Balancing Generalization and Robustness in Adversarial Training via Steering through Clean and Adversarial Gradient Directions

Haoyu Tong
State Key Laboratory of Integrated Service Networks (ISN)
Xidian University
Xi'an, China
haoyutong@stu.xidian.edu.cn

Xiaoyu Zhang*
State Key Laboratory of Integrated Service Networks (ISN)
Xidian University
Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies
Xi'an, China
xiaoyuzhang@xidian.edu.cn

Yulin Jin
State Key Laboratory of Integrated Service Networks (ISN)
Xidian University
Xi'an, China
jyl990903@163.com

Jian Lou
State Key Laboratory of Integrated Service Networks (ISN)
Xidian University
Xi'an, China
jian.lou@hoiying.net

Kai Wu
School of Artificial Intelligence,
Xidian University
Xi'an, China
kwu@xidian.edu.cn

Xiaofeng Chen
State Key Laboratory of Integrated Service Networks (ISN)
Xidian University
Xi'an, China
xfchen@xidian.edu.cn

Abstract

Adversarial training (AT) is a fundamental method to enhance the robustness of Deep Neural Networks (DNNs) against adversarial examples. While AT achieves improved robustness on adversarial examples, it often leads to reduced accuracy on clean examples. Considerable effort has been devoted to handling the trade-off from the perspective of *input space*. However, we demonstrate that the trade-off can also be illustrated from the perspective of the *gradient space*. In this paper, we propose Adversarial Training with Adaptive Gradient Reconstruction (*AGR*), a novel approach that balances generalization (accuracy on clean examples) and robustness (accuracy on adversarial examples) in adversarial training via steering through clean and adversarial gradient directions. We first introduce an ingenious technique named Gradient Orthogonal Projection in the case of negative correlation gradients to adjust the adversarial gradient direction to reduce the degradation of generalization. Then we present a gradient interpolation scheme in the case of positive correlation gradients for efficiently increasing the generalization without compromising the robustness of the final obtained. Rigorous theoretical analysis proves that our *AGR* has lower generalization error upper bounds indicating its effectiveness. Comprehensive experiments empirically demonstrate that *AGR* achieves excellent capability of balancing generalization and robustness, and is compatible with various adversarial training methods to achieve superior performance. Our codes are available at: https://github.com/RUIYUN-ML/AGR.

CCS Concepts

• **Computing methodologies** → **Computer vision**; • **Security and privacy** → *Human and societal aspects of security and privacy*.

Keywords

Adversarial Training, Deep Neural Networks

ACM Reference Format:
Haoyu Tong, Xiaoyu Zhang, Yulin Jin, Jian Lou, Kai Wu, and Xiaofeng Chen. 2024. Balancing Generalization and Robustness in Adversarial Training via Steering through Clean and Adversarial Gradient Directions. In *Proceedings of the 32nd ACM International Conference on Multimedia (MM '24), October 28-November 1, 2024, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3664647.3680963

1 Introduction

Deep neural networks (DNNs) have gained widespread adoption in myriad multimedia processing fields thanks to their exceptional performance, such as image recognition [16, 20], text generation [1, 2], speech recognition [34, 61] and object detection [19, 50]. However, due to widely recognized vulnerabilities caused by various attacks, the security concerns and privacy protection issues for AI-backed multimedia systems are increasingly aggravating [12, 31, 32, 60]. Currently, a wide range of real-world applications have been shown to be vulnerable to adversarial examples (AEs) [13, 18, 21, 40], which adds tiny perturbations to clean examples that are imperceptible to human perception and give a false prediction. There exist considerable well-established methods for adversarial example generation such as FGSM [13], PGD [35], C&W [6], AutoAttack [9], *etc.* The introduction of adversarial examples has spurred the development of numerous techniques to ensure the trustworthiness of multimedia processing [14, 58, 59], research on the adversarial attack defense mechanisms has been a trending topic in the multimedia

*Corresponding author.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MM '24, October 28-November 1, 2024, Melbourne, VIC, Australia
© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 979-8-4007-0686-8/24/10
https://doi.org/10.1145/3664647.3680963

field [15, 46], among which adversarial training [35, 45, 48, 55] has been widely accepted as an effective defense.

Although AT and its variants have notably improved model robustness, they inevitably compromise generalization performance compared to standard training methodologies and vice versa. Recently, there has been a host of research dedicated to the trade-off between generalization and robustness [30, 37–39, 42, 44], such as leveraging the redundant capacity [63], instance reweighting [56, 57], adding a regularisation term for the loss function [41, 49], or leveraging mixed data [62]. However, most of the aforementioned works utilize only the information brought by the *input space* to improve the model generalization or robustness, paying little attention to the information in the *gradient space* and even less attention to the dynamics of the *gradient space*, which is also very important in influencing the model performance. In order to thoroughly investigate the trade-off issue from the perspective of *gradient space*, we conducted an experiment with the results depicted in Figure 1. We observe that there exist various levels of correlation between the gradient direction of the clean and adversarial robust loss during the adversarial training process, and even a negative correlation between their directions in some cases, which exactly explains that the adversarial training updating process leads to the gradual rise of the robustness and the degradation of the generalization. We perceive this finding as an indication of the existence of the trade-off problem, which can be explained from the perspective of the difference in gradient directions between clean loss and adversarial robust loss. In light of this phenomenon of trade-off observed from the gradient direction, we ask the intuitive question: *"Can we exploit these gradient directions to control or improve the trade-off in the training process dynamically?"* As an answer to the question, we are tackling the trade-off problem from a novel perspective by dividing the clean and adversarial directions of the gradient into two cases during adversarial training.

- In the negative correlation case, since the reduction in generalization comes from this situation, it is possible to adjust the direction of the adversarial gradient to remove that negative effect and maintain a positive correlation with the original gradient direction.
- In the positive correlation case, correlation-based gradient interpolation can be employed to enhance generalization, paying high attention to increasing generalization and low attention to maintaining robustness in high correlation and vice versa.

Based on the above inspiration, we propose a novel adversarial training method based on adjusting the direction of the gradients called Adversarial Training with Adaptive Gradient Reconstruction (*AGR*) that aims to improve the generalization without compromising the robustness of the final obtained. Our framework divides the parameter update process into two cases based on the cosine similarity between gradient pairs (see Figure 2). **1) Negative correlation case**, we propose the Gradient Orthogonal Projection (GOP) to decompose the adversarial gradient orthogonality into two parts, one aligned with the negative direction of the clean gradient and the other orthogonal to it. This achieves the desired effect that the new gradient direction causes no degradation of generalization; **2) Positive correlation case**, we propose an adaptive gradient interpolation scheme named Gradient Interpolation Based on Cosine Similarity (GICS), which makes use of the similarity of the gradients as the weights of the interpolation. With this interpolation scheme, the adversarial training process can dynamically pay more (or less) attention to robustness and less (or more) attention to generalization in the presence of low (or high) correlation, achieved by assigning interpolation weights corresponding to the gradient. This ensures efficient and dynamic improvement of generalization without compromising the robustness of the final obtained in adversarial training. The main contributions of our work can be summarized as follows:

- We analyze the trade-off problem that exists during adversarial training from a novel perspective, *i.e.*, *gradient space*, utilizing this perspective as a breakthrough point to further control or improve the trade-off.
- We propose an innovative adversarial training method named *AGR* to increase the generalization without compromising the robustness of the final obtained. Additionally, this approach is compatible with most of the existing adversarial training methods to achieve outstanding performance.
- We theoretically prove that our *AGR* has lower generalization error bounds compared to other AT methods. We also conducted extensive experiments to evaluate the *AGR* against five state-of-the-art adversarial attacks, demonstrating superior performance in handling the trade-off.

2 Related Work

Adversarial Attacks. Adversarial examples [40] have become a prevalent attack method imposing tiny perturbations to the model inputs that drive the target model mispredicts outputs. There has been a considerable amount of literature on the methods of their generation since their discovery. One of the earliest and most well-known methods is the Fast Gradient Sign Method (FGSM) [13]. Most subsequent adversarial attacks have been proposed based on this approach. Iterative FGSM (I-FGSM) [23] acts as an extended variant of the FGSM, which adds small perturbations by iteratively using the FGSM. PGD [35] is to randomly initialize a point on the input neighborhood as the starting point and then apply I-FGSM. Unlike FGSM [13], C&W [6] is an optimization-based attack that transforms the attack problem into a minimization problem with constraints, instead of making use of the model's gradient information to generate perturbations.

