parameters, demonstrating that PILO's superior performance is intrinsically linked to its targeted parameter efficiency.

Generalization to Transformer Architectures To verify that our approach is not limited to CNNs, we evaluated PILO on Vision Transformer (ViT) and Swin Transformer back-

Dataset	Evaluation Setting	Rank r_{adv} 18	19	20	21	22
CUB-200	Clean	49.60%	**52.00%**	48.74%	50.02%	45.60%
	PGD-10	20.61%	**21.59%**	19.95%	20.85%	18.24%
DOG-120	Clean	47.94%	**49.02%**	48.38%	48.58%	47.45%
	PGD-10	16.74%	**17.04%**	16.66%	16.88%	16.36%
CIFAR100	Clean	65.81%	66.16%	66.35%	**66.55%**	65.36%
	PGD-10	48.83%	50.48%	**51.37%**	50.11%	47.87%

Table 4: The choice of adversarial rank r_{adv} reveals a clear trade-off between model capacity and overfitting. Optimal performance is achieved at moderate ranks, validating our hypothesis that targeting a limited principal subspace is key. Best results are in **bold**.

Model	Evaluation Setting	CIFAR-10 AUTOLORA	PILO	Diff.
ViT-B/16	Clean	87.63%	87.25%	-0.38%
	PGD-10	69.24%	**71.05%**	**+1.81%**
Swin-Tiny	Clean	82.60%	**82.88%**	+0.28%
	PGD-10	66.62%	**67.86%**	**+1.24%**

Table 5: PILO's benefits generalize effectively to Transformer architectures, improving robust accuracy over the AUTOLORA baseline on both ViT and Swin models. This suggests our core principle is architecture-agnostic.

Rank r_{adv}	ResNet18 Ratio	ResNet50 Ratio
18	36.36%	28.92%
19	38.38%	30.53%
20	40.40%	32.13%
21	42.42%	33.74%
22	44.45%	35.35%

Table 6: The ratio of trainable parameters in PILO's adversarial branch relative to the full feature extractor's parameters. Optimal performance is achieved while using less than 41% of the original parameters.

bones. The results in Table 5 show that PILO consistently improves adversarial robustness over AUTOLORA while maintaining comparable clean accuracy. This successful application to fundamentally different architectures suggests that our core principle, that adversarial vulnerability is concentrated in the principal components of linear transformation layers, is a general phenomenon in deep neural networks. It confirms that PILO is not an architecture-specific technique but a broadly applicable and robust fine-tuning paradigm.

Performance Across Different Pre-training Strengths We also investigated whether PILO's effectiveness depends on the initial robustness of the pre-trained model. Table 7 compares various fine-tuning methods on ResNet18 models pre-trained with different adversarial budgets (ϵ_{pt}). The results show two key trends: first, as expected, a stronger pre-training budget provides a better foundation and improves performance for all methods. Second, and more importantly,

Method	Evaluation Setting	ϵ_{pt} used for Pre-training 1/255	2/255	4/255	8/255
TRADES	Clean	83.18%	84.15%	84.80%	85.29%
	PGD-10	71.11%	72.18%	73.99%	74.51%
TWINS	Clean	85.57%	86.62%	86.26%	87.28%
	PGD-10	71.35%	73.36%	74.08%	74.67%
AUTO-LORA	Clean	86.11%	86.96%	87.54%	86.96%
	PGD-10	73.28%	74.53%	75.71%	76.05%
PILO	Clean	**86.96%**	**87.12%**	**87.87%**	**88.03%**
	PGD-10	**73.65%**	**74.87%**	**75.84%**	**76.28%**

Table 7: Comparison of robust fine-tuning methods on CIFAR-10 using ResNet18 models pre-trained regardless of the initial robustness level (ϵ_{pt}) of the pre-trained ResNet18 model. Best results in each column are in **bold**.

