

Revisiting and Exploring Efficient Fast Adversarial Training via LAW: Lipschitz Regularization and Auto Weight Averaging

Xiaojun Jia, *Student Member, IEEE*, Yuefeng Chen, Xiaofeng Mao, Ranjie Duan, Jindong Gu, Rong Zhang, Hui Xue and Xiaochun Cao *Senior Member, IEEE*

Abstract—Fast Adversarial Training (FAT) not only improves the model robustness but also reduces the training cost of standard adversarial training. However, fast adversarial training often suffers from Catastrophic Overfitting (CO), which results in poor robustness performance. Catastrophic Overfitting describes the phenomenon of a sudden and significant decrease in robust accuracy during the training of fast adversarial training. Many effective techniques have been developed to prevent Catastrophic Overfitting and improve the model robustness from different perspectives. However, these techniques adopt inconsistent training settings and require different training costs, i.e., training time and memory costs, leading to unfair comparisons. In this paper, we conduct a comprehensive study of over 10 fast adversarial training methods in terms of adversarial robustness and training costs. We revisit the effectiveness and efficiency of fast adversarial training techniques in preventing Catastrophic Overfitting from the perspective of model local nonlinearity and propose an effective Lipschitz regularization method for fast adversarial training. Furthermore, we explore the effect of data augmentation and weight averaging in fast adversarial training and propose a simple yet effective auto weight averaging method to improve robustness further. By assembling these techniques, we propose a FGSM-based fast adversarial training method equipped with Lipschitz regularization and Auto Weight averaging, abbreviated as FGSM-LAW. Experimental evaluations on four benchmark databases demonstrate the superiority of the proposed method over state-of-the-art fast adversarial training methods and the advanced standard adversarial training methods.

Index Terms—Fast adversarial training, Catastrophic overfitting, Lipschitz regularization, Auto weight averaging.

I. INTRODUCTION

Deep Neural Networks (DNNs) [1]–[4] can be easily fooled by Adversarial Examples (AEs) [5]–[9] which are carefully

crafted by adding imperceptible perturbations to benign samples. Adversarial Training (AT) has been demonstrated as one of the most effective methods to improve the model robustness against adversarial examples. In Standard Adversarial Training (SAT) [10]–[16], multi-step attack methods such as Projected Gradient Descent (PGD) are often applied to generate adversarial examples, which are computationally costly. This makes it challenging to apply AT on large-scale datasets. To reduce training expenses, Fast Adversarial Training (FAT) [17] has been proposed, which uses a single-step attack such as Fast Gradient Sign Method (FGSM) to approximate the multi-step attack for adversarial example generation. Recent work [18] has revealed that FGSM-AT does not increase the model’s resilience against adversarial examples. In detail, Wong *et al.* [18] first discover that FGSM-based AT is prone to Catastrophic Overfitting (CO), where model robustness initially increases during FGSM-AT but then sharply decreases after a few training epochs.

Some fast adversarial training variants [19]–[22] have been proposed to prevent Catastrophic Overfitting and improve the model robustness. Sample initialization [18], [20] and loss regularization [23] are two commonly used strategies to prevent Catastrophic Overfitting and improve the model robustness. Sample initialization [24], [25] provide better initialized perturbations of FGSM for improving the quality of the adversarial examples. Loss regularization [26] adopts regularity to guide the generation of stronger adversaries and smooths the loss surface for better robustness. Although these techniques can alleviate Catastrophic Overfitting effectively, some of them may introduce more training burden, violating the original intention of fast adversarial training. These inconsistent training time requirements can also lead to unfair comparisons.

To conduct a comprehensive study, we investigate the used techniques of over 10 papers on the fast adversarial training methods, which are summarized in Supplementary Material. We find the robustness improvement of these techniques is still correlated with training costs even in fast adversarial training. For example, Andriushchenko *et al.* [27] propose a gradient alignment regularization method that can relieve Catastrophic Overfitting and enhance model robustness. But calculating the regularization term requires a significant amount of time. And Zheng *et al.* [24] propose to use the adversarial perturbations from the last epoch as the sample initialization to improve model robustness, but this method requires extra computa-

Xiaojun Jia is with State Key Laboratory of Information Security, Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China, and also with School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China. (e-mail: jiaxiaojun@iie.ac.cn)

Yuefeng Chen, Xiaofeng Mao, Ranjie Duan, Rong Zhang and Hui Xue are with the Security Department of Alibaba Group. (e-mail: yuefeng.chenyf@alibaba-inc.com, mxfl64419@alibaba-inc.com, ranjieduan@gmail.com, stone.zhangr@alibaba-inc.com, hui.xueh@alibaba-inc.com)

Jindong Gu is with Department of Engineering Science in the University of Oxford, Oxford OX1 3PJ, United Kingdom. (e-mail: jindong.gu@outlook.com)

Xiaochun Cao (Corresponding) is with School of Cyber Science and Technology, Shenzhen Campus, Sun Yat-sen University, Shenzhen 518107, China (e-mail: caoxiaochun@mail.sysu.edu.cn)

Xiaochun Cao is the corresponding author

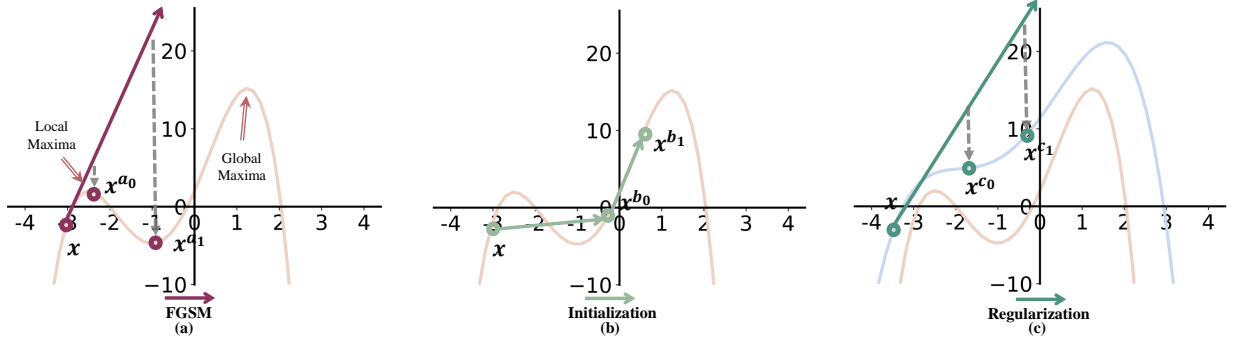


Fig. 1. Generation process of adversarial examples generated by FGSM in FAT on the maximization loss function. (a) Using a zero initialization. (b) Using the sample initialization. (c) Using the regularization.

tional costs and high memory consumption to restore previous perturbations. Therefore, conducting empirical studies to evaluate different FAT-related techniques quantitatively would be essential.

In this paper, we focus on exploring and finding efficient and effective FAT-related strategies. We roughly divide existing FAT-related techniques into two categories: sample initialization and regularization. In addition, we also explore the effect of several techniques on the robustness improvement for fast adversarial training, *i.e.*, data augmentation, and weight averaging. We conduct extensive experiments to evaluate the effectiveness and computational cost of various techniques in each category and conclude the following findings:

- **Sample Initialization** Prior/learning-based initialization prevents Catastrophic Overfitting and achieves better model robustness, which requires more training costs. Random sample initialization with an appropriate step size also prevents Catastrophic Overfitting but achieves limited model robustness improvement without extra training cost.
- **Regularization** Regularization is unnecessary during the inner maximization step in fast adversarial training, but regularization plays a critical role in the outer loss minimization step in fast adversarial training.
- **Data Augmentation** Different from SAT [28], fast adversarial training can achieve better model robustness when equipped with several sophisticated data augmentations.
- **Weight Averaging** In contrast to SAT [29], original fast adversarial training (FGSM-RS) cannot boost robustness through the mere implementation of weight averaging technique.

Based on the findings above, **a)** for sample initialization, considering the time consumption, we adopt random initialization. **b)** For regularization, motivated by Lipschitz constraint, we propose a novel regularization approach only use in minimization optimization to achieve better performance and less time overhead. **c)** For data augmentation, we find data augmentation can improve the quality of adversarial examples and adopt the Cutout for fast adversarial training. **d)** For weight averaging, we discover that the reason WA fails on original fast adversarial training (FGSM-RS) is that model

parameters accumulate non-robust parameters trained on the low-quality adversarial examples. Then we also propose a simple yet effective WA method to automatically select the robust parameters to conduct WA. By assembling the above techniques, we conclude our fast adversarial training method equipped with Lipschitz regularization and Auto Weight averaging, called FGSM-LAW. Our main contributions are in three aspects:

- We explore existing FAT-related techniques and find efficient and effective fast adversarial training techniques for further research.
- We propose a novel regularization approach motivated by the Lipschitz constraint, which can further improve the model robustness for fast adversarial training.
- We find WA fails to boost the original fast adversarial training and propose a simple yet effective WA method for original fast adversarial training (FGSM-RS) to improve model robustness further.
- By assembling these techniques, we conclude our fast adversarial training method, called FGSM-LAW. Compared with state-of-the-art fast adversarial training methods across various network architectures and datasets, FGSM-LAW achieves higher robustness performance with less training costs.