Adversarial Defense. In order to address and mitigate the potential threat of adversarial example attacks, manifold defense methods have been introduced and proposed. These include adversarial training [35, 45, 48, 55], adversarial detection [7, 51], certified adversarial robustness [8, 24, 26] and adversarial purification [3, 36, 52]. Among these techniques, adversarial training has emerged as a particularly effective defense method that involves training a model to improve its robustness by utilizing adversarial examples as training data. There are numerous approaches to adversarial training, each with its own unique strengths. Among them, PGD-AT [35] is the earliest and most widely accepted method, which improves model robustness by using maximization with PGD attack. TRADES [55] proposes a surrogate loss by analyzing the upper bound of the robust error, which pushes the decision boundary away from the examples. AWP [48] reveals a clear correlation between weight

landscape and robust generalization gap. By smoothing the weight loss landscape, the generalization gap can be effectively reduced to improve robustness. CFA [47] investigates the effects of adversarial perturbations on different classes, and improves the robustness by customizing the perturbation configurations of different classes during adversarial training.

3 Methodology

3.1 Preliminaries

Consider a K-class classification task on a dataset $\mathbb{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subseteq \mathcal{X} \times \mathcal{Y}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ represents examples drawn from a defined unknown distribution, and $\mathcal{Y}$ represents all possible labels corresponding to the examples in $\mathcal{X}$. The prediction of input data $\boldsymbol{x}$ on model f is denoted as $f(\boldsymbol{w}; \boldsymbol{x})$, where $\boldsymbol{w} \in \mathbb{R}^p$ represents the weight of the model. For an adversarial example classification problem, we use $\boldsymbol{x}'$ to denote the adversarial example of $\boldsymbol{x}$.

The complete clean loss function in standard training is defined as:

$$\mathcal{L}_{std}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \ell(f(\boldsymbol{w}; \boldsymbol{x}_i), y_i), \tag{1}$$

where n is the number of the training data, $f(\boldsymbol{w}; \cdot)$ is the model (neural network), $\ell(\cdot)$ is the loss function (*e.g.*, the cross-entropy (CE) loss).

For adversarial training, we denote the adversarial (or robust) loss as:

$$\mathcal{L}_{adv}(\boldsymbol{w}) = \frac{1}{n}\sum_{i=1}^{n} \max_{\|\boldsymbol{x}_i' - \boldsymbol{x}_i\|_p < \epsilon} \ell(f(\boldsymbol{w}; \boldsymbol{x}_i'), y_i), \tag{2}$$

where $\boldsymbol{x}'$ is considered an adversarial example that falls within the ϵ-ball, bounded by L_p-norm and centered at the clean example $\boldsymbol{x}$.

3.2 Overview

Adversarial training has been demonstrated to be an effective technique for enhancing a model's robustness against adversarial examples. However, one potential drawback of this approach is that it may decline the model classification accuracy on clean examples. This is because the model may become overly focused on defending against adversarial attacks, which can negatively impact its ability to correctly classify clean examples.

To address this issue, we propose a novel adversarial training approach with Adaptive Gradient Reconstruction (*AGR*). Specifically, *AGR* introduces orthogonal projection and gradient interpolation in adversarial training, which modifies the gradient of robust loss during the adversarial training process. As a result, *AGR* improves standard accuracy without compromising the robustness of the final obtained. Essentially, *AGR* allows the model to better balance its attention between defending against adversarial attacks and correctly classifying clean examples.

We first explore the application of orthogonal projection methods in adversarial training. This exploration provides a metric for distinguishing between positive and negative correlation gradient pairs. For convenience, we denote the gradient of the latter concerning $\mathcal{L}_{std}(\boldsymbol{w})$ and $\mathcal{L}_{adv}(\boldsymbol{w})$ as $\mathcal{L}_n$ and $\mathcal{L}_{adv}$, respectively.

Definition 3.1. The **cosine similarity** between the gradients $\nabla\mathcal{L}_n$ and $\nabla\mathcal{L}_{adv}$ is $\Psi(\nabla\mathcal{L}_n, \nabla\mathcal{L}_{adv}) = \frac{\nabla\mathcal{L}_n \cdot \nabla\mathcal{L}_{adv}}{\|\nabla\mathcal{L}_n\|_2 \|\nabla\mathcal{L}_{adv}\|_2}$.

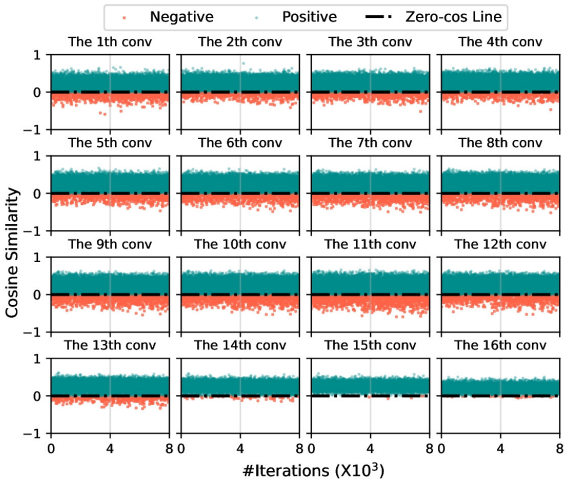

Figure 1: The cosine similarity of $\nabla\mathcal{L}_n$ and $\nabla\mathcal{L}_{adv}$ of weights of convolutional layers of PreActResNet-18 trained on CIFAR10 by TRADES. Below each scatter plot is the value of cosine similarity in the corresponding iteration rounds.

When the $\Psi(\cdot) < 0$, it indicates that the gradient is in the opposite direction for standard and adversarial training. We refer to this gradient pair as a negative correlation gradient pair. Whereas, when $\Psi(\cdot) > 0$, we call the gradient pair a positive correlation gradient pair. We also conduct a simple experiment in Figure 1 to show that there are many cases where the $\Psi(\cdot) < 0$ during the model training process.

In the adversarial training process, updating the model in the direction of $\nabla\mathcal{L}_{adv}$ will significantly affect its predictive performance at point $\boldsymbol{x}'$, but will result in a smaller change to the prediction of $\boldsymbol{x}'$ along the orthogonal to $\nabla\mathcal{L}_{adv}$. We consider the effect of both directions $\nabla\mathcal{L}_{adv}$ and $\nabla\mathcal{L}_n$ on the performance of the model simultaneously. Let us use cosine similarity as a measure between gradient pairs (i.e., $\Psi(\nabla\mathcal{L}_n, \nabla\mathcal{L}_{adv})$), then we can divide the gradient pairs into two types, positive correlation gradient pairs or negative correlation gradient pairs, corresponding to $\Psi(\nabla\mathcal{L}_n, \nabla\mathcal{L}_{adv}) \geq 0$ or $\Psi(\nabla\mathcal{L}_n, \nabla\mathcal{L}_{adv}) < 0$, respectively. Subsequently, we propose the gradient orthogonal projection and gradient interpolation methods to improve the generalization for the two cases (See Figure 2). To consider the orthogonalization process in more detail, we will orthogonalize the gradient for each parameter, we have:

$$\nabla\mathcal{L}_{adv} = [\nabla\mathcal{L}_{adv}^{(1)}, \nabla\mathcal{L}_{adv}^{(2)}, ..., \nabla\mathcal{L}_{adv}^{(c)}], \tag{3}$$

$$\nabla\mathcal{L}_n = [\nabla\mathcal{L}_n^{(1)}, \nabla\mathcal{L}_n^{(2)}, ..., \nabla\mathcal{L}_n^{(c)}], \tag{4}$$

where c is the number of the parameters.

3.3 Gradient Orthogonal Projection

As depicted in Figure 1, we demonstrate that there exists a few instances where the $\Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)}) < 0$ during adversarial training. It suggests that this update moving will increase the robustness but decrease the generalization. To reduce the loss of generalization, we propose to use "orthogonalize" to refine the gradient $\nabla\mathcal{L}_{adv}^{(i)}$.