PILO consistently outperforms all baselines across every pre-training level. This demonstrates that PILO is not merely exploiting an artifact of a specific checkpoint; rather, it is a superior fine-tuning method that more effectively preserves and enhances the model's intrinsic robustness, regardless of its starting point.

5 Conclusion

This paper introduced PILO, a parameter-efficient framework for robust transfer learning designed to address the critical issue of catastrophic forgetting of robustness on limited downstream data. Our work is founded on the principle that adversarial vulnerability is disproportionately concentrated in the principal components of a pre-trained model's weights. By operationalizing this insight, PILO employs a principled, asymmetric decomposition strategy: it performs a surgical, spectrally-guided update on the principal components to build robustness, while a separate, auxiliary low-rank branch handles adaptation to natural data. This approach serves as a powerful form of implicit spectral regularization, effectively mitigating the gradient conflicts that plague monolithic fine-tuning methods. Our extensive experiments show that PILO consistently establishes a new state-of-the-art, outperforming previous approaches in adversarial robustness while using as little as 29-40% of the trainable parameters, thereby enhancing both model resilience and computational efficiency.

Limitations and Future Work We identify two primary limitations that suggest avenues for future work. First, the reliance on singular value decomposition (SVD) as a pre-processing step, while feasible for current models, may pose computational challenges when scaling to extremely large foundation models. Future research could explore more efficient algorithms for approximating the principal subspace. Second, the optimal adversarial rank, r_{adv}, is currently determined via hyperparameter search. Developing methods to automatically or dynamically select this rank would improve the framework's efficiency and ease of use, representing another valuable direction for future work.

Acknowledgments

This work was supported in part by the Chengdu Science and Technology Program under Grant 2025-YF12-00009-RC, in part by the Sichuan Science and Technology Program under Grant 2024ZYD0147, Natural Science Foundation of Xinjiang Uygur Autonomous Region under Grant 2024D01A18, the National Natural Science Foundation of China (NSFC) under Grant 62376228, in part by the Natural Science Foundation of Sichuan Province under Grant 2026NSFSC1468, and in part by the Open Project Program of the Network and Data Security Key Laboratory of Sichuan Province under Grant 2026NDSF002.

References

[Andriushchenko *et al.*, 2020] 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*, pages 484–501. Springer, 2020.

[Bommasani *et al.*, 2021] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*, 2021.

[Buch *et al.*, 2018] Varun H Buch, Irfan Ahmed, and Mahiben Maruthappu. Artificial intelligence in medicine: current trends and future possibilities. *British Journal of General Practice*, 68(668):143–144, 2018.

[Carlini and Wagner, 2017] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017.

[Chen *et al.*, 2017] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM workshop on artificial intelligence and security*, pages 15–26, 2017.

[Croce and Hein, 2020a] Francesco Croce and Matthias Hein. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International conference on machine learning*, pages 2196–2205. PMLR, 2020.

[Croce and Hein, 2020b] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020.

[Deng *et al.*, 2009] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[Eykholt *et al.*, 2018] Kevin Eykholt, Ivan Evtimov, Earlence Fernandes, Bo Li, Amir Rahmati, Chaowei Xiao, Atul Prakash, Tadayoshi Kohno, and Dawn Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1625–1634, 2018.

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

[Griffin *et al.*, 2007] Gregory Griffin, Alex Holub, Pietro Perona, et al. Caltech-256 object category dataset. Technical report, Technical Report 7694, California Institute of Technology Pasadena, 2007.

[He *et al.*, 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*, pages 770–778, 2016.

[He *et al.*, 2019] Kaiming He, Ross Girshick, and Piotr Dollár. Rethinking imagenet pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4918–4927, 2019.

[Hendrycks *et al.*, 2019] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International conference on machine learning*, pages 2712–2721. PMLR, 2019.

[Houlsby *et al.*, 2019] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR, 2019.

[Howard and Ruder, 2018] Jeremy Howard and Sebastian Ruder. Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*, 2018.