II. RELATED WORK

A. Adversarial Attack Methods

DNNs are vulnerable to adversarial examples. Several studies [9], [13], [30]–[36] focus on generating adversarial examples. In detail, Goodfellow *et al.* [37] first propose to adopt Fast Gradient Sign Method (FGSM) to attack DNNs, which makes use of the model gradient to generate adversarial examples. Then, a strong adversarial attack method, *i.e.*, Projected Gradient Descent (PGD), is proposed by Madry *et al.* [38]. It is an iterative extension of FGSM, which conducts FGSM with the project operation multiple times in a small attack step. To improve the attack performance, Carlini *et al.* [39] introduce three distance metrics that are used in the previous literature and propose some powerful adversarial attack methods, called C&W. Andriushchenko *et al.* [40] propose a score-based

black-box adversarial attack method based on a random search, called Square. Recently, Croce *et al.* [41] propose a fast adaptive boundary attack method that aims to find the minimal adversarial perturbation, called FAB. Moreover, Croce *et al.* [42] propose two powerful adversarial attack methods (APGD-CE and APGD-DLR), which can automatically select the attack step size. They combine the two previous adversarial attack methods (Square and FAB) with their own attack methods, called AutoAttack (AA).

B. Fast Adversarial Training

Even though adversarial training have been proposed to be one of the most effective defense methods, it requires high costs to generate adversarial examples, resulting in its limited practical application. To reduce the training costs, Goodfellow *et al.* [37] propose to adopt FGSM to generate adversarial examples for training, called FGSM-AT. Tramèr *et al.* [17] propose to initialize the clean samples with the Gaussian distribution to conduct FGSM-AT, called R+FGSM. Wong *et al.* [18] indicate that both FGSM-AT and R+FGSM easily encounter catastrophic overfitting (CO) and propose to initialize the clean samples with the uniform distribution instead of Gaussian distribution to conduct FGSM-AT, called FGSM-RS. Several fast adversarial training methods are proposed to improve the model robustness from the regularization perspective. In detail, Anndriushchenko *et al.* [27] propose a gradient alignment regularization to prevent CO and improve the model robustness. Also, Sriramanan *et al.* [23] propose a ℓ_2 distance regularization to improve the quality of adversarial examples. Sriramanan *et al.* [26] propose a Nuclear-Norm regularization to enforce function smoothing. Tsiligkaridis *et al.* [43] explore the correlation of catastrophic overfitting with low distortion for Frank-Wolfe attacks and propose an adaptive step fast adversarial training, called FW-ADAPT. Zhang *et al.* [44] revisit fast adversarial training from bi-level optimization (BLO) and propose a BLO-based fast adversarial training method, called FAST-BAT. Moreover, some research focuses on sample initialization to improve the performance of FGSM-AT. Specifically, Jia *et al.* [20] propose a learnable sample initialization generated by a generative model to boost fast adversarial training. Jia *et al.* [25] propose to use the prior adversarial perturbation as the sample initialization to improve robustness.

III. TECHNIQUES TO PREVENT OVERFITTING

A. Rethinking Catastrophic Overfitting

Catastrophic Overfitting (CO) indicates the phenomenon that the robust accuracy dramatically and suddenly decreases during the training parse of fast adversarial training, which is first discovered by [18]. A series of fast adversarial training variants have been proposed to prevent Catastrophic Overfitting and improve the robustness from two aspects, *i.e.*, sample initialization and regularization. Rethink that FGSM attack can be regarded as a closed-form solution to the maximization optimization problem, *i.e.*, $\delta = \arg \max_{\|\delta\|_\infty < \epsilon} \langle \nabla_{\mathbf{x}} \ell(\mathbf{x}, \mathbf{y}; \theta), \delta \rangle$. If the loss function is locally linear, the output of $\nabla_{\mathbf{x}} \ell(\mathbf{x}, \mathbf{y}; \theta)$ is constant within

the ℓ_∞ -ball around the input sample \mathbf{x} , which can provide the attacker with the optimal adversarial perturbation to attack the model. On the contrary, if the loss function is locally nonlinear, as shown in Fig 1 (a), the generated adversarial examples based on one large step FGSM may not be able to reach the local maximum region of the loss function, resulting in low-quality adversarial examples. [20], [25] have indicated that Catastrophic Overfitting happens when the adversarial example quality becomes worse. As shown in Fig 1 (b), better initialization can improve the quality of adversarial examples, which can reduce the effect of local nonlinearity to prevent Catastrophic Overfitting. Adding regularization to the loss function is another way of reducing local nonlinearity. In detail, as shown in Fig 1 (c), using the regularization can promote the flatness of the loss surface, which mitigates local nonlinearity and meanwhile improves the quality of adversarial examples generated by FGSM. Hence, Catastrophic Overfitting is directly related to the quality of the solution to the inner maximization and it is intrinsically caused by the model's local non-linearity. Initialization and regularization improve the quality of the adversarial examples to prevent Catastrophic Overfitting.

B. Sample Initialization

The sample initialization used in fast adversarial training is quite different. For a given clean sample \mathbf{x} initialized by η , fast adversarial training methods adopt FGSM to generate the adversarial perturbation δ . It can be defined as:

$$\delta = \Pi_{[-\epsilon, \epsilon]} [\eta + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \eta; \theta), \mathbf{y}))], \quad (1)$$

where ϵ represents the maximal perturbation, α represents the attack step size, $f(\cdot; \theta)$ represents the trained model with the weight parameters θ , \mathbf{y} represents the ground truth label. And the $\mathcal{L}(f(\mathbf{x} + \eta; \theta), \mathbf{y})$ is the loss function for training.

In detail, the primary FGSM-AT proposed by [37] conduct FGSM with a zero initialization to generate adversarial examples, *i.e.*, $\eta = 0$. Tramèr *et al.* [17] apply FGSM with a normal initialization in the half perturbation to generate adversarial examples for training, called **FGSM-NR**. It can be defined as:

$$\delta = \Pi_{[-\epsilon, \epsilon]} \left[\frac{\epsilon}{2} \cdot \text{Normal}(0, 1) + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \frac{\epsilon}{2} \cdot \text{Normal}(0, 1); \theta), \mathbf{y})) \right], \quad (2)$$

where α is set to $\epsilon/2$ for training. After that, Andriushchenko *et al.* [27] propose to implement FGSM with a uniform initialization in the whole perturbation to generate adversarial examples for training, called **FGSM-RS**. It can be defined as:

$$\delta = \Pi_{[-\epsilon, \epsilon]} [\epsilon \cdot \text{Uniform}(-1, 1) + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \epsilon \cdot \text{Uniform}(-1, 1); \theta), \mathbf{y}))], \quad (3)$$

where α is set to 1.25ϵ for training. Moreover, in the recent research ([23], [26]), they adopt a Bernoulli initialization in

the half perturbation to perform fast adversarial training, called **FGSM-BR**. It can be defined as:

$$\delta = \Pi_{[-\epsilon, \epsilon]} \left[\frac{\epsilon}{2} \cdot \text{Bernoulli}(-1, 1) + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \frac{\epsilon}{2} \cdot \text{Bernoulli}(-1, 1); \theta), \mathbf{y})) \right], \quad (4)$$

where α is set to ϵ for training. Also, Zheng *et al.* [24] propose to adopt the previous adversarial perturbation from the last epoch to initialize the clean sample, called **ATTA**. It can be defined as:

$$\delta_{E_{t+1}} = \Pi_{[-\epsilon, \epsilon]} [\delta_{E_t, \mathbf{x}} + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \delta_{E_t, \mathbf{x}}; \theta), \mathbf{y}))], \quad (5)$$

where α is set to ϵ for training, $\delta_{E_t, \mathbf{x}}$ is the generated adversarial perturbation by FGSM attack at the t -th epoch on the input sample \mathbf{x} . Besides, Jia *et al.* [25] adopt series of prior-guided adversarial perturbations as the adversarial initialization which are prior from the previous batch, the previous epoch and the momentum of all previous epochs to perform FGSM, called **FGSM-PGI**. It can be defined as:

$$\mathbf{g}_c = \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \boldsymbol{\eta}_{E_{t-1}}; \theta), \mathbf{y})), \quad (6)$$

$$\mathbf{g}_{E_t} = \mu \cdot \mathbf{g}_{E_{t-1}} + \mathbf{g}_c, \quad (7)$$

$$\delta_{E_t} = \Pi_{[-\epsilon, \epsilon]} [\boldsymbol{\eta}_{E_{t-1}} + \alpha \cdot \mathbf{g}_c], \quad (8)$$

$$\boldsymbol{\eta}_{E_t} = \Pi_{[-\epsilon, \epsilon]} [\boldsymbol{\eta}_{E_{t-1}} + \alpha \cdot \text{sign}(\mathbf{g}_{E_t})]. \quad (9)$$

where α is set to ϵ for training, \mathbf{g}_c is regarded as the signed gradient and \mathbf{g}_{E_t} is the signed gradient momentum in the t -th epoch. Moreover, Jia *et al.* [20] propose to adopt a generative model to generate the sample initialization for training, called **FGSM-SDI**. It can be defined as:

$$\delta = \Pi_{[-\epsilon, \epsilon]} [\epsilon \cdot g(\mathbf{x}, \mathbf{x}_g; \mathbf{w}) + \alpha \cdot \text{sign}(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \epsilon \cdot g(\mathbf{x}, \mathbf{x}_g; \mathbf{w}); \theta), \mathbf{y}))], \quad (10)$$

where α is set to ϵ for training, \mathbf{x}_g represents gradient information of \mathbf{x} on the trained model and $g(\cdot, \cdot; \mathbf{w})$ represents the generative model with the weight parameters \mathbf{w} .

We summarize existing initialization methods used in fast adversarial training and divide them into three categories: **a) random-based initialization**, **b) prior-based initialization** and **c) learning-based initialization**.