We now introduce the **G**radient **O**rthogonal **P**rojection (GOP) clearly and concisely, using a formulaic approach. We first orthogonally project $\nabla\mathcal{L}_{adv}^{(i)}$ along $\nabla\mathcal{L}_n^{(i)}$ to obtain the new direction $\boldsymbol{g}_1^i$

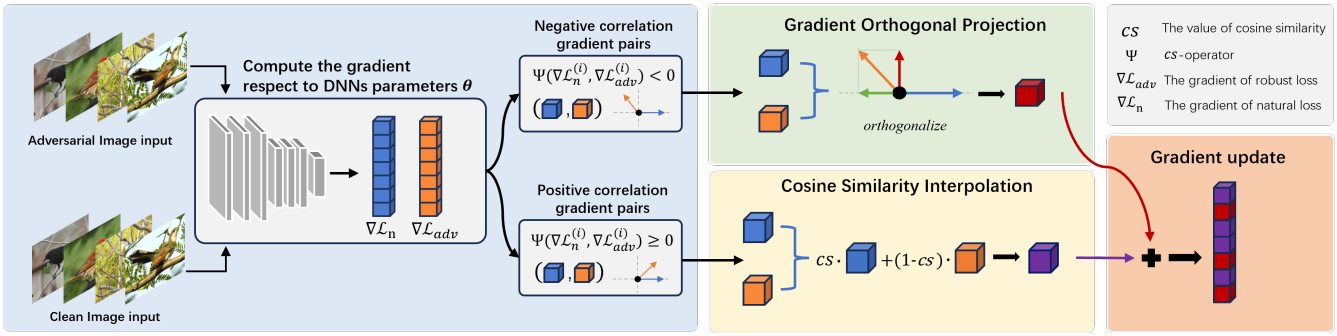

Figure 2: Overview of the proposed *AGR*.

as follows:

$$\boldsymbol{g}_1^i = \frac{\langle \nabla\mathcal{L}_{adv}^{(i)}, \nabla\mathcal{L}_n^{(i)} \rangle}{\langle \nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_n^{(i)} \rangle} \nabla\mathcal{L}_n^{(i)}. \tag{5}$$

In this scenario, the vectors of $\boldsymbol{g}_1^i$ and $\nabla\mathcal{L}_n^{(i)}$ are exactly opposite in direction, leading to reduced predictive accuracy on the model's clean examples. So we discard the component $\boldsymbol{g}_1^i$ and choose the component $\nabla\mathcal{L}_{adv}^{(i)}$ that orthogonal to $\nabla\mathcal{L}_n^{(i)}$ as the moving direction, we denote $Proj(\cdot)$ as follows:

$$\boldsymbol{g}_2^i = \nabla\mathcal{L}_{adv}^{(i)} - \boldsymbol{g}_1^i. \tag{6}$$

It is worth noting that we take $\boldsymbol{g}_2^i$ as the gradient of the i-th parameter update for this model does not cause much decrease in robustness as it is positively correlated with $\nabla\mathcal{L}_{adv}^{(i)}$ (*i.e.*, $\Psi(\boldsymbol{g}_2^i, \nabla\mathcal{L}_{adv}) > 0$).

In the case where $\Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)}) > 0$, let's reconsider using orthogonal decomposition for $\nabla\mathcal{L}_{adv}^{(i)}$ with respect to $\nabla\mathcal{L}_n^{(i)}$. The direction of $\boldsymbol{g}_1^i$ aligns with $\nabla\mathcal{L}_n^{(i)}$, indicating it doesn't hinder the model's prediction on clean examples. Since $\boldsymbol{g}_2^i$ is orthogonal to $\nabla\mathcal{L}_n^{(i)}$, it has minimal impact on clean prediction performance but positively correlates with $\nabla\mathcal{L}_{adv}^{(i)}$, enhancing robustness. Thus, discarding any component via gradient orthogonal projection negatively affects generalization or robustness (e.g., discarding $\boldsymbol{g}_1^i$ reduces generalization). Therefore, applying gradient orthogonal projection when $\Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)}) > 0$ is inappropriate, and we next consider a gradient interpolation method for this scenario.

While there exists a series of works that have used the GOP to solve a variety of problems. For example, Bryniarski *et al.* [5] explores a new attack technique aimed at constructing adversarial examples that satisfy multiple constraints simultaneously through GOP. Farajtabar *et al.* [11] proposes to use it to solve the *catastrophic forgetting* in continual learning. Li *et al.* [27] introduces subspace learning to federated unlearning via the orthogonal projection. To the best of our knowledge, we are the first work to pursue the improvement of generalization through gradient orthogonal projection.

3.4 Gradient Interpolation

Let's focus on improving generalization during training. After performing gradient orthogonal projection in adversarial training, the model achieves higher standard accuracy and consistent robustness compared to AT. However, it is worth noting that the increase in standard accuracy is often not very significant. Figure 1 shows that during the training process, only a small fraction of iterations have $\Psi(\cdot)$ values less than 0.

Overall, there are still a large number of positive correlations in the direction of the gradient here, and we will make reasonable utilization of this gradient information to further handle the trade-off. Our main idea is to consider gradient interpolation based on cosine similarity between $\nabla\mathcal{L}_{adv}^{(i)}$ and $\nabla\mathcal{L}_n^{(i)}$. Through interpolation, we can obtain a gradient that is very correlated with both gradient vectors, which can effectively balance the robustness and generalization. Thus, we propose the **G**radient **I**nterpolation Based on **C**osine **S**imilarity (GICS), the interpolation of the formula can be written as,

$$G = cs \cdot \nabla\mathcal{L}_n^{(i)} + (1 - cs) \cdot \nabla\mathcal{L}_{adv}^{(i)}, \tag{7}$$

where cs is the cosine similarity between $\nabla\mathcal{L}_{adv}^{(i)}$ and $\nabla\mathcal{L}_n^{(i)}$.

The above interpolation method is able to dynamically handle the balance between generalization and robustness by using the cosine similarity between $\nabla\mathcal{L}_{adv}^{(i)}$ and $\nabla\mathcal{L}_n^{(i)}$ as interpolation weighting factors. When the cosine similarity between gradients is larger, due to the larger cs this interpolation method will consider $\nabla\mathcal{L}_n^{(i)}$ more and efficiently gain generalization without loss of robustness. When the cosine similarity between gradients becomes small, the interpolation method will give priority to $\nabla\mathcal{L}_{adv}^{(i)}$ to prevent the decrease in robustness. Thus, it achieves the effect of improving generalization without loss of robustness. And when $\Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)})$ is less than 0, we do not opt for using gradient interpolation. We reason that the gradient pairs are negatively correlated at this moment, and the interpolated gradient will retain a small similarity to the clean gradients or adversarial robust gradients. This similarity makes it challenging to improve the generalization and robustness of the model.

No matter which interpolation method, it will lead to a certain difference between the new gradient direction and the original $\nabla\mathcal{L}_{adv}^{(i)}$, which will cause the model to fail to obtain the robustness in time, so we choose to take GICS after the model obtains the maximum robustness (*e.g.*, in the subsequent experiments we start using GICS at 150 epochs) and in order to finally maintain higher robustness, we choose to perform a gradient clipping on the interpolated gradient in l_2 norm, *i.e.*, the *Clip* operation is formulated as $G/max(1, \frac{\|G\|_2}{C})$, for a clipping threshold C. It is well documented that catastrophic overfitting of AT is related to the magnitude of the gradient paradigm, and reducing the magnitude

of the gradient paradigm can effectively suppress catastrophic forgetting [29]. To better facilitate understanding, the pseudo-code for *AGR* is presented in Supplementary.

4 Theoretical Analysis

In this section, we provide a general result (Theorem 4.1) to demonstrate the effectiveness of *AGR* in improving generalization while maintaining robustness during each iteration. Then, we use Rademacher complexity to bound the generalization error of *AGR*. Now, we consider generalization in terms of two parts of *AGR* to obtain the following results:

THEOREM 4.1. *Let $\nabla\mathcal{L}_n, \nabla\mathcal{L}_{adv}$ denote the clean and adversarial gradients, respectively. For arbitrary cases of $-1 \le \Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)}) < 0$ and $0 \le \Psi(\nabla\mathcal{L}_n^{(i)}, \nabla\mathcal{L}_{adv}^{(i)}) \le 1$, AGR with gradient $\boldsymbol{G}$ for one iteration, it holds that $\boldsymbol{G}$ induces a descent in both $\mathcal{L}_n$ and $\mathcal{L}_{adv}$.*

Theorem 4.1 shows that both in gradient orthogonal projection and cosine similarity interpolation operations, it is at least guaranteed that the direction of the gradient update causes no increase in clean loss and adversarial robust loss. Detailed notations and proof can be found in Supplementary. Next, we consider the following setting of Rademacher complexity.

There has an unknown distribution $\mathcal{D}$ over the input space $\mathcal{Z}$, from which we drawn N examples i.i.d from $\mathcal{D}$ to form the standard training dataset $S = \{z_1, z_2, ..., z_N\}$ where $z_i = (x_i, y_i)$. Similarly, we denote the adversarial training dataset S' drawn from the distribution $\mathcal{T}$. We formulate the population risk and empirical risk as:

$$R_{\mathcal{D}}(f) = E_{(x,y)\sim\mathcal{D}}[\ell(f(\boldsymbol{\theta}, x), y)], \tag{8}$$

$$R_S(f) = \frac{1}{n}\sum_{i=1}^{N}\ell(f(\boldsymbol{\theta}, x_i), y_i), \tag{9}$$

where $\ell(\cdot,\cdot)$ represents the loss function.