[Hu *et al.*, 2022] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022.

[Khosla *et al.*, 2011] Aditya Khosla, Nityananda Jayadevaprakash, Bangpeng Yao, and Fei-Fei Li. Novel dataset for fine-grained image categorization: Stanford dogs. In *Proc. CVPR workshop on fine-grained visual categorization (FGVC)*, volume 2, 2011.

[Kornblith *et al.*, 2019] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2661–2671, 2019.

[Krizhevsky *et al.*, 2009] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

[Kumar *et al.*, 2022] Ananya Kumar, Aditi Raghunathan, Robbie Jones, Tengyu Ma, and Percy Liang. Fine-tuning can distort pretrained features and underperform out-of-distribution. *arXiv preprint arXiv:2202.10054*, 2022.

[Kurakin *et al.*, 2018] Alexey Kurakin, Ian J Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Artificial intelligence safety and security*, pages 99–112. Chapman and Hall/CRC, 2018.

[Lester *et al.*, 2021] Brian Lester, Rami Al-Rfou, and Noah Constant. The power of scale for parameter-efficient prompt tuning. *arXiv preprint arXiv:2104.08691*, 2021.

[Li and Liang, 2021] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation. *arXiv preprint arXiv:2101.00190*, 2021.

[Liu *et al.*, 2021] Xiao Liu, Kaixuan Ji, Yicheng Fu, Weng Lam Tam, Zhengxiao Du, Zhilin Yang, and Jie Tang. P-tuning v2: Prompt tuning can be comparable to fine-tuning universally across scales and tasks. *arXiv preprint arXiv:2110.07602*, 2021.

[Liu *et al.*, 2023] Ziquan Liu, Yi Xu, Xiangyang Ji, and Antoni B Chan. Twins: A fine-tuning framework for improved transferability of adversarial robustness and generalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16436–16446, 2023.

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

[Meng *et al.*, 2024] Fanxu Meng, Zhaohui Wang, and Muhan Zhang. Pissa: Principal singular values and singular vectors adaptation of large language models. *Advances in Neural Information Processing Systems*, 37:121038–121072, 2024.

[Papernot *et al.*, 2016] Nicolas Papernot, Patrick McDaniel, and Ian Goodfellow. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

[Pfeiffer *et al.*, 2020] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. *arXiv preprint arXiv:2007.07779*, 2020.

[Raffel *et al.*, 2020] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[Rebuffi *et al.*, 2017] Sylvestre-Alvise Rebuffi, Hakan Bilen, and Andrea Vedaldi. Learning multiple visual domains with residual adapters. *Advances in neural information processing systems*, 30, 2017.

[Rice *et al.*, 2020] Leslie Rice, Eric Wong, and Zico Kolter. Overfitting in adversarially robust deep learning. In *International conference on machine learning*, pages 8093–8104. PMLR, 2020.

[Salman *et al.*, 2020] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 33:3533–3545, 2020.

[Shafahi *et al.*, 2019] Ali Shafahi, Parsa Saadatpanah, Chen Zhu, Amin Ghiasi, Christoph Studer, David Jacobs, and Tom Goldstein. Adversarially robust transfer learning. *arXiv preprint arXiv:1905.08232*, 2019.

[Wah *et al.*, 2011] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. 2011.

[Xie *et al.*, 2020] Cihang Xie, Mingxing Tan, Boqing Gong, Jiang Wang, Alan L Yuille, and Quoc V Le. Adversarial examples improve image recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 819–828, 2020.

[Xu *et al.*, 2023] Xilie Xu, Jingfeng Zhang, and Mohan Kankanhalli. Autolora: A parameter-free automated robust fine-tuning framework. *arXiv e-prints*, pages arXiv–2310, 2023.

[Yosinski *et al.*, 2014] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? *Advances in neural information processing systems*, 27, 2014.

[Zhang *et al.*, 2019] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.