Random-based Initialization Most fast adversarial training methods [17], [26], [37] adopt random-based initialization methods to conduct AT. They always use a random sample-independent distribution to initialize the sample for the adversarial example generation, *i.e.*, $\boldsymbol{\eta} \sim \text{Random}(\epsilon)$, where $\text{Random}(\epsilon)$ represents a random distribution related to the maximum perturbation strength ϵ . In detail, Goodfellow *et al.* [37] conduct FGSM with a zero initialization to generate adversarial examples, *i.e.*, $\boldsymbol{\eta} = 0$. Tramèr *et al.* [17] apply FGSM with a normal initialization in the half perturbation to generate adversarial examples for training, *i.e.*, $\boldsymbol{\eta} = \frac{\epsilon}{2} \cdot \text{Normal}(0, 1)$, called FGSM-NR. Wong *et al.* [18] propose to implement FGSM with a uniform initialization in the whole perturbation to generate adversarial examples for training, *i.e.*, $\boldsymbol{\eta} = \epsilon \cdot \text{Uniform}(-1, 1)$, called FGSM-RS. Moreover, in the recent researchs [23], [26], they adopt a Bernoulli initialization

TABLE I
COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18 WITH DIFFERENT INITIALIZATION TECHNIQUES. NUMBER IN BOLD INDICATES THE BEST.

Method	Type	Clean	PGD-50	C&W	AA	Training Time (min)
FGSM-NR [17]	Random-based	86.41	43.99	45.4	41.29	51
FGSM-RS [18]	Random-based	86.65	44.12	45.65	42.2	51
FGSM-BR [23], [26]	Random-based	86.82	44.23	45.89	42.37	51
ATTA [24]	Prior-based	86.30	44.97	43.8	42.84	51
FGSM-PGI [25]	Prior-based	86.61	45.69	44.8	43.26	51
FGSM-SDI [20]	Learning-based	85.25	51.79	50.29	47.91	83

in the half perturbation to perform fast adversarial training, *i.e.*, $\boldsymbol{\eta} = \frac{\epsilon}{2} \cdot \text{Bernoulli}(-1, 1)$, called FGSM-BR.

Prior-based Initialization Several works [24], [25] also propose to use the previous adversarial perturbations as the sample initialization, which can be called prior-based initialization, *i.e.*, $\boldsymbol{\eta} \sim \text{Prior}(\epsilon, \mathbf{x})$, where $\text{Prior}(\epsilon, \cdot)$ represents the function related to the maximum perturbation strength ϵ and the sample \mathbf{x} , *i.e.*, sample-dependent initialization. Zheng *et al.* [24] propose to adopt the previous adversarial perturbation from the last epoch to initialize the clean sample, *i.e.*, $\boldsymbol{\eta} = \delta_E$, called ATTA. And Jia *et al.* [25] propose to use some prior-guided adversarial perturbations as the adversarial initialization which is prior from the previous epoch and the momentum of all previous epochs to perform FGSM, *i.e.*, $\boldsymbol{\eta} = \delta_P$, called FGSM-PGI. Note that in this section, for a fair comparison, we adopt the FGSM-PGI without the proposed regularization in their original paper to conduct experiments.

Learning-based Initialization Jia *et al.* [20] propose to adopt a sample-dependent learnable initialization to conduct AT. In detail, they adopt a generative model to generate the sample initialization for training, called FGSM-SDI, *i.e.*, $\boldsymbol{\eta} = \alpha \cdot g(\mathbf{x}, \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \mathbf{y}); \mathbf{w})$, where α is set to ϵ for training, and $g(\cdot, \cdot; \mathbf{w})$ represents the generative model with the weight parameters \mathbf{w} .

Discussion Previous works adopt different sample initialization for AT to prevent Catastrophic Overfitting and improve the model robustness. As shown in Fig. 1, using a zero initialization, FGSM with a large step size obtains a worse solution to the inner maximization problem. But, using sample initialization, FGSM can achieve a better solution to the inner maximization problem. The quality of the solutions to the inner maximization problem is related to the Catastrophic Overfitting. As shown in Fig. 2, FGSM-AT with a small step size (e.g., $\alpha = 4/255$) does not meet Catastrophic Overfitting and the attack success rate and model robustness are directly related in fast adversarial training. It indicates that when the quality of the inner maximization problem becomes worse, Catastrophic Overfitting could happen. Using a zero initialization, FGSM with a larger step size more easily meets Catastrophic Overfitting. Using sample initialization can prevent Catastrophic Overfitting when FGSM with a large step size. We summarize the existing initialization methods and make some meaningful findings based on the quality of the solutions to the inner maximization problem. First, sample-dependent initialization methods (prior-based and learning-

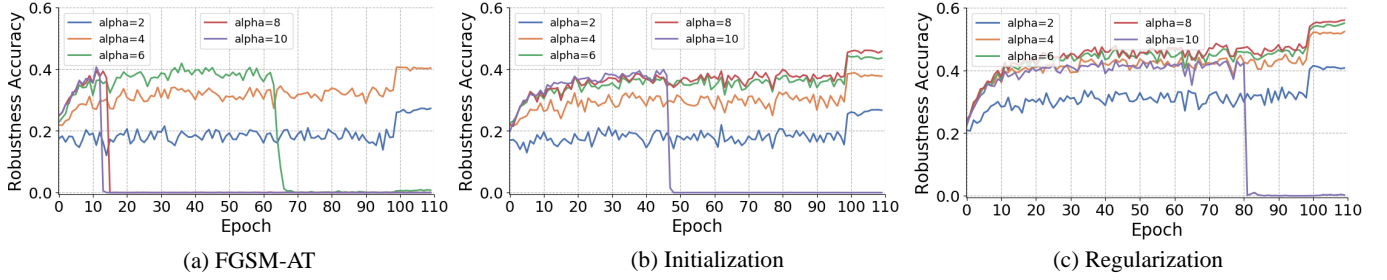


Fig. 2. The robustness accuracy under PGD-10 of FGSM-AT, FGSM with Bernoulli random initialization and FGSM with the regularization of different step sizes on the CIFAR-10 during the training phase.

based initialization) can achieve better robustness performance than sample-independent initialization methods (random-based initialization). The learning-based initialization achieves the best robustness performance. Second, surprisingly, random-based initialization methods with the appropriate step size prevent Catastrophic Overfitting and achieve approximately the same robustness improvement performance, which has not been explored in the previous works. We adopt ResNet18 for AT with different initialization to conduct experiments on CIFAR10. The result is shown in Table I, which demonstrates the correctness of our findings. Compared with the previous random initialization, the advanced ATTA and FGSM-PGI need more memory to store the historical adversarial perturbations, which could require more memory costs. And FGSM-SDI needs more time consumption to train an extra generator. We use random initialization based on the Bernoulli initialization and the appropriate step size to conduct fast adversarial training since it can achieve comparable robustness with less training costs.

C. Regularization

Regularization used in AT can be divided into two categories: a) regularization worked in the inner maximization and b) regularization worked in the outer minimization. The former is related to the inner maximization to improve the quality of the adversarial examples. The latter is related to the outer minimization to improve the model robustness. Existing fast adversarial training variants adopt the regularization in min-max optimization or only in minimization optimization. It can be defined as:

$$\begin{aligned} \delta^* &= \arg\max_{\delta \in \Omega} [\mathcal{L}(f(\mathbf{x} + \delta; \theta), \mathbf{y}) + \gamma \cdot R(\mathbf{x}, \mathbf{x} + \delta; \theta)] \\ \min_{\theta} \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}} [\mathcal{L}(f(\mathbf{x} + \delta^*; \theta), \mathbf{y}) + \mu \cdot R(\mathbf{x}, \mathbf{x} + \delta^*; \theta)], \end{aligned} \quad (11)$$

where $\gamma, \mu \in [0, 1]$ control the use of regularization and $R(\mathbf{x}, \mathbf{x} + \delta; \theta)$ represents the regularization.

Regularization only in minimization optimization Andriushchenko *et al.* [27] propose a gradient alignment regularization to measure the local linearity. They adopt the regularization only in the minimization optimization, *i.e.*, $\gamma = 0$ and $\mu = 1$. It can be defined as:

$$\begin{aligned} R(\mathbf{x}, \mathbf{x} + \delta; \theta) \\ = 1 - \cos(\nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x}; \theta), \mathbf{y}), \nabla_{\mathbf{x}} \mathcal{L}(f(\mathbf{x} + \eta; \theta), \mathbf{y})), \end{aligned} \quad (12)$$

where η represents the sample initialization. Although the FGSM-GA achieves model robustness improvement, it re-

quires extra training time to compute the alignment regularization on the gradient.

Regularization in min-max optimization Sriramanan *et al.* [23] propose a more effective guided regularization for improved optimization and function smoothing. They adopt regularization in min-max optimization. It works in the inner maximization to generate adversarial examples and the outer minimization to improve the model robustness, *i.e.*, $\gamma = 1$ and $\mu = 1$. It can be defined as:

$$R(\mathbf{x}, \mathbf{x} + \delta; \theta) = \lambda \cdot \|f(\mathbf{x} + \delta; \theta) - f(\mathbf{x}; \theta)\|_2^2, \quad (13)$$

where λ represents the a hyper-parameter that determines the smoothness of the loss surface. Sriramanan *et al.* [26] propose a Nuclear-Norm regularization for enforcing function smoothing in min-max optimization, *i.e.*, $\gamma = 1$ and $\mu = 1$. It can be defined as:

$$R(\mathbf{x}, \mathbf{x} + \delta; \theta) = \lambda \cdot \|f(\mathbf{x} + \delta; \theta) - f(\mathbf{x}; \theta)\|_*, \quad (14)$$

where $\|\cdot\|_*$ represents the Nuclear Norm which is the sum of the singular values.