In our setting, *AGR* utilizes the standard training dataset and adversarial training dataset during adversarial training denoted as S and S', respectively. Further, *AGR* maintains the value of $\mathcal{L}_{std}$ without decreasing during the optimization of $\mathcal{L}_{adv}$, thus we could view f_{AGR} as learned from S and S',

$$f_{AGR} = \arg\min_{f\in\mathcal{F}} R_{S+S'}(f) = \arg\min_{f\in\mathcal{F}} \lambda R_S(f) + (1-\lambda)R_{S'}(f), \tag{10}$$

where $\mathcal{F}$ represents the function space and λ denotes the proportion between the standard training dataset and the adversarial training dataset. Following this, we give the definition of empirical Rademacher complexity and derive the generalization error of *AGR* based on Rademacher complexity.

Definition 4.2. Given a unknown distribution $\mathcal{D}$ and a function space $\mathcal{F}$, let $S = \{z_i\}_{i=1}^N$ denotes the training dataset drawn i.i.d from $\mathcal{D}$ and $\{\sigma_i\}_{i=1}^N$ be the independent random variables set drawn uniformly from $\{-1, 1\}$. Then, the empirical Rademacher complexity of $\mathcal{F}$ on the set S is defined to be:

$$\hat{\mathfrak{R}}_S(\mathcal{F}) = \mathrm{E}_\sigma \sup_{f\in\mathcal{F}}\Big[\frac{1}{N}\sum_{i=1}^{N}\sigma_i f(z_i)\Big]. \tag{11}$$

THEOREM 4.3. *Assume that $\mathcal{F}$ is a function space with the range [0,1], let $\boldsymbol{D}^{N_s} = \{z_n^s\}_{n=1}^{N_s}$ and $\boldsymbol{D}^{N_a} = \{z_n^a\}_{n=1}^{N_a}$ be two datasets of i.i.d sampled from the standard example distribution $\mathcal{D}$ and adversarial example distribution $\mathcal{T}$. Then, given $\lambda \in [0, 1)$ and for any $\epsilon > 0$, with probability at least $1-\epsilon$,*

$$\begin{aligned}
R_{\mathcal{D}}(f_{AGR}) - R_{S+S'}(f_{AGR}) &\le 2\lambda\hat{\mathfrak{R}}_S(\mathcal{F}) + 3\lambda\sqrt{\frac{\ln(2/\epsilon)}{2N_s}} \\
&+ (1-\lambda)D_{\mathcal{F}}(\mathcal{D},\mathcal{T}) + 2(1-\lambda)\hat{\mathfrak{R}}_{S'}(\mathcal{F}) \\
&+ 3(1-\lambda)\sqrt{\frac{\ln(2/\epsilon)}{2N_a}} + \sqrt{\frac{\ln(1/\epsilon)}{2}\Big(\frac{\lambda^2}{N_s} + \frac{(1-\lambda)^2}{N_a}\Big)} \\
&\le 2c\lambda B\frac{(\sqrt{2d\log 2}+1)\prod_{j=1}^{d}M_F(j)}{\sqrt{N_s}} + 3\lambda\sqrt{\frac{\ln(2/\epsilon)}{2N_s}} \\
&+ (1-\lambda)D_{\mathcal{F}}(\mathcal{D},\mathcal{T}) + 3(1-\lambda)\sqrt{\frac{\ln(2/\epsilon)}{2N_a}} \\
&+ 2c(1-\lambda)B\frac{(\sqrt{2d\log 2}+1)\prod_{j=1}^{d}M_F(j)}{\sqrt{N_a}} \\
&+ \sqrt{\frac{\ln(1/\epsilon)}{2}\Big(\frac{\lambda^2}{N_s} + \frac{(1-\lambda)^2}{N_a}\Big)}.
\end{aligned} \tag{12}$$

where $D_{\mathcal{F}}(\cdot,\cdot)$ represents the integral probability metric proposed by [54], $M_F(j)$ denotes the maximum value of the Frobenius norm for each parameter matrix W_j, and d is the depth of networks. Similarly, we bound the generalization error in standard adversarial training as:

$$\begin{aligned}
R_{\mathcal{D}}(f_{SA}) - R_{S'}(f_{SA}) &\le 2\hat{\mathfrak{R}}_{S'} + D_{\mathcal{F}}(\mathcal{D},\mathcal{T}) + 3\sqrt{\frac{\ln(2/\epsilon)}{2N_a}} + \sqrt{\frac{\ln(1/\epsilon)}{2N_a}} \\
&\le D_{\mathcal{F}}(\mathcal{D},\mathcal{T}) + 3\sqrt{\frac{\ln(2/\epsilon)}{2N_a}} + \sqrt{\frac{\ln(1/\epsilon)}{2N_a}} \\
&+ 2cB\frac{(\sqrt{2d\log 2}+1)\prod_{j=1}^{d}M_F(j)}{\sqrt{N_a}}.
\end{aligned} \tag{13}$$

In our setting, the ratio of clean to adversarial examples for *AGR* was kept at 1:1 (*i.e.*, $\lambda = 0.5$ and $N_S = N_a$). For conventional adversarial training methods, which usually utilize only the adversarial training dataset, its generalization error can be viewed as Eq. 13, while the error of *AGR* is shown in Eq. 12. It is worth noting that the TRADES method also utilizes the clean examples, and in subsequent experiments, we point out that *AGR* with TRADES as a special case of TRADES is dynamically adjusting the hyperparameters of the loss function of TRADES. Overall, the generalization error of f_{AGR} and f_{SA} are bounded by the empirical training risk, distributed error, and estimation error. The empirical training risk can be optimized to be infinitely small. The distribution error of f_{SA} is two times that of f_{AGR}. The remaining term in the inequality with respect to N_a and N_s is denoted the estimation error, which tends to 0 when the training dataset size is infinitely large. Therefore, our *AGR* can achieve a much smaller generalization error. The proof of Theorem 4.3 can be found in Supplementary.

5 Experiments

In this section, to verify the effectiveness, efficiency, and feasibility of our proposed *AGR*, we conduct extensive comparative experiments. Firstly, we show that our approach combined with currently available adversarial training methods (*i.e.*, AT [35], TRADES

Table 1: The standard accuracy (SA) and robust accuracy of PreAct-ResNet-18 and WideResNet-34 trained on CIFAR10, CIFAR100, and Tiny Imagenet datasets under $l_\infty = 8/255$ against white-box attacks across different defense mechanisms, *i.e.*, AT [35], TRADES [55], MART [45], and AGR (%).

Model (Architecture)	Methods	SA	FGSM	PGD-20	PGD-100	CW_∞	AutoAttack
CIFAR10 (PreAct-ResNet-18)	AT	$84.86_{\pm 0.014}$	$\mathbf{56.17}_{\pm 0.011}$	$46.30_{\pm 0.021}$	$45.89_{\pm 0.008}$	$46.03_{\pm 0.024}$	$44.01_{\pm 0.010}$
	With AGR	$\mathbf{86.22}_{\pm 0.004}$	$55.72_{\pm 0.015}$	$\mathbf{46.85}_{\pm 0.026}$	$\mathbf{46.36}_{\pm 0.022}$	$\mathbf{46.43}_{\pm 0.013}$	$\mathbf{44.91}_{\pm 0.009}$
	TRADES	$82.26_{\pm 0.165}$	$57.29_{\pm 0.036}$	$51.12_{\pm 0.051}$	$50.69_{\pm 0.042}$	$50.93_{\pm 0.029}$	$47.06_{\pm 0.033}$
	With AGR	$\mathbf{83.30}_{\pm 0.012}$	$\mathbf{58.31}_{\pm 0.027}$	$\mathbf{52.80}_{\pm 0.031}$	$\mathbf{52.54}_{\pm 0.033}$	$\mathbf{52.64}_{\pm 0.024}$	$\mathbf{48.78}_{\pm 0.011}$
	MART	$82.33_{\pm 0.032}$	$58.23_{\pm 0.005}$	$51.29_{\pm 0.014}$	$50.82_{\pm 0.010}$	$51.18_{\pm 0.021}$	$46.28_{\pm 0.011}$
	With AGR	$\mathbf{83.29}_{\pm 0.010}$	$\mathbf{58.26}_{\pm 0.039}$	$\mathbf{51.64}_{\pm 0.060}$	$\mathbf{51.21}_{\pm 0.057}$	$\mathbf{51.46}_{\pm 0.042}$	$\mathbf{47.34}_{\pm 0.034}$
CIFAR10 (WideResNet-34)	AT	$87.03_{\pm 0.012}$	$\mathbf{58.64}_{\pm 0.076}$	$49.06_{\pm 0.118}$	$48.53_{\pm 0.063}$	$48.44_{\pm 0.049}$	$47.71_{\pm 0.075}$
	With AGR	$\mathbf{87.53}_{\pm 0.007}$	$58.42_{\pm 0.007}$	$\mathbf{49.51}_{\pm 0.009}$	$\mathbf{49.14}_{\pm 0.014}$	$\mathbf{49.19}_{\pm 0.266}$	$\mathbf{48.36}_{\pm 0.018}$
	TRADES	$85.73_{\pm 0.012}$	$58.38_{\pm 0.014}$	$50.39_{\pm 0.022}$	$49.88_{\pm 0.020}$	$50.24_{\pm 0.004}$	$48.41_{\pm 0.007}$
	With AGR	$\mathbf{86.50}_{\pm 0.023}$	$\mathbf{60.31}_{\pm 0.034}$	$\mathbf{52.56}_{\pm 0.013}$	$\mathbf{52.04}_{\pm 0.011}$	$\mathbf{52.32}_{\pm 0.026}$	$\mathbf{50.44}_{\pm 0.020}$
	MART	$85.14_{\pm 0.018}$	$59.17_{\pm 0.058}$	$50.50_{\pm 0.160}$	$49.96_{\pm 0.142}$	$50.17_{\pm 0.074}$	$47.27_{\pm 0.061}$
	With AGR	$\mathbf{86.30}_{\pm 0.036}$	$\mathbf{60.83}_{\pm 0.125}$	$\mathbf{51.37}_{\pm 0.350}$	$\mathbf{50.87}_{\pm 0.316}$	$\mathbf{51.19}_{\pm 0.232}$	$\mathbf{48.74}_{\pm 0.117}$
CIFAR100 (PreAct-ResNet-18)	AT	$58.18_{\pm 0.048}$	$27.69_{\pm 0.018}$	$22.04_{\pm 0.029}$	$21.75_{\pm 0.024}$	$21.93_{\pm 0.013}$	$20.19_{\pm 0.009}$
	With AGR	$\mathbf{59.18}_{\pm 0.016}$	$\mathbf{27.82}_{\pm 0.012}$	$\mathbf{22.18}_{\pm 0.005}$	$\mathbf{21.78}_{\pm 0.008}$	$\mathbf{22.06}_{\pm 0.014}$	$\mathbf{20.33}_{\pm 0.011}$
	TRADES	$53.82_{\pm 0.012}$	$\mathbf{29.84}_{\pm 0.014}$	$\mathbf{27.02}_{\pm 0.011}$	$\mathbf{26.91}_{\pm 0.024}$	$\mathbf{27.13}_{\pm 0.021}$	$\mathbf{23.29}_{\pm 0.010}$
	With AGR	$\mathbf{54.53}_{\pm 0.009}$	$29.39_{\pm 0.012}$	$26.85_{\pm 0.006}$	$26.79_{\pm 0.014}$	$26.82_{\pm 0.004}$	$23.27_{\pm 0.006}$
	MART	$53.68_{\pm 0.015}$	$29.32_{\pm 0.132}$	$25.35_{\pm 0.166}$	$25.14_{\pm 0.154}$	$25.23_{\pm 0.116}$	$21.67_{\pm 0.096}$
	With AGR	$\mathbf{54.21}_{\pm 0.105}$	$\mathbf{29.48}_{\pm 0.009}$	$\mathbf{26.21}_{\pm 0.005}$	$\mathbf{25.95}_{\pm 0.032}$	$\mathbf{26.11}_{\pm 0.016}$	$\mathbf{22.91}_{\pm 0.022}$
CIFAR100 (WideResNet-34)	AT	$60.93_{\pm 0.079}$	$31.61_{\pm 0.0033}$	$26.05_{\pm 0.008}$	$25.65_{\pm 0.014}$	$25.88_{\pm 0.031}$	$24.33_{\pm 0.023}$
	With AGR	$\mathbf{61.98}_{\pm 0.026}$	$\mathbf{31.68}_{\pm 0.027}$	$\mathbf{26.41}_{\pm 0.013}$	$\mathbf{25.96}_{\pm 0.016}$	$\mathbf{25.99}_{\pm 0.007}$	$\mathbf{24.37}_{\pm 0.009}$
	TRADES	$57.10_{\pm 0.020}$	$31.23_{\pm 0.029}$	$26.96_{\pm 0.060}$	$26.75_{\pm 0.053}$	$26.93_{\pm 0.046}$	$24.55_{\pm 0.044}$
	With AGR	$\mathbf{58.17}_{\pm 0.003}$	$\mathbf{32.30}_{\pm 0.014}$	$\mathbf{28.34}_{\pm 0.012}$	$\mathbf{28.12}_{\pm 0.011}$	$\mathbf{28.21}_{\pm 0.004}$	$\mathbf{25.84}_{\pm 0.016}$
	MART	$57.29_{\pm 0.151}$	$30.33_{\pm 0.029}$	$26.03_{\pm 0.062}$	$25.88_{\pm 0.047}$	$25.96_{\pm 0.051}$	$23.92_{\pm 0.044}$
	With AGR	$\mathbf{58.10}_{\pm 0.224}$	$\mathbf{30.52}_{\pm 0.019}$	$\mathbf{27.61}_{\pm 0.033}$	$\mathbf{27.12}_{\pm 0.036}$	$\mathbf{27.24}_{\pm 0.025}$	$\mathbf{25.04}_{\pm 0.011}$
Tiny Imagenet (PreAct-ResNet-18)	AT	$31.90_{\pm 0.023}$	$\mathbf{11.09}_{\pm 0.016}$	$\mathbf{8.36}_{\pm 0.012}$	$\mathbf{8.28}_{\pm 0.011}$	$\mathbf{8.12}_{\pm 0.008}$	$6.47_{\pm 0.014}$
	With AGR	$\mathbf{35.56}_{\pm 0.004}$	$10.60_{\pm 0.014}$	$7.75_{\pm 0.018}$	$7.76_{\pm 0.007}$	$7.88_{\pm 0.008}$	$\mathbf{6.48}_{\pm 0.006}$
	TRADES	$31.06_{\pm 0.067}$	$\mathbf{12.04}_{\pm 0.017}$	$\mathbf{9.99}_{\pm 0.011}$	$\mathbf{10.08}_{\pm 0.013}$	$\mathbf{9.74}_{\pm 0.024}$	$\mathbf{7.20}_{\pm 0.019}$
	With AGR	$\mathbf{33.62}_{\pm 0.059}$	$11.47_{\pm 0.011}$	$9.73_{\pm 0.009}$	$9.62_{\pm 0.012}$	$9.65_{\pm 0.018}$	$6.48_{\pm 0.010}$
	MART	$29.19_{\pm 0.120}$	$12.65_{\pm 0.044}$	$11.25_{\pm 0.012}$	$10.80_{\pm 0.021}$	$11.03_{\pm 0.013}$	$7.75_{\pm 0.018}$
	With AGR	$\mathbf{31.10}_{\pm 0.082}$	$\mathbf{12.94}_{\pm 0.012}$	$\mathbf{11.29}_{\pm 0.006}$	$\mathbf{11.20}_{\pm 0.009}$	$\mathbf{11.06}_{\pm 0.015}$	$\mathbf{8.01}_{\pm 0.014}$
Tiny Imagenet (WideResNet-34)	AT	$33.92_{\pm 0.032}$	$10.66_{\pm 0.019}$	$7.57_{\pm 0.022}$	$7.55_{\pm 0.023}$	$7.23_{\pm 0.018}$	$6.54_{\pm 0.009}$
	With AGR	$\mathbf{37.30}_{\pm 0.032}$	$\mathbf{11.30}_{\pm 0.017}$	$\mathbf{8.42}_{\pm 0.004}$	$\mathbf{8.30}_{\pm 0.007}$	$\mathbf{8.18}_{\pm 0.005}$	$\mathbf{6.98}_{\pm 0.013}$
	TRADES	$32.60_{\pm 0.043}$	$12.22_{\pm 0.021}$	$9.65_{\pm 0.034}$	$9.85_{\pm 0.029}$	$9.47_{\pm 0.026}$	$7.63_{\pm 0.013}$
	With AGR	$\mathbf{33.57}_{\pm 0.016}$	$\mathbf{12.89}_{\pm 0.008}$	$\mathbf{11.55}_{\pm 0.011}$	$\mathbf{11.44}_{\pm 0.020}$	$\mathbf{11.25}_{\pm 0.015}$	$\mathbf{8.27}_{\pm 0.006}$
	MART	$30.79_{\pm 0.136}$	$13.42_{\pm 0.019}$	$11.82_{\pm 0.013}$	$11.53_{\pm 0.025}$	$11.02_{\pm 0.020}$	$7.37_{\pm 0.014}$
	With AGR	$\mathbf{32.96}_{\pm 0.093}$	$\mathbf{13.55}_{\pm 0.012}$	$\mathbf{11.89}_{\pm 0.007}$	$\mathbf{11.65}_{\pm 0.023}$	$\mathbf{11.47}_{\pm 0.008}$	$\mathbf{7.72}_{\pm 0.011}$

($\lambda = 1/6$) [55], MART [45], AWP [48], Avmixup [25] can significantly improve standard accuracy while ensuring robustness against various attacks. Furthermore, we conducted many ablation studies to confirm the effectiveness of individual components of our methodology. We use the standard accuracy and robustness accuracy as metrics to measure model performance.