Discussion We conclude that the regularization in min-max optimization achieves better robustness. It greatly improves model robustness. In inner maximization, the regularization loss is used to find the least smooth domain of the loss surface. In outer minimization, the regularization is used to improve the local smoothness of the loss surface. Although the regularization is effective to improve the model robustness, it also requires a computation overhead. On the other hand, adversarial examples almost exist in the least smooth domain of the loss surface [28], [45] which has the same effect as inner maximization in regularization. Based on this, we argue that it is not necessary to use regularization in the inner maximization. To verify this opinion, we conduct experiments on CIFAR-10 with ResNet18 comparing regularization in the min-max optimization and regularization solely in the minimization. We adopt FGSM with BR initialization as the baseline. The results are shown in Table II. Compared with fast adversarial training with the regularization in min-max optimization, Fast adversarial training with the regularization only in the minimization achieves better performance under PGD attack scenarios and comparable robustness performance under C&W and AA attack scenarios. The regularization only in the minimization requires less training time.

D. The Proposed Regularization

Previous works [46], [47] have demonstrated that the robustness is related to the local Lipschitzness, *i.e.*, smaller local

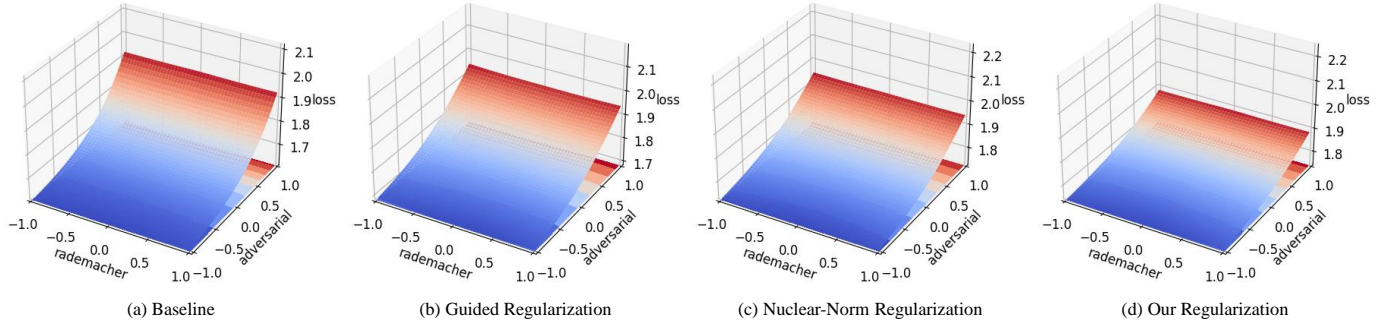


Fig. 3. Visualization of the loss landscape on CIFAR10 of different regularization methods, *i.e.*, the baseline FGSM-BR, FGSM-BR with the guided regularization(min-max), FGSM-BR with the nuclear-norm regularization(min-max) and FGSM-BR with our proposed regularization. We randomly select 1000 test images for CIFAR-10 testing dataset and plot the cross-entropy loss along a spatial variation composed of two directions, *i.e.*, a Rademacher (random) direction r_1 and an adversarial direction r_2 . The Rademacher (random) direction is calculated as: $r_1 \sim \text{Rademacher}(\eta)$, where η is 8/255. The adversarial direction is calculated as: $r_2 = \eta \text{sign}(\nabla_x f(\mathbf{x}; \theta))$. Note the same adversarial attack (PGD-50) are conducted on the same images for the visualization.

TABLE II

COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18 WITH DIFFERENT REGULARIZATION METHODS. NUMBER IN BOLD INDICATES THE BEST.

Method	Clean	PGD-50	C&W	AA	Training Time (min)
Baseline	86.82	44.23	45.89	42.37	51
Guided Regularization(min-max)	83.33	50.96	48.64	46.56	102
Guided Regularization(min)	84.20	51.26	47.87	45.74	73
Nuclear-Norm Regularization(min-max)	84.9	49.79	49.14	47.36	102
Nuclear-Norm Regularization(min)	84.94	50.78	48.83	47.06	73
Ours(min)	82.97	54.29	50.41	48.05	73

TABLE III

TEST ROBUSTNESS (%) ON THE CIFAR-10 DATABASE USING RESNET18 WITH DIFFERENT DATA AUGMENTATION TECHNIQUES. NUMBER IN BOLD INDICATES THE BEST.

Method	Clean	PGD-10	PGD-20	PGD-50	C&W	AA
Baseline	86.82	46.62	44.76	44.23	45.89	42.37
FGSM-BR+Mixup	86.94	46.78	44.93	44.41	45.52	42.34
FGSM-BR+Cutout	84.93	48.31	46.9	46.43	46.91	43.79
FGSM-BR+AutoAugment	85.79	48.97	47.24	46.57	46.06	42.25
FGSM-BR+CutMix	81.12	50.45	49.28	48.95	45.36	42.87

Lipschitzness leads to stronger robustness. We propose a novel regularization approach motivated by the Lipschitz constraint. It can be defined as:

$$\frac{\lambda}{\|\delta - \eta\|_2^2} \cdot (\|f(\mathbf{x} + \delta; \theta) - f(\mathbf{x} + \eta; \theta)\|_2^2 + \|f_F(\mathbf{x} + \delta; \theta) - f_F(\mathbf{x} + \eta; \theta)\|_2^2), \quad (15)$$

where $f_F(\cdot; \theta)$ represents the feature output of $f(\cdot; \theta)$ and λ is the hyper-parameter. We also compared the proposed regularization with the previous regularization methods, the results are shown in Table II. The proposed regularization achieves the best performance under all attack scenarios. We also visualize the loss landscape of each regularization methods. As shown in Fig. 3, it can be observed that using the regularization methods can make the cross-entropy loss more linear in the adversarial direction. Our proposed regularization method achieves the most linear cross-entropy loss in the adversarial direction. It indicates that our method can improve the local linearity of the trained model better. Moreover, We provide proof to prove the effectiveness of **the proposed regularization** in Supplementary Material.

IV. TECHNIQUES TO IMPROVE MODEL ROBUSTNESS

A. Data Augmentation

Several works [28], [48] have explored the impact of more sophisticated data augmentation techniques on the robustness of PGD-based AT. However, the impact of the data

augmentation techniques on model robustness of FGSM-based AT has not been studied before. In this paper, we introduce the sophisticated data augmentation techniques (Cutout [49], Mixup [50], CutMix [51], and AutoAugment [52]) into existing fast adversarial training techniques and explore the impact of them on model robustness of fast adversarial training methods. In detail, we use FGSM-BR as the baseline and compare it with FGSM-BR with data augmentation techniques. The results are shown in Fig. 4. In detail, we adopt the PGD-10, PGD-20, PGD-50, C&W and AA to evaluate the FGSM-BR with different data augmentation methods. The result is shown in Table III. It is clear that using Cutout, Mixup, CutMix and AutoAugment can improve model robustness. FGSM-BR combined with Cutout achieves the best robustness performance in all attack scenarios. Moreover, Cutout also achieves comparable clean accuracy on the clean images to the vanilla fast adversarial training (FGSM-BR). Hence, it can be observed that combined with data augmentation, fast adversarial training can achieve better robustness performance in all the attack scenarios. Particularly, under the PGD-10 attack, Cutout improves the performance by about 1.5%. Compared with the vanilla FGSM-BR, using data augmentation methods can improve the quality of the solution to the inner maximization problem, *i.e.*, the higher attack success rate. Unlike PGD-based AT [28], using data augmentation techniques can improve the robustness of FGSM-based adversarial training.

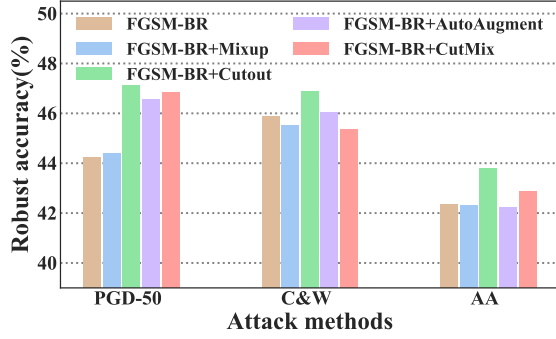


Fig. 4. The adversarial robustness performance of FAT methods with different data augmentation methods on CIFAR-10 under several attack scenarios.

B. Model Weight Averaging

As a technique to improve the model generalization, model weight averaging (WA) [53] obtains a WA model by computing exponentially weighted moving averages of model parameters of each training iteration. Formally, given a decay rate τ and the model parameters θ of an iteration, the WA model is updated with $\tilde{\theta} \leftarrow \tau \cdot \tilde{\theta} + (1 - \tau) \cdot \theta$. The obtained WA model $\tilde{\theta}$ at the end of training often generalizes better to the unseen test data. Recent work [45], [54] has indicated that the WA model shows higher robustness by achieving a flatter adversarial loss landscape. In particular, Wong *et al.* [29] show that the WA model obtained in SAT also shows higher adversarial robustness. However, the impact of WA on the fast adversarial training has not been explored in previous works. We make the first exploration by combining the FGSM-RS with the vanilla EM [28] and SE-EMA [29]. As shown in the Table IV, simply applying the existing WA to original fast adversarial training (FGSM-RS) does not improve the robustness.

C. The Proposed Model Weight Averaging

In this work, we attribute the failure of WA on original fast adversarial training (FGSM-RS) to the weak quality of adversarial examples generated by FGSM. The existing model weight averaging (WA) will accumulate model parameters of all iterations regardless of the quality of adversarial examples. However, as discussed in Sec. III-A, the gradient directions can be undesired when the quality of adversarial examples in fast adversarial training is low. The model parameter updates with the inaccurate gradients can be accumulated in WA model.