5.1 Experimental Setup

Datasets. For all experiments, we evaluate the standard accuracy and robustness of the proposed *AGR* on CIFAR10 [22], CIFAR-100 [22], and Tiny Imagenet [10], which are three well-known datasets of natural images used in computer vision research. These datasets were divided into two parts, the training set, and the validation set, with a ratio of 5:1. For the data augmentations, we apply 32 × 32 random crops with 4-pixel zero padding, random horizontal flip, and cutout.

Training. For CIFAR10/100 and Tiny Imagenet, we utilize PreAct-ResNet-18 [17] and WideResNet-34 [53] architecture as the primary model for evaluation. For the setting of hyperparameters for adversarial training, we train the PreAct-ResNet-18 and WideResNet-34 for 200 epochs by SGD with momentum 0.9, and weight decay of 5×10^{-4}. The learning rate is initially 0.1, divided by 10 at the 100-th and 150-th calendar times. Regarding the generation of adversarial examples, we use PGD-10 [35] with the value of ϵ to 8/255, the step size to 2/255, and applied a limit of l_∞ constraint over 10 steps. More detailed settings can be found in the Supplementary.

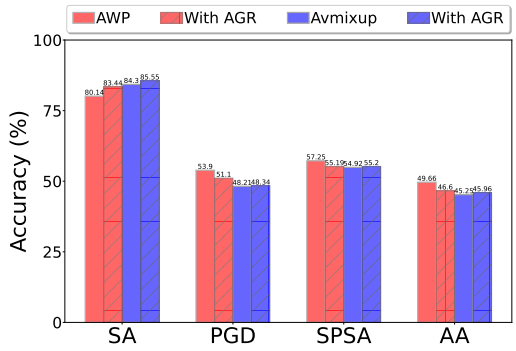

Figure 3: The standard accuracy (SA) and three robust accuracy of PreAct-ResNet-18 on CIFAR10 across AWP, Avmixup, and *AGR*.

Attacks. For white-box attacks, we consider the four typical attacks below: FGSM [13], PGD-20 [35], PGD-100 [35], and CW_∞ [6]. For black-box attacks, we choose SPSA attack [43] and AutoAttack [9], which contains a black box attack called square attack [4] and three white-box attacks. To ensure consistency in the experimental results, the mean of three experiment repetitions was employed for all results.

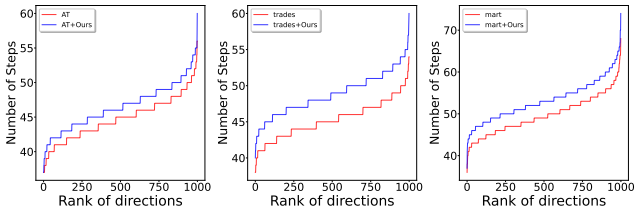

Figure 4: The decision boundary of *AGR*.

Table 2: The robust accuracy of PreAct-ResNet-18 on CIFAR-10/100 and TinyImagenet datasets under under l_2 and l_1 threat models (%).

Method	CIFAR10		CIFAR100		Tiny Imagenet	
	l_2 (128/255)	l_1 (2000/255)	l_2 (128/255)	l_1 (2000/255)	l_2 (128/255)	l_1 (2000/255)
AT	59.73$_{\pm 0.018}$	47.96$_{\pm 0.013}$	32.80$_{\pm 0.009}$	30.26$_{\pm 0.012}$	**14.40**$_{\pm 0.015}$	26.31$_{\pm 0.014}$
Ours	**60.76**$_{\pm 0.013}$	**48.84**$_{\pm 0.011}$	**33.81**$_{\pm 0.016}$	**31.09**$_{\pm 0.012}$	14.31$_{\pm 0.008}$	**26.52**$_{\pm 0.010}$
TRADES	61.89$_{\pm 0.042}$	45.82$_{\pm 0.036}$	34.88$_{\pm 0.013}$	32.68$_{\pm 0.017}$	16.32$_{\pm 0.020}$	28.35$_{\pm 0.031}$
Ours	**63.14**$_{\pm 0.035}$	**47.54**$_{\pm 0.037}$	**35.20**$_{\pm 0.028}$	**33.94**$_{\pm 0.013}$	**17.38**$_{\pm 0.017}$	**29.55**$_{\pm 0.025}$
MART	61.65$_{\pm 0.011}$	48.01$_{\pm 0.017}$	34.37$_{\pm 0.015}$	31.43$_{\pm 0.013}$	16.17$_{\pm 0.009}$	27.42$_{\pm 0.010}$
Ours	**61.94**$_{\pm 0.049}$	**48.11**$_{\pm 0.035}$	**34.50**$_{\pm 0.013}$	**31.97**$_{\pm 0.007}$	**17.55**$_{\pm 0.009}$	**27.87**$_{\pm 0.011}$

5.2 Main Results

Impact on standard accuracy. Table 1 shows that our proposed *AGR* method, when combined with diverse adversarial training methods, can significantly enhance the standard accuracy of the model on CIFAR10, CIFAR100, and Tiny Imagenet. Specifically, when applied in conjunction with adaptive gradients reconstruction, *AGR* achieves an impressive standard accuracy of 86.22% on CIFAR10 with PreAct-ResNet-18. This is a significant improvement compared to the baseline AT, which only manages to achieve 84.86% accuracy on clean images, resulting in a gap of 1.36%. Despite the gap in standard accuracy, the robustness accuracy of both methods is quite comparable, with only a marginal difference of 0.49%. The identical experimental results can also be observed on the Tiny Imagenet dataset.

Robustness against white-box attacks. To verify the reliability of our approach, we conducted a thorough evaluation of its robustness against a range of white-box attacks. Specifically, we considered various attacks with the same norm constraints (e.g., $l_\infty = 8/255$). The results, as presented in Table 1, demonstrate that our proposed method, which incorporates adaptive gradient reconstruction, consistently maintains exceptional robustness across all evaluated attacks. Even in the worst-case scenario, where the model was trained with TRADES on CIFAR100, there was only a 0.45% reduction in robustness. Over the Tiny Imagenet dataset, we also demonstrate the effectiveness of our method, which performs well in improving generalization. In addition to this, we further consider other adversarial training methods in conjunction with *AGR*, such as AWP [48], Avmixup [25]. The results are shown in Figure 3, where the proposed method is effective in improving generalization while maintaining robustness.

To demonstrate the robustness of our method against different white-box adversarial attacks, we evaluate our method against FGSM, PGD-100, and C&W. As shown in Table 1, our method is efficient in maintaining robustness against diverse adversarial attacks. We can observe that the combination of our *AGR* method also maintains reasonable robustness against unseen perturbations.

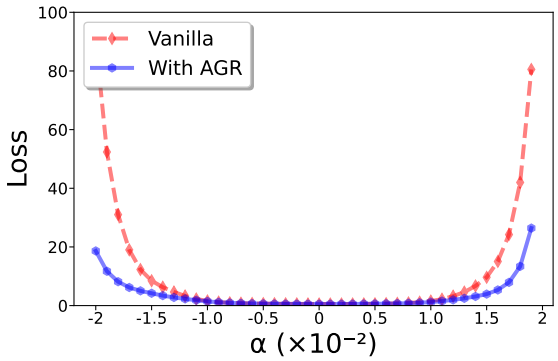

Figure 5: Comparison of the weight loss landscape for vanilla AT and AGR of PreAct-ResNet-18 trained on CIFAR10. These curves are the change in loss when moving model weight in the direction of a randomly sampled from a Gaussian distribution with the step size of α.

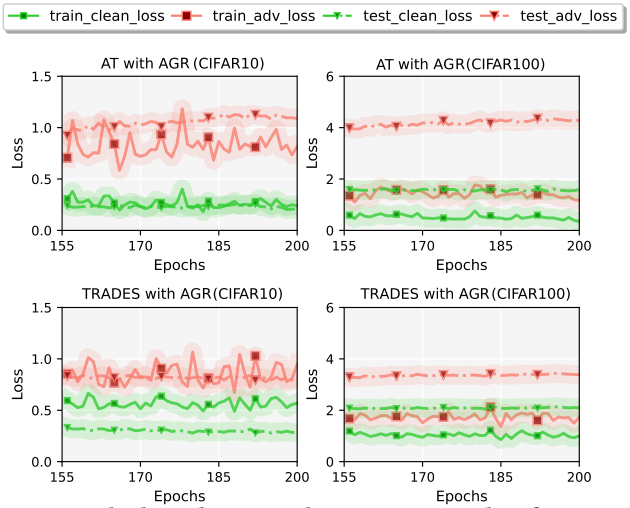

Figure 6: The loss change with respect to epochs of AT-AGR and TRADES-AGR of PreAct-ResNet-18 trained on CIFAR10 and CIFAR100.

Robustness against black-box attacks. We also conducted two black-box attack tests on our approach, *i.e.*, query-based attack SPSA and AutoAttack, assessing robustness by a black-box attack (Square Attack) and three white-box attacks. Table 1 and Figure 3 illustrate our approach's effectiveness in defending against black-box attacks.

Visualization results. To visualize that our method improves generalization, we show the decision boundary of *AGR* in Figure 4 and the weight loss landscape of the model trained with vanilla AT and *AGR* in Figure 5. For decision boundary, we randomly select an image and generate 1K random directions, applying continuously perturbations with a fixed step size ($\alpha = 1/255$) along each direction until the model's prediction for the image changes. We record the number of steps required to change the prediction result of the example in 1K directions in Figure 4 in ascending order. The decision boundary of our method is further away from the example than the baseline. Weight loss landscape is a commonly used measure to describe the generalization gap in standard training [28, 33]. We can observe that our approach has a much flatter loss compared to

Table 3: Comparison of DNNs model trained with/without proposed GOP in different adversarial training methods trained on CIFAR10 (%).

Method	Standard acc	Robust acc	GOP frequency
Vanilla AT	$86.14_{\pm 0.011}$	$46.42_{\pm 0.018}$	-
+ GOP	$\mathbf{86.22}_{\pm 0.004}$	$\mathbf{46.85}_{\pm 0.026}$	2.03
TRADES	$82.98_{\pm 0.014}$	$52.60_{\pm 0.024}$	-
+ GOP	$\mathbf{83.30}_{\pm 0.012}$	$\mathbf{52.80}_{\pm 0.031}$	14.93
MART	$82.94_{\pm 0.008}$	$51.43_{\pm 0.032}$	-
+ GOP	$\mathbf{83.29}_{\pm 0.010}$	$\mathbf{51.64}_{\pm 0.060}$	6.09

Table 4: Comparison of the effects of three interpolation methods on the generalization and robustness of vanilla AT trained on CIFAR10. Case 1 is to average the gradients of standard and robust loss, Case 2 is to mix the two gradients in a certain proportion(*i.e.*, $\tau = 0.9$), while Case 3 is to apply GICS (%).

Method	AT		TRADES		MART	
	Clean	PGD	Clean	PGD	Clean	PGD
+ Case 1	$86.04_{\pm 0.005}$	$46.70_{\pm 0.013}$	$\mathbf{83.32}_{\pm 0.011}$	$52.64_{\pm 0.017}$	$83.28_{\pm 0.008}$	$51.55_{\pm 0.015}$
+ Case 2	$85.10_{\pm 0.010}$	$48.42_{\pm 0.019}$	$82.48_{\pm 0.012}$	$52.84_{\pm 0.017}$	$82.54_{\pm 0.011}$	$52.55_{\pm 0.022}$
+ Case 3 (Ours)	$\mathbf{86.22}_{\pm 0.004}$	$46.85_{\pm 0.026}$	$83.30_{\pm 0.012}$	$52.80_{\pm 0.031}$	$\mathbf{83.29}_{\pm 0.010}$	$51.64_{\pm 0.060}$

vanilla AT, which indicates better generalization. As depicted in Figure 6, we show that our method achieves a remarkably smooth loss trained by vanilla AT and TRADES on CIFAR-10/100 datasets. The loss change result on Tiny Imagenet can be found in Supplementary.

Robustness against unseen attacks. We have conducted evaluations for other threat models. Table 2 reports the adversarial robustness using PRN-18 under l_2 and l_1 threat models, which indicates that the proposed method is also effective for other threat models.

5.3 Ablation Study

We aim to provide a thorough analysis of the proposed *AGR* method by conducting substantial ablation studies on each individual component. Our primary objective is to evaluate the effectiveness of the Gradient Orthogonal Projection (GOP) and Gradient Interpolation Based on Cosine Similarity (GICS) of *AGR* for standard accuracy enhancement, respectively.

Evaluation on gradient orthogonal projection. To evaluate GOP's ability to improve standard accuracy, we tested models trained with and without GOP, comparing it to vanilla AT, TRADES, and MART under consistent experimental settings. The results, as shown in Table 3, indicate that GOP significantly enhances standard accuracy in all adversarial training. Notably, TRADES and MART with GOP show greater improvement than vanilla AT, likely due to the frequency of orthogonal projection during training. In AT-GOP, GOP frequency is 2.03%, while in TRADES and MART, it reaches 14.93% and 6.09%, respectively.

Evaluation on gradient interpolation. We conducted a comparative experiment with three gradient interpolation methods: average gradients, fixed scale interpolation, and our GICS (Section 3.4). Table 4 shows that our method significantly improves standard accuracy, but in some cases, it results in decreased robustness compared to Case 2.

In particular, we point out that the interpolation method in TRADES is actually changing the hyperparameters λ of the loss function of TRADES, whereas our method (*i.e.*, Case 3) dynamically adjusts λ so that when robustness is lost too much, interpolation yields a larger λ to ensure that robustness is not lost. In this way, we can get a better balance between standard accuracy and robustness accuracy compared to fixed interpolation weights. The results of Case 2 with different τ values are shown in Figure 7. As τ increases, robust accuracy increases while standard accuracy decreases, indicating a static trade-off where the gradient focuses more on robustness over generalization. In Case 3, this trade-off is dynamically managed by adjusting the interpolation weight based on the gradient directions of clean and adversarial losses. As shown by the dotted line in Figure 7, Case 3 achieves a better balance between robustness and generalization compared to Case 2, often providing higher generalization with comparable robustness.

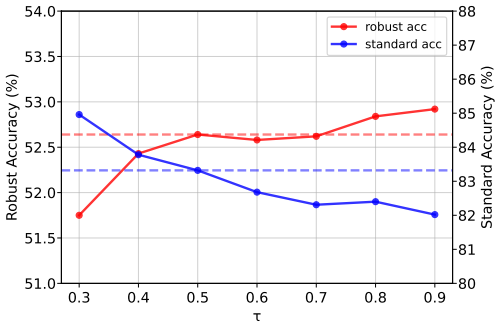

Figure 7: Comparison of robust and standard accuracy for TRADES-GI of PreAct-ResNet-18 on CIFAR10 under different τ values in Case 2. The red and blue dashed lines indicate TRADES-AGR results in Case 3.

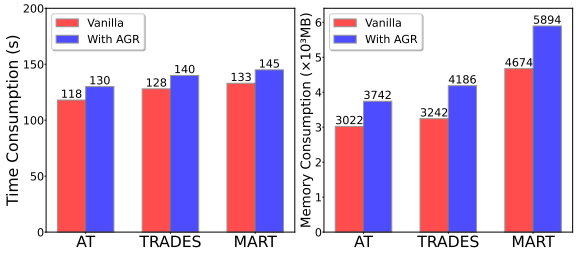

Figure 8: Comparison of time and memory consumption for PreActResNet-18 trained with AT, TRADES, MART, and *AGR* on CIFAR10. (Left) shows seconds per epoch, (Right) shows memory usage by each method.

Evaluation on time and memory consumption. Here we demonstrated the time and GPU memory consumption of our proposed method compared to the vanilla AT method in Figure 8. Time loss and memory usage are still increased compared to the original method.

6 Conclusion

In this paper, we perform a comprehensive study on how to balance generalization and robustness in adversarial training. From a novel perspective, we proposed Adversarial Training with Adaptive Gradients Reconstruction (*AGR*) to implement the gradient information of clean and adversarial examples to dynamically handle the trade-off between generalization and robustness in order to improve standard accuracy. Extensive experiments show that our method has an excellent performance in improving generalization while maintaining robustness.

Acknowledgments

This work is supported by the National Natural Science Foundation of China (No. 62102300, 62206207, 61960206014, and 62121001), Guangdong Provincial Key Laboratory of Novel Security Intelligence Technologies (No. 2022B1212010005), Fundamental Research Funds for the Central Universities (No. ZYTS24140), and '111 Center' (No. B16037).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] AI@Meta. 2024. Llama 3 Model Card. (2024). https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md

[3] Zeyuan Allen-Zhu and Yuanzhi Li. 2022. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*. IEEE, 977–988.

[4] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. 2020. Square attack: a query-efficient black-box adversarial attack via random search. In *European conference on computer vision*. Springer, 484–501.