To overcome the shortcoming of the existing WA methods, we propose to exclude some training iterations from the WA model update trajectory. Namely, the WA model will not update in an iteration as long as the quality of adversarial examples in the iteration is low. Specifically, we propose a simple yet effective metric to automatically evaluate the quality of adversarial examples. The metric will be used to determine whether the WA model will be updated in the current training iteration. Concretely, the metric is defined as the ratio of the robust accuracy on adversarial examples to

TABLE IV
COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18 WITH DIFFERENT MODEL WEIGHT AVERAGING TECHNIQUES UNDER $\ell_\infty = 8/225$. NUMBER IN BOLD INDICATES THE BEST.

Method	Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Training Time (min)
FGSM-RS [18]	83.82	0.0	0.0	0.0	0.0	0.0	51
Vanilla EMA [28]	84.9	0.0	0.0	0.0	0.0	0.0	52
SE-EMA [29]	84.8	0.0	0.0	0.0	0.0	0.0	52
Auto-EMA(Ours)	83.62	51.26	50.21	49.9	49.42	46.53	52

the corresponding clean accuracy to evaluate the adversarial example quality. Formally, it can be formulated as:

$$\Delta = \frac{Acc(f(\mathbf{x}_b + \delta_b; \theta), \mathbf{y}_b)}{Acc(f(\mathbf{x}_b; \theta), \mathbf{y}_b)}, \quad (16)$$

where \mathbf{x}_b is the input samples within a batch, δ_b is the generated perturbations within a batch, \mathbf{y}_b is the corresponding labels within a batch and $Acc(\cdot)$ is the accuracy of trained model on training samples. When the proposed metric Δ is above a certain threshold T , the trained parameters θ is added to the average model parameters $\tilde{\theta}$, i.e., $\tilde{\theta} \leftarrow \tau \cdot \tilde{\theta} + (1 - \tau) \cdot \kappa \cdot \theta$, where $\kappa = 1(\Delta \leq T)$. In summary, our proposed method improves EMA by automatically selecting iterations to update WA model, called Auto-EMA. We conduct original FGSM-RS with our Auto-EMA and compare it with Vanilla EMA and SE-EMA. The result is shown in Table IV. It can be observed that EMA and SE-EMA also meets CO but the proposed Auto-EMA can prevent CO and achieves the best robustness.

TABLE V
TEST ROBUSTNESS (%) ON THE CIFAR-10 DATABASE USING RESNET18 WITH EQUIPPED WITH LIPSCHITZ REGULARIZATION WITH DIFFERENT HYPER-PARAMETER λ . NUMBER IN BOLD INDICATES THE BEST.

Hyper-Parameter	Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Training Time (min)
Baseline	86.82	46.62	44.76	44.23	45.89	42.37	51
$\lambda = 8$	81.61	56.31	55.56	55.48	50.83	48.77	81
$\lambda = 10$	81.91	56.73	56.03	55.9	51.32	49.12	81
$\lambda = 12$	81.34	57.07	56.57	56.36	51.4	49.22	81
$\lambda = 14$	81.40	57.05	56.31	56.11	51.08	48.71	81
$\lambda = 16$	81.25	56.98	56.25	56.15	51.03	48.96	81
$\lambda = 18$	80.93	57.06	56.45	56.31	51.25	49.10	81

TABLE VI
TEST ROBUSTNESS (%) ON THE CIFAR-10 DATABASE USING RESNET18 EQUIPPED WITH AUTO WEIGHT AVERAGING WITH DIFFERENT HYPER-PARAMETERS Δ . NUMBER IN BOLD INDICATES THE BEST.

Hyper-Parameter	Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Training Time (min)
Baseline	86.82	46.62	44.76	44.23	45.89	42.37	51
$\Delta = 0.8$	80.40	56.97	56.4	56.26	51.35	48.98	81
$\Delta = 0.82$	81.34	57.07	56.57	56.36	51.4	49.22	81
$\Delta = 0.84$	82.91	56.12	55.27	55.12	51.17	48.63	81
$\Delta = 0.86$	82.99	56.12	55.20	55.0	51.0	48.67	81
$\Delta = 0.88$	83.08	56.03	54.97	54.95	50.75	48.23	81
$\Delta = 0.9$	83.44	55.52	54.64	54.46	50.8	48.32	81

TABLE VII

COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10, CIFAR-100 AND TINY IMAGENET DATABASES USING DIFFERENT ADVERSARIAL TRAINING METHODS UNDER $\ell_\infty = 8/255$. NUMBER IN BOLD INDICATES THE BEST.

Dataset	CIFAR-10				Training Time(min)	CIFAR-100				Training Time(min)	Tiny ImageNet				Training Time(min)
Method	Clean	PGD-50	C&W	AA		Clean	PGD-50	C&W	AA		Clean	PGD-50	C&W	AA	
PGD-AT [58]	82.65	52.27	51.28	48.93	265	57.5	28.9	27.6	25.58	284	45.28	15.4	14.28	12.84	1833
FGSM-RS [18]	83.82	0.02	0.00	0.00	51	60.55	0.19	0.25	0.00	55	45.18	0.00	0.00	0.00	330
FGSM-PGI [25]	81.72	54.17	50.75	49.0	73	58.81	30.88	27.72	25.42	77	45.88	21.6	17.44	15.50	458
FGSM-CKPT [19]	90.29	39.15	41.13	37.15	76	60.93	15.24	16.6	14.34	84	49.98	8.68	9.24	8.10	495
FGSM-SDI [20]	85.25	51.79	50.29	47.91	83	60.82	30.08	27.3	25.19	99	47.64	19.16	16.02	14.10	565
NuAT [26] ¹	81.38	52.48	50.63	48.70	104	59.62	20.09	21.59	11.55	115	42.42	13.2	11.32	9.56	660
GAT [26]	80.41	51.76	49.07	46.56	109	56.07	23.0	21.93	19.51	119	41.84	13.8	11.48	9.74	663
FGSM-GA [27]	84.43	46.08	46.75	42.63	178	55.1	18.84	18.96	16.45	187	43.44	18.36	16.2	14.28	1054
Free-AT (m=8) [59]	80.75	44.48	43.73	41.17	215	52.63	22.16	20.68	18.57	229	40.06	8.2	8.08	7.34	1375
FGSM-LAW (ours)	81.36	56.36	51.4	49.22	81	58.79	31.33	28.47	25.78	86	47.74	24.06	18.65	16.43	514

TABLE VIII

TEST ROBUSTNESS (%) ON THE CIFAR-10 DATABASE USING WIDERESNET34-10 WITH DIFFERENT ADVERSARIAL TRAINING METHODS. NUMBER IN BOLD INDICATES THE BEST.

CIFAR-10	Clean	PGD-10	PGD-20	PGD-50	AA	Training Time (h)
PGD-AT [58]	85.17	56.1	55.07	54.87	51.67	31.9
FGSM-RS [18]	74.3	42.3	41.2	40.9	38.4	5.8
FGSM-PGI [25]	85.09	57.72	56.86	56.4	50.11	8.3
FGSM-CKPT [19]	91.84	44.7	42.72	42.25	40.46	8.7
FGSM-SDI [20]	86.4	55.89	54.95	54.6	51.17	9.4
NuAT [26]	85.30	55.8	54.68	53.75	50.06	11.8
GAT [23]	85.17	56.3	55.23	54.97	50.01	12.9
FGSM-GA [27]	82.1	48.9	47.1	46.9	45.7	20.3
Free-AT [59]	80.1	47.9	46.7	46.3	43.9	23.7
FGSM-LAW (Ours)	84.78	59.88	59.26	59.12	52.29	9.2

V. EXPERIMENTS

To evaluate the proposed FGSM-LAW, We adopt several datasets, *i.e.*, CIFAR-10 [55], CIFAR-100 [55], Tiny ImageNet [56] and ImageNet [56], which are widely used to evaluate the model robustness.

A. Default Training Setting

Following the default settings of AT [20], [57], we apply the same training hyper-parameters to conduct experiments. All models with ReLU activation function are trained for 110 epochs with the batch size 128. We use SGD momentum optimizer with the weight decay of 5×10^{-4} . The learning rate initialized to 0.1 decays with a factor of 0.1 at 100 and 105 epochs. All experiments are conducted under the ℓ_∞ distance metric with the maximal perturbation of $8/255$. We adopt PGD attack of 50-steps (**PGD-50**), C&W attack of 20 steps (**C&W**) and AutoAttack (**AA**) to evaluate the trained models. Note that we report the results of the last checkpoint, and we also report the the results of the best checkpoint with the best accuracy under PGD-10 attack in the Supplementary Material. All experiments are conducted on a single NVIDIA Tesla V100 for a fair comparison.

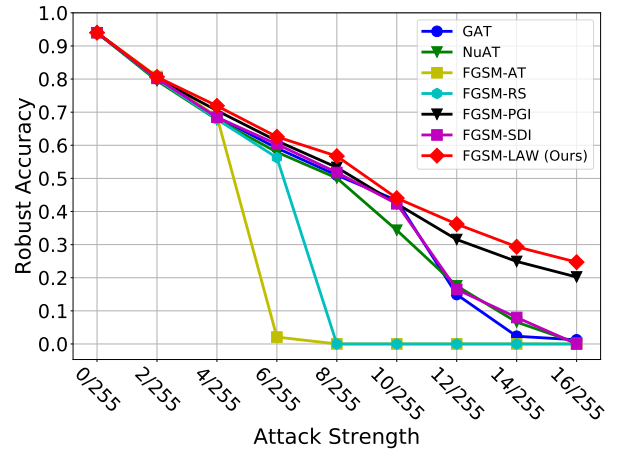


Fig. 5. Robust accuracy VS attack strength. The adversarial robustness performance of different advanced fast adversarial training methods for CIFAR-10 with the same training setting and evaluation attack strengths under PGD-50.

B. Detailed hyper-parameter settings

There are two hyper-parameters in the proposed FGSM-LAW, *i.e.*, the hyper-parameter of the Lipschitz regularization λ and the hyper-parameter of the auto weight averaging Δ . Using ResNet18 on CIFAR-10, we adopt the FGSM-BR combined with Lipschitz regularization and auto weight averaging to determine the optimal hyper-parameters, respectively. As for the hyper-parameter λ , the results are shown in Table V. It can be observed that when $\lambda = 12$, the proposed Lipschitz regularization can achieve the best model robustness under all attack scenarios. In detail, under AA attack, the proposed Lipschitz regularization achieves an accuracy of about 48%. Hence, the hyper-parameter λ of the proposed Lipschitz regularization is set to 12. As for the hyper-parameter Δ , the results are shown in Table VI. It can be observed that when $\Delta = 0.82$, the proposed auto weight averaging can achieve the best model robustness under C&W and AA attack scenarios. Considering

¹For a fair comparison, we adopt the standard ResNet18 to conduct experiments, but the original NuAT adopts ResNet18 with extra the ReLU layer at each block from the code they provide.

TABLE IX

COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18 WITH DIFFERENT ADVERSARIAL TRAINING METHODS BY USING A CYCLIC LEARNING RATE STRATEGY. NUMBER IN BOLD INDICATES THE BEST.

Method	Clean Best/Last	PGD-10 Best/Last	PGD-20 Best/Last	PGD-50 Best/Last	AA Best/Last	Training Time (min)
PGD-AT [58]	80.12/80.12	51.59/51.59	50.83/50.83	50.7/50.7	46.83/46.83	78
FGSM-RS [18]	83.75/83.75	48.05/48.05	46.47/46.47	46.11/46.11	42.92/42.92	15
FGSM-PGI [25]	80.68/80.68	52.48/53.48	51.69/51.69	51.5/51.5	46.65/46.65	22
FGSM-CKPT [19]	89.08/89.08	40.47/40.47	38.2/38.2	37.69/37.69	35.66/35.66	23
FGSM-SDI [20]	82.08/82.08	51.63/51.63	50.65/50.65	50.33/50.33	46.21/46.21	24
NuAT [26]	76.23/76.23	51.52/51.52	50.81/50.81	50.64/50.64	46.33/46.33	30
GAT [23]	81.91/81.91	50.43/50.43	49.82/49.82	49.62/49.62	45.24/45.24	33
FGSM-GA [27]	80.83/80.83	48.76/48.76	47.83/47.83	47.54/47.54	43.06/43.06	53
Free-AT (m=8) [59]	75.22/75.22	44.67/44.67	43.97/43.97	43.72/43.72	40.30/40.30	58
FGSM-LAW (ours)	80.44/80.44	53.54/53.54	52.91/52.91	52.73/52.73	46.93/46.93	24

TABLE X

TEST ROBUSTNESS (%) ON THE IMAGENET DATABASE USING RESNET50 WITH DIFFERENT ADVERSARIAL TRAINING METHODS. NUMBER IN BOLD INDICATES THE BEST.

Method	Clean	AA	Time (hour)
PGD-AT [38]	64.02	34.96	211.2
Free-AT [59]	59.96	28.58	127.7
FGSM-RS [18]	55.62	26.24	44.5
FGSM-LAW (Ours)	66.14	30.12	62.3

the clean accuracy on clean samples, the hyper-parameter Δ of the proposed auto weight averaging is set to 0.82.

C. Comparisons With State-of-the-Art Methods

To evaluate the proposed method, we compare the proposed method with several state-of-the-art fast adversarial training methods *i.e.*, Free-AT [59], FGSM-RS [18], FGSM-GA [27], GAT [23], FGSM-CKPT [19], NuAT [26], FGSM-SDI [20], FGSM-PGI [25] and an advanced SAT method (*i.e.*, PGD-AT [58]). Note that we adopt the training settings reported in their original works to train these AT models.

1) *Results on CIFAR-10*: As for CIFAR-10, we use ResNet18 as the backbone. The results are shown in Table VII. It can be observed that compared with the advanced PGD-AT, the proposed method achieves better model robustness under all attack scenarios and requires much less training time. Compared with other advanced fast adversarial training methods, as for the best and last checkpoints, the proposed method achieves the best model robustness under all attack scenarios. Specifically, under the PGD-50 attack, the previous most effective fast adversarial training methods achieve an accuracy of about 54%, while the proposed method achieves an accuracy of about 56%. In terms of the training cost, the proposed method requires the minimum training time except for FGSM-RS which meets Catastrophic Overfitting. The training time of the proposed method is the same as FGSM-SDI, which is about 1.3 times faster than GAT and NuAT,

2.2 times faster than FGSM-GA and 2.7 times faster than Free-AT. FGSM-PGI requires huge memory costs to store the historical adversarial perturbations. But our method requires the same memory cost as the other fast adversarial training methods. Different from previous fast adversarial training variants, the proposed method not only prevents Catastrophic Overfitting but also improves the model robustness without extra training time and memory cost. Moreover, following the default settings, we compare the proposed method with several advanced fast adversarial training methods under larger ℓ_∞ distance metric. As shown in Fig. 5, previous fast adversarial training methods could lead to Catastrophic Overfitting but the proposed method can prevent Catastrophic Overfitting and achieve significant adversarial robustness improvement.

Besides, we also adopt a larger architecture (WideResNet34-10) as the backbone to conduct comparative experiments on CIFAR-10. The results are shown in Table VIII. It can be observed that compared with PGD-AT, the proposed FGSM-LAW achieves better model robustness. Particularly, under AA attack, PGD-AT achieves an accuracy of about 51%, while the proposed FGSM-LAW achieves an accuracy of about 52%. Compared with previous fast adversarial training methods, the proposed method achieves the best robustness performance under all attack scenarios. In terms of the training cost, we observe similar phenomenons as ResNet18 used as the backbone.

Moreover, the comparative experiments are also implemented by using a cyclic learning rate strategy on CIFAR-10. Following the training settings ([18], [25]), the maximum learning rate of FGSM-GA [27] and FGSM-CKPT [19] is set to 0.3. The maximum learning rate of other fast adversarial training methods is set to 0.2. The results are shown in Table IX. It can be observed that compared with PGD-AT, the proposed FGSM-LAW achieves better model robustness performance under all adversarial attack scenarios and require less training cost. And compared with previous fast adversarial training methods, the proposed FGSM-LAW can achieve the best adversarial robustness under all adversarial attack scenarios on the best and last checkpoint. In terms of the training

TABLE XI
COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18. NUMBER IN BOLD INDICATES THE BEST.

	Method	Clean	PGD-10	PGD-20	PGD-50	C&W	AA	Training Time(min)
CIFAR-10	NuAT-WA [26]	82.21	54.76	54.07	53.87	52.07	50.54	95
	FGSM-LAW (ours)	81.8	58.13	57.6	57.47	52.46	50.64	81
CIFAR-100	NuAT-WA [26]	61.6	27.57	24.04	21.77	22.58	15.11	105
	FGSM-LAW (ours)	59.2	31.95	31.4	31.3	28.27	25.72	86

TABLE XII
COMPARISONS OF CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 OF THE BEST AND LAST CHECKPOINTS USING DIFFERENT ADVERSARIAL TRAINING WITH DIFFERENT MODEL WEIGHT AVERAGING METHODS UNDER $\ell_\infty = 8/225$. NUMBER IN BOLD INDICATES THE BEST. .

Method	Clean Best/Last	PGD-10 Best/Last	PGD-20 Best/Last	PGD-50 Best/Last	C&W Best/Last	AA Best/Last	Training Time (min)
FGSM-PGI [25]	81.72/81.72	55.18/55.18	54.36/54.36	54.17/54.17	50.75/50.75	49.0/49.0	73
FGSM-PGI-EMA [25]	80.6/82.3	55.73/55.18	55.02/54.48	55.0/54.14	50.86/50.74	49.23/49.08	74
FGSM-PGI-Auto-EMA (ours)	80.71/81.0	55.87/55.85	55.32/55.31	55.11/55.15	51.07/51.0	49.66/49.56	74

cost, we observe similar phenomenons as the models trained with the multi-step learning rate strategy.

2) *Results on CIFAR-100*: As for CIFAR-100, ResNet18 is used as the backbone network. The result is shown in Table VII. Compared with PGD-AT, the proposed method achieves better model robustness under most attack scenarios and costs less training time. Even under AA attack, PGD-AT obtains an accuracy of about 25.48%, but our method obtains an accuracy of about 25.78%. Our method achieves the best model robustness under all attack scenarios among the fast adversarial training methods. In terms of the training cost, similar results are observed on CIFAR-10.

3) *Results on Tiny ImageNet and ImageNet*: As for Tiny ImageNet, following the previous works [20], [25], [27] Pre-ActResNet18 is used as the backbone network. The result is shown in Table VII. Compared with PGD-AT, the proposed method achieves better adversarial robustness under all attack scenarios with less training time. In detail, under AA attack, PGD-AT achieves an accuracy of about 12.84%, but the proposed method achieves an accuracy of about 16.43%. And compared with previous fast adversarial training methods, the proposed method achieves the best model robustness under all attack scenarios. In terms of the training cost, similar results are observed on CIFAR-10 and CIFAR-100. Following the training settings [18], [59], we adopt ResNet50 as the backbone to conduct comparative experiments. The maximum perturbation strength ϵ is set to $4/255$. We compare the proposed FGSM-LAW with PGD-AT [38], Free-AT [59] and FGSM-RS [18]. The results are shown in Table X. Compared with PGD-AT, our FGSM-LAW achieves higher clean accuracy and comparable robustness under AA attack. Compared with Free-AT and FGSM-RS, the proposed FGSM-LAW achieves the best clean and robust accuracy.