[5] Oliver Bryniarski, Nabeel Hingun, Pedro Pachuca, Vincent Wang, and Nicholas Carlini. 2021. Evading adversarial example detection defenses with orthogonal projected gradient descent. *arXiv preprint arXiv:2106.15023* (2021).

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

[7] Liang Chen, Yong Zhang, Yibing Song, Lingqiao Liu, and Jue Wang. 2022. Self-supervised learning of adversarial example: Towards good generalizations for deepfake detection. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18710–18719.

[8] Jeremy Cohen, Elan Rosenfeld, and Zico Kolter. 2019. Certified adversarial robustness via randomized smoothing. In *international conference on machine learning*. PMLR, 1310–1320.

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

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

[11] Mehrdad Farajtabar, Navid Azizan, Alex Mott, and Ang Li. 2020. Orthogonal gradient descent for continual learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR, 3762–3773.

[12] Yue Fu, Qingqing Ye, Rong Du, and Haibo Hu. 2024. Interactive Trimming against Evasive Online Data Manipulation Attacks: A Game-Theoretic Approach. *arXiv preprint arXiv:2403.10313* (2024).

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

[14] Xiaolan Gu, Ming Li, and Li Xiong. 2023. DP-BREM: differentially-private and byzantine-robust federated learning with client momentum. *arXiv preprint arXiv:2306.12608* (2023).

[15] Jun Guo, Xingyu Zheng, Aishan Liu, Siyuan Liang, Yisong Xiao, Yichao Wu, and Xianglong Liu. 2023. Isolation and induction: Training robust deep neural networks against model stealing attacks. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4178–4189.

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

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Identity mappings in deep residual networks. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 630–645.

[18] Zhanhao Hu, Siyuan Huang, Xiaopei Zhu, Fuchun Sun, Bo Zhang, and Xiaolin Hu. 2022. Adversarial texture for fooling person detectors in the physical world. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 13307–13316.

[19] Zhou Huang, Hang Dai, Tian-Zhu Xiang, Shuo Wang, Huai-Xin Chen, Jie Qin, and Huan Xiong. 2023. Feature shrinkage pyramid for camouflaged object detection with transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5557–5566.

[20] Yan Jin, Mengke Li, Yang Lu, Yiu-ming Cheung, and Hanzi Wang. 2023. Long-Tailed Visual Recognition via Self-Heterogeneous Integration with Knowledge Excavation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 23695–23704.

[21] Zelun Kong, Junfeng Guo, Ang Li, and Cong Liu. 2020. Physgan: Generating physical-world-resilient adversarial examples for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 14254–14263.

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

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

[24] Mathias Lecuyer, Vaggelis Atlidakis, Roxana Geambasu, Daniel Hsu, and Suman Jana. 2019. Certified robustness to adversarial examples with differential privacy. In *2019 IEEE symposium on security and privacy (SP)*. IEEE, 656–672.

[25] Saehyung Lee, Hyungyu Lee, and Sungroh Yoon. 2020. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 272–281.

[26] Bai Li, Changyou Chen, Wenlin Wang, and Lawrence Carin. 2019. Certified adversarial robustness with additive noise. *Advances in neural information processing systems* 32 (2019).

[27] Guanghao Li, Li Shen, Yan Sun, Yue Hu, Han Hu, and Dacheng Tao. 2023. Subspace based Federated Unlearning. *arXiv preprint arXiv:2302.12448* (2023).

[28] Hao Li, Zheng Xu, Gavin Taylor, Christoph Studer, and Tom Goldstein. 2018. Visualizing the loss landscape of neural nets. *Advances in neural information processing systems* 31 (2018).

[29] Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. 2022. Subspace adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13409–13418.

[30] Yanxi Li and Chang Xu. 2023. Trade-Off Between Robustness and Accuracy of Vision Transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7558–7568.

[31] Shen Lin, Xiaoyu Zhang, Chenyang Chen, Xiaofeng Chen, and Willy Susilo. 2023. Erm-ktp: Knowledge-level machine unlearning via knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20147–20155.

[32] Ruixuan Liu, Tianhao Wang, Yang Cao, and Li Xiong. 2024. PreCurious: How Innocent Pre-Trained Language Models Turn into Privacy Traps. *arXiv preprint arXiv:2403.09562* (2024).

[33] Yong Liu, Siqi Mai, Xiangning Chen, Cho-Jui Hsieh, and Yang You. 2022. Towards efficient and scalable sharpness-aware minimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12360–12370.

[34] Pingchuan Ma, Alexandros Haliassos, Adriana Fernandez-Lopez, Honglie Chen, Stavros Petridis, and Maja Pantic. 2023. Auto-avsr: Audio-visual speech recognition with automatic labels. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 1–5.

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

[36] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. 2022. Diffusion models for adversarial purification. *arXiv preprint arXiv:2205.07460* (2022).

[37] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. 2022. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*. PMLR, 17258–17277.

[38] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. 2021. Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *ICML 2021 Workshop on Adversarial Machine Learning*.

[39] Rahul Rade and Seyed-Mohsen Moosavi-Dezfooli. 2021. Reducing excessive margin to achieve a better accuracy vs. robustness trade-off. In *International Conference on Learning Representations*.

[40] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).

[41] Jihoon Tack, Sihyun Yu, Jongheon Jeong, Minseon Kim, Sung Ju Hwang, and Jinwoo Shin. 2022. Consistency regularization for adversarial robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 8414–8422.

[42] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. 2018. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152* (2018).

[43] Jonathan Uesato, Brendan O'donoghue, Pushmeet Kohli, and Aaron Oord. 2018. Adversarial risk and the dangers of evaluating against weak attacks. In *International Conference on Machine Learning*. PMLR, 5025–5034.

[44] Haotao Wang, Tianlong Chen, Shupeng Gui, TingKuei Hu, Ji Liu, and Zhangyang Wang. 2020. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. *Advances in Neural Information Processing Systems* 33 (2020), 7449–7461.

[45] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*.

[46] Zhaoxin Wang, Handing Wang, Cong Tian, and Yaochu Jin. 2023. Adversarial Training of Deep Neural Networks Guided by Texture and Structural Information. In *Proceedings of the 31st ACM International Conference on Multimedia*. 4958–4967.

[47] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. 2023. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8193–8201.

[48] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems* 33 (2020), 2958–2969.

[49] Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems* 33 (2020), 6256–6268.

[50] Chang Xu, Jian Ding, Jinwang Wang, Wen Yang, Huai Yu, Lei Yu, and Gui-Song Xia. 2023. Dynamic Coarse-to-Fine Learning for Oriented Tiny Object Detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7318–7328.

[51] KiYoon Yoo, Jangho Kim, Jiho Jang, and Nojun Kwak. 2022. Detection of word adversarial examples in text classification: Benchmark and baseline via robust density estimation. *arXiv preprint arXiv:2203.01677* (2022).

[52] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. 2021. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*. PMLR, 12062–12072.

[53] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).

[54] Chao Zhang, Lei Zhang, and Jieping Ye. 2012. Generalization bounds for domain adaptation. *Advances in neural information processing systems* 25 (2012).

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

[56] Jianfu Zhang, Yan Hong, and Qibin Zhao. 2023. Memorization weights for instance reweighting in adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 11228–11236.

[57] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan Kankanhalli. 2020. Geometry-aware instance-reweighted adversarial training. *arXiv preprint arXiv:2010.01736* (2020).

[58] Xiaoyu Zhang, Xiaofeng Chen, Joseph K Liu, and Yang Xiang. 2019. DeepPAR and DeepDPA: privacy preserving and asynchronous deep learning for industrial IoT. *IEEE Transactions on Industrial Informatics* 16, 3 (2019), 2081–2090.

[59] Xiaoyu Zhang, Yulin Jin, Tao Wang, Jian Lou, and Xiaofeng Chen. 2022. Purifier: Plug-and-play backdoor mitigation for pre-trained models via anomaly activation suppression. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4291–4299.

[60] Xiaoyu Zhang, Shen Lin, Chao Chen, and Xiaofeng Chen. 2024. MODA: Model Ownership Deprivation Attack in Asynchronous Federated Learning. *IEEE Transactions on Dependable and Secure Computing* 21, 4 (2024), 4220–4235. https://doi.org/10.1109/TDSC.2023.3348204

[61] Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037* (2023).

[62] Long Zhao, Ting Liu, Xi Peng, and Dimitris Metaxas. 2020. Maximum-entropy adversarial data augmentation for improved generalization and robustness. *Advances in Neural Information Processing Systems* 33 (2020), 14435–14447.

[63] Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. 2023. Improving Generalization of Adversarial Training via Robust Critical Fine-Tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4424–4434.