4) *Comparisons with state-of-the-art robustness model*: **Comparisons With NuAT-WA**. NuAT-WA uses CyclicLR

to improve robustness, achieving state-of-the-art results in adversarial robustness. To conduct a fair comparison, we adopt the same training settings to conduct FGSM-LAW. The results are shown in Table XI. Compared to NuAT-WA, our proposed method demonstrates better adversarial robustness under all attack scenarios on both CIFAR-10 and CIFAR-100 datasets. Additionally, the proposed method incurs a lower training cost than NuAT-WA.

Comparisons With FGSM-PGI-WA. The proposed auto weight averaging method can be combined with FGSM-PGI to improve the adversarial robustness of the original FGSM-PGI. We adopt FGSM-PGI with the original weight averaging as the baseline. The results are shown in Table XII. It can be observed that FGDM-PGI combined with the proposed Auto-EMA can achieve the best adversarial robustness under all attack scenarios. In detail, under AA attack, FGSM-PGI-EMA achieves an accuracy of about 49.08%, while the proposed FGSM-PGI-Auto-EMA achieves an accuracy of about 49.56% with the same training time, which validates the effectiveness of the proposed method.

TABLE XIII
ABLATION STUDY ON CIFAR-10. CLEAN AND ROBUST ACCURACY (%) AND TRAINING TIME ON THE CIFAR-10 DATABASE USING RESNET18 UNDER $\ell_\infty = 8/225$. NUMBER IN BOLD INDICATES THE BEST.

L-Regular (Ours)	Cutout	A-WA (Ours)	Clean	PGD-50	C&W	AA	Training Time (min)
			86.82	44.23	45.89	42.37	51
✓			82.97	54.29	50.41	48.05	73
✓	✓		82.55	55.08	50.69	48.63	80
✓		✓	82.56	55.64	51.22	48.80	74
✓	✓	✓	81.34	56.36	51.4	49.22	81

D. Ablation Study

In the proposed method, we propose a Lipschitz regularization (L-regular) term and an auto weight averaging (A-WA)

with Cutout augmentation to conduct fast adversarial training. To validate the effectiveness of each element, we conduct an ablation study on CIFAR-10 by using ResNet18. The result is shown in Table XIII. We use the clean accuracy on clean images and the robust accuracy on different adversarial attack methods as the evaluation metric on the last checkpoint. Analyses are summarized as follows. First, only using the proposed regularization, the performance of the model robustness under all attack scenarios significantly improves. Incorporating our regularization with cutout augmentation, the performance of the model robustness further improves. Incorporating our regularization with the proposed weight averaging, the proposed method can better model robustness. Second, using all terms can obtain the best robust performance, which indicates that the three elements are compatible, and combining them can achieve the best performance.

VI. CONCLUSION

This work conducts a thorough analysis of existing works in the field of fast adversarial training in terms of adversarial robustness and training cost. We revisit the effectiveness and efficiency of the fast adversarial training techniques in preventing catastrophic overfitting. To further improve the performance of fast adversarial training, we propose a novel regularization method motivated by the Lipschitz constraint. Moreover, we explore the effect of data augmentation and weight averaging in fast adversarial training and propose an effective auto weight averaging method for fast adversarial training to improve robustness. By assembling these techniques, we conclude our fast adversarial training method equipped with the proposed Lipschitz regularization and our auto weight averaging, called FGSM-LAW. Extensive experimental evaluations demonstrate that the proposed method outperforms state-of-the-art fast adversarial training methods with less training costs.

REFERENCES

- [1] L. Deng, G. E. Hinton, and B. Kingsbury, "New types of deep neural network learning for speech recognition and related applications: an overview," in *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2013, Vancouver, BC, Canada, May 26-31, 2013*. IEEE, 2013, pp. 8599–8603. [Online]. Available: <https://doi.org/10.1109/ICASSP.2013.6639344>
- [2] T. N. Sainath, B. Kingsbury, A. Mohamed, G. E. Dahl, G. Saon, H. Soltau, T. Beran, A. Y. Aravkin, and B. Ramabhadran, "Improvements to deep convolutional neural networks for LVCSR," in *2013 IEEE Workshop on Automatic Speech Recognition and Understanding, Olomouc, Czech Republic, December 8-12, 2013*. IEEE, 2013, pp. 315–320. [Online]. Available: <https://doi.org/10.1109/ASRU.2013.6707749>
- [3] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," *Commun. ACM*, vol. 60, no. 6, pp. 84–90, 2017. [Online]. Available: <https://doi.org/10.1145/3065386>
- [4] Z. Zhao, P. Zheng, S. Xu, and X. Wu, "Object detection with deep learning: A review," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 30, no. 11, pp. 3212–3232, 2019. [Online]. Available: <https://doi.org/10.1109/TNNLS.2018.2876865>
- [5] Y. Dong, F. Liao, T. Pang, H. Su, J. Zhu, X. Hu, and J. Li, "Boosting adversarial attacks with momentum," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 9185–9193. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Dong_Boosting_Adversarial_Attacks_CVPR_2018_paper.html
- [6] Y. Fan, B. Wu, T. Li, Y. Zhang, M. Li, Z. Li, and Y. Yang, "Sparse adversarial attack via perturbation factorization," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12367. Springer, 2020, pp. 35–50. [Online]. Available: https://doi.org/10.1007/978-3-030-58542-6_3
- [7] X. Jia, X. Wei, X. Cao, and X. Han, "Adv-watermark: A novel watermark perturbation for adversarial examples," in *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, C. W. Chen, R. Cucchiara, X. Hua, G. Qi, E. Ricci, Z. Zhang, and R. Zimmermann, Eds. ACM, 2020, pp. 1579–1587. [Online]. Available: <https://doi.org/10.1145/3394171.3413976>
- [8] R. Duan, X. Ma, Y. Wang, J. Bailey, A. K. Qin, and Y. Yang, "Adversarial camouflage: Hiding physical-world attacks with natural styles," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 997–1005. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Duan_Adversarial_Camouflage_Hiding_Physical-World_Attacks_With_Natural_Styles_CVPR_2020_paper.html
- [9] X. Wang and K. He, "Enhancing the transferability of adversarial attacks through variance tuning," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*. Computer Vision Foundation / IEEE, 2021, pp. 1924–1933. [Online]. Available: https://openaccess.thecvf.com/content_CVPR2021/html/Wang_Enhancing_the_Transferability_of_Adversarial_Attacks_Through_Variance_Tuning_CVPR_2021_paper.html
- [10] Q. Cai, C. Liu, and D. Song, "Curriculum adversarial training," pp. 3740–3747, 2018. [Online]. Available: <https://doi.org/10.24963/ijcai.2018/520>
- [11] Y. Wang, D. Zou, J. Yi, J. Bailey, X. Ma, and Q. Gu, "Improving adversarial robustness requires revisiting misclassified examples," in *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. [Online]. Available: <https://openreview.net/forum?id=rklOg6EFwS>
- [12] Y. Wang, X. Ma, J. Bailey, J. Yi, B. Zhou, and Q. Gu, "On the convergence and robustness of adversarial training," vol. 97, pp. 6586–6595, 2019. [Online]. Available: <http://proceedings.mlr.press/v97/wang19i.html>
- [13] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. S. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 11278–11287. [Online]. Available: <http://proceedings.mlr.press/v119/zhang20z.html>
- [14] J. Cui, S. Liu, L. Wang, and J. Jia, "Learnable boundary guided adversarial training," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 15701–15710. [Online]. Available: <https://doi.org/10.1109/ICCV48922.2021.01543>
- [15] C. Sitawarin, S. Chakraborty, and D. A. Wagner, "SAT: improving adversarial training via curriculum-based loss smoothing," in *AISeC@CCS 2021: Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security, Virtual Event, Republic of Korea, 15 November 2021*, N. Carlini, A. Demontis, and Y. Chen, Eds. ACM, 2021, pp. 25–36. [Online]. Available: <https://doi.org/10.1145/3474369.3486878>
- [16] X. Mao, G. Qi, Y. Chen, X. Li, R. Duan, S. Ye, Y. He, and H. Xue, "Towards robust vision transformer," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 12032–12041. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01173>
- [17] F. Tramèr, A. Kurakin, N. Papernot, I. J. Goodfellow, D. Boneh, and P. D. McDaniel, "Ensemble adversarial training: Attacks and defenses," 2018. [Online]. Available: <https://openreview.net/forum?id=rkZvSe-RZ>
- [18] E. Wong, L. Rice, and J. Z. Kolter, "Fast is better than free: Revisiting adversarial training," 2020. [Online]. Available: <https://openreview.net/forum?id=BJx040EFvH>
- [19] H. Kim, W. Lee, and J. Lee, "Understanding catastrophic overfitting in single-step adversarial training," in *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*. AAAI Press, 2021, pp. 8119–8127. [Online]. Available: <https://ojs.aaai.org/index.php/AAAI/article/view/16989>

- [20] X. Jia, Y. Zhang, B. Wu, J. Wang, and X. Cao, "Boosting fast adversarial training with learnable adversarial initialization," *IEEE Trans. Image Process.*, vol. 31, pp. 4417–4430, 2022. [Online]. Available: <https://doi.org/10.1109/TIP.2022.3184255>
- [21] T. Li, Y. Wu, S. Chen, K. Fang, and X. Huang, "Subspace adversarial training," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 13 399–13 408. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01305>
- [22] P. de Jorge Aranda, A. Bibi, R. Volpi, A. Sanyal, P. H. S. Torr, G. Rogez, and P. K. Dokania, "Make some noise: Reliable and efficient single-step adversarial training," 2022. [Online]. Available: https://papers.nips.cc/paper_files/paper/2022/hash/5434a6b40f8f65488e722bc33d796c8b-Abstract-Conference.html
- [23] G. Sriraman, S. Addepalli, A. Baburaj, and V. B. R., "Guided adversarial attack for evaluating and enhancing adversarial defenses," 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/ea3ed20b6b101a09085ef09c97da1597-Abstract.html>
- [24] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash, "Efficient adversarial training with transferable adversarial examples," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2020, Seattle, WA, USA, June 13-19, 2020*. Computer Vision Foundation / IEEE, 2020, pp. 1178–1187. [Online]. Available: https://openaccess.thecvf.com/content_CVPR_2020/html/Zheng_Efficient_Adversarial_Training_With_Transferable_Adversarial_Examples_CVPR_2020_paper.html
- [25] X. Jia, Y. Zhang, X. Wei, B. Wu, K. Ma, J. Wang, and X. Cao, "Prior-guided adversarial initialization for fast adversarial training," vol. 13664, pp. 567–584, 2022. [Online]. Available: https://doi.org/10.1007/978-3-031-19772-7_33
- [26] G. Sriraman, S. Addepalli, A. Baburaj, and V. B. R., "Towards efficient and effective adversarial training," pp. 11821–11833, 2021. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/62889e73828c756c961c5a6d6c01a463-Abstract.html>
- [27] M. Andriushchenko and N. Flammarion, "Understanding and improving fast adversarial training," 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/b8ce47761ed7b3b6f48b583350b7f9e4-Abstract.html>
- [28] S. Rebuffi, S. Goyal, D. A. Calian, F. Stimberg, O. Wiles, and T. A. Mann, "Fixing data augmentation to improve adversarial robustness," *CoRR*, vol. abs/2103.01946, 2021. [Online]. Available: <https://arxiv.org/abs/2103.01946>
- [29] H. Wang and Y. Wang, "Self-ensemble adversarial training for improved robustness," 2022. [Online]. Available: <https://openreview.net/forum?id=oU3aTsmerQV>
- [30] T. Miyato, S. Maeda, M. Koyama, and S. Ishii, "Virtual adversarial training: A regularization method for supervised and semi-supervised learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 8, pp. 1979–1993, 2019. [Online]. Available: <https://doi.org/10.1109/TPAMI.2018.2858821>
- [31] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song, "Robust physical-world attacks on deep learning visual classification," in *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*. Computer Vision Foundation / IEEE Computer Society, 2018, pp. 1625–1634. [Online]. Available: http://openaccess.thecvf.com/content_cvpr_2018/html/Eykholt_Robust_Physical-World_Attacks_CVPR_2018_paper.html
- [32] J. Gu, H. Zhao, V. Tresp, and P. H. S. Torr, "Adversarial examples on segmentation models can be easy to transfer," *CoRR*, vol. abs/2111.11368, 2021. [Online]. Available: <https://arxiv.org/abs/2111.11368>
- [33] B. Yin, W. Wang, T. Yao, J. Guo, Z. Kong, S. Ding, J. Li, and C. Liu, "Adv-makeup: A new imperceptible and transferable attack on face recognition," pp. 1252–1258, 2021. [Online]. Available: <https://doi.org/10.24963/ijcai.2021/173>
- [34] Y. Hu, J. Chen, B. Kung, K. Hua, and D. S. Tan, "Naturalistic physical adversarial patch for object detectors," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 7828–7837. [Online]. Available: <https://doi.org/10.1109/ICCV48922.2021.00775>
- [35] R. Duan, Y. Chen, D. Niu, Y. Yang, A. K. Qin, and Y. He, "Advdrop: Adversarial attack to dnn by dropping information," in *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021, Montreal, QC, Canada, October 10-17, 2021*. IEEE, 2021, pp. 7486–7495. [Online]. Available: <https://doi.org/10.1109/ICCV48922.2021.00741>
- [36] Z. Hu, S. Huang, X. Zhu, F. Sun, B. Zhang, and X. Hu, "Adversarial texture for fooling person detectors in the physical world," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 13 297–13 306. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.01295>
- [37] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," 2015. [Online]. Available: <http://arxiv.org/abs/1412.6572>
- [38] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," 2018. [Online]. Available: <https://openreview.net/forum?id=rJzIBfZAb>
- [39] N. Carlini and D. A. Wagner, "Towards evaluating the robustness of neural networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*. IEEE Computer Society, 2017, pp. 39–57. [Online]. Available: <https://doi.org/10.1109/SP.2017.49>
- [40] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXIII*, ser. Lecture Notes in Computer Science, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds., vol. 12368. Springer, 2020, pp. 484–501. [Online]. Available: https://doi.org/10.1007/978-3-030-58592-1_29
- [41] F. Croce and M. Hein, "Minimally distorted adversarial examples with a fast adaptive boundary attack," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 2196–2205. [Online]. Available: <http://proceedings.mlr.press/v119/croce20a.html>
- [42] —, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 2206–2216. [Online]. Available: <http://proceedings.mlr.press/v119/croce20b.html>
- [43] T. Tsiligkaridis and J. Roberts, "Understanding and increasing efficiency of frank-wolfe adversarial training," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 2022, pp. 50–59. [Online]. Available: <https://doi.org/10.1109/CVPR52688.2022.00015>
- [44] Y. Zhang, G. Zhang, P. Khanduri, M. Hong, S. Chang, and S. Liu, "Revisiting and advancing fast adversarial training through the lens of bi-level optimization," in *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, ser. Proceedings of Machine Learning Research, K. Chaudhuri, S. Jegelka, L. Song, C. Szepesvári, G. Niu, and S. Sabato, Eds., vol. 162. PMLR, 2022, pp. 26 693–26 712. [Online]. Available: <https://proceedings.mlr.press/v162/zhang22ak.html>
- [45] D. Wu, S. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," 2020. [Online]. Available: <https://proceedings.neurips.cc/paper/2020/hash/1ef91c212e30e14bf125e9374262401f-Abstract.html>
- [46] T. Weng, H. Zhang, P. Chen, J. Yi, D. Su, Y. Gao, C. Hsieh, and L. Daniel, "Evaluating the robustness of neural networks: An extreme value theory approach," 2018. [Online]. Available: <https://openreview.net/forum?id=BkUHIMZ0b>
- [47] B. Wu, J. Chen, D. Cai, X. He, and Q. Gu, "Do wider neural networks really help adversarial robustness?" pp. 7054–7067, 2021. [Online]. Available: <https://proceedings.neurips.cc/paper/2021/hash/3937230de3c8041e4da6ac3246a888e8-Abstract.html>
- [48] S. Goyal, C. Qin, J. Uesato, T. A. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," *CoRR*, vol. abs/2010.03593, 2020. [Online]. Available: <https://arxiv.org/abs/2010.03593>
- [49] T. Devries and G. W. Taylor, "Improved regularization of convolutional neural networks with cutout," *CoRR*, vol. abs/1708.04552, 2017. [Online]. Available: <http://arxiv.org/abs/1708.04552>
- [50] H. Zhang, M. Cissé, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," 2018. [Online]. Available: <https://openreview.net/forum?id=r1Ddp1-Rb>
- [51] E. D. Cubuk, B. Zoph, D. Mané, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation policies from data," *CoRR*, vol. abs/1805.09501, 2018. [Online]. Available: <http://arxiv.org/abs/1805.09501>
- [52] P. Izmailov, D. Podoprikin, T. Garipov, D. P. Vetrov, and A. G. Wilson, "Averaging weights leads to wider optima and

- better generalization,” pp. 876–885, 2018. [Online]. Available: <http://auai.org/uai2018/proceedings/papers/313.pdf>
- [54] T. Chen, Z. Zhang, S. Liu, S. Chang, and Z. Wang, “Robust overfitting may be mitigated by properly learned smoothening,” in *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. [Online]. Available: <https://openreview.net/forum?id=qZzy5urZw9>
- [55] A. Krizhevsky, G. Hinton *et al.*, “Learning multiple layers of features from tiny images,” 2009.
- [56] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei, “Imagenet: A large-scale hierarchical image database,” in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), 20-25 June 2009, Miami, Florida, USA*. IEEE Computer Society, 2009, pp. 248–255. [Online]. Available: <https://doi.org/10.1109/CVPR.2009.5206848>
- [57] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, “Bag of tricks for adversarial training,” 2021. [Online]. Available: <https://openreview.net/forum?id=Xb8xvrtB8Ce>
- [58] L. Rice, E. Wong, and J. Z. Kolter, “Overfitting in adversarially robust deep learning,” in *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 8093–8104. [Online]. Available: <http://proceedings.mlr.press/v119/rice20a.html>
- [59] A. Shafahi, M. Najibi, A. Ghiasi, Z. Xu, J. P. Dickerson, C. Studer, L. S. Davis, G. Taylor, and T. Goldstein, “Adversarial training for free!” pp. 3353–3364, 2019. [Online]. Available: <https://proceedings.neurips.cc/paper/2019/hash/7503cfacd12053d309b6bed5c89de212-Abstract.html>