Towards Fairness-Aware Adversarial Learning

Yanghao Zhang   Tianle Zhang   Ronghui Mu   Xiaowei Huang   Wenjie Ruan*

University of Liverpool, UK

{yanghao.zhang, tianle.zhang, ronghui.mu, xiaowei.huang}@liverpool.ac.uk, w.ruan@trustai.uk

Abstract

Although adversarial training (AT) has proven effective in enhancing the model's robustness, the recently revealed issue of fairness in robustness has not been well addressed, i.e. the robust accuracy varies significantly among different categories. In this paper, instead of uniformly evaluating the model's average class performance, we delve into the issue of robust fairness, by considering the worst-case distribution across various classes. We propose a novel learning paradigm, named Fairness-Aware Adversarial Learning (FAAL). As a generalization of conventional AT, we redefine the problem of adversarial training as a min-max-max framework, to ensure both robustness and fairness of the trained model. Specifically, by taking advantage of distributional robust optimization, our method aims to find the worst distribution among different categories, and the solution is guaranteed to obtain the upper bound performance with high probability. In particular, FAAL can fine-tune an unfair robust model to be fair within only two epochs, without compromising the overall clean and robust accuracies. Extensive experiments on various image datasets validate the superior performance and efficiency of the proposed FAAL compared to other state-of-the-art methods.

1. Introduction

Deep learning models have undoubtedly achieved remarkable success across various domains, such as computer vision [31, 46] and natural language processing [41]. However, they still remain susceptible to deliberate adversarial manipulations of input data [15, 19, 20, 44, 48, 49]. Adversarial training techniques [11, 22, 23, 29, 37] have emerged as a potential solution, aiming to enhance a model's resilience against such vulnerabilities. These techniques have demonstrated a promising ability to enhance a model's overall robustness, yet the intricate connection between robustness and fairness, as revealed by [43], demonstrates that the robust accuracy of the models can vary considerably
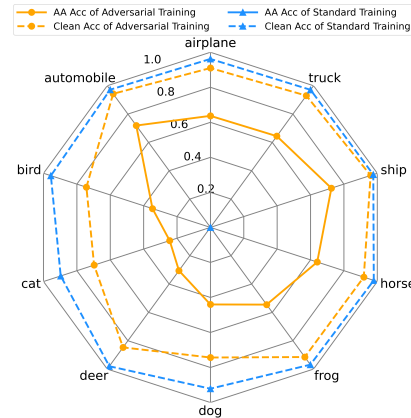
---
*Corresponding author

Figure 1. Class-wise accuracy of the Wide-ResNet34-10 model on CIFAR-10 dataset, where AA accuracy represents the robust accuracy against AutoAttack.

across different categories or classes. Consider a scenario where an autonomous driving system attains commendable average robust accuracy in recognizing road objects; despite this success, the system might demonstrate robustness against categories like inanimate objects (*with high accuracy*) while displaying vulnerability to *crucial* categories such as "human" (*with low accuracy*). This *disparity* or *unfair robustness* could potentially *endanger* drivers and pedestrians. Hence, it is vital to ensure consistent, equitable model performance against adversarial attacks by assessing worst-case robustness beyond average levels. This provides a more accurate evaluation than the average performance, recognizing the model's limitations while ensuring reliability across diverse categories in real-world applications.

Figure 1 provides an example, displaying the class-wise clean accuracy and robust accuracy against AutoAttack [12] using the Wide-ResNet34-10 [45] models on CIFAR-10 dataset, where both models are trained via standard training and adversarial training, respectively. It comes as no surprise that the model with standard training is vulnerable to adversarial attacks, yet it manages to attain comparable performance across various classes. In contrast, the adversarially trained model exhibits a noticeable *bias*, confidently classifying "*automobile*" but hesitating with "*cat*".

That is, the robust accuracy *diverges* significantly across classes. Furthermore, it is noteworthy that even though the class "*cat*" attains the lowest clean and robust accuracy, the most significant disparity between clean and robust accuracy arises in the case of the class "*deer*". This finding highlights an inconsistency between the clean and adversarial performances, where the robust accuracy on different classes illustrates more severe diverges in the model.

This phenomenon is called the "*robust fairness*" issue, which is first revealed in [43] referring to the gap between average robustness and worst-class robustness. Recently, some pioneering solutions have been proposed to address the robust fairness issue [26, 35, 40]. These efforts either tackle the robust fairness from the adversarial example generation or adjust the class weights empirically. However, essentially, most of these methods can be regarded as instances of *reweighting*, albeit diverse heuristic strategies are adopted. It is worth noting that such issues are absent in models trained without adversarial training, thus it is obvious that the underlying issue stems from the adversarial perturbations during the adversarial training. Inspired by but beyond these studies, we are prompted to consider why reweighting strategies are effective in mitigating the robust fairness issue. Intuitively, a model trained adversarially without reweighting fails to achieve high worst-class robust accuracy because it treats all classes equally, yet neglects the fact that the replaced adversarial examples may introduce bias to the final model. Conversely, reweighting can be perceived as a means of inducing a form of group distribution shift. This shift disrupts the uniform optimization of different classes, compelling the model to acquire resistance against these distribution shifts. This, in turn, leads to an enhancement in robust fairness. In line with this, we adopt an alternative approach to address this issue in the adversarial learning paradigm, where the following assumption is made:

> *The robust fairness issue in the conventional AT is due to the unknown group (class) distribution shift induced by the generated adversarial perturbations, which results in the overfitting problem.*

As opposed to the heuristic assignment for the reweighting item in the prior works, instead, we are expecting to leverage some optimization techniques for reweighting, such that may bring better results to resolve this overfitting problem. Hence Distributional Robust Optimization (DRO) [3, 13] naturally emerges as a viable choice. Rather than assuming a fixed uniform data distribution, DRO acknowledges the inherent distributional uncertainty in real-world data, offering a more resilient and adaptable model structure. Therefore, this paper delves into the exploration and adaption of DRO, as a sensible solution for the robust fairness challenge. Specifically, after finding the adversarial

example in adversarial training, instead of manually or empirically adjusting the weights for each class, we resort to learning the class-wise distributionally adversarial weights with the pre-defined constraints via DRO. By learning with these weights, the model will be guided to acquire the capacity to resist unknown group distribution shifts. The contributions of this paper are summarized as follows:

- We investigate the robust fairness issue from the perspective of group/class distributional shift, by taking the recent advances of Distributional Robust Optimization (DRO), which ultimately falls into a reweighting problem. To the best of our knowledge, this work is the first attempt to address the challenge of robust fairness through distributional robust optimization.
- We introduce a novel learning paradigm, named FAAL (**F**airness-**A**ware **A**dversarial **L**earning). This innovative approach extends the conventional min-max adversarial training framework into a *min-max-max* formulation. The intermediate maximization is dedicated to dealing with the robust fairness issue, by learning with the class-wise distributionally adversarial weights.
- Comprehensive experiments are conducted on CIFAR-10 and CIFAR-100 datasets across different models. We empirically validate that the proposed method is able to fine-tune a robust model with intensive bias into a model with *both* fairness and robustness within only *two* epochs.

2. Related Work

2.1. Robust Fairness

It is noted that in traditional machine learning, the definition of *fairness* [1, 16] might be different from the *robust fairness* we want to tackle, where the focus of this paper is on mitigating the fairness issue under the scenario against adversarial attacks.

Several works [26, 28, 35, 40, 43] are explored to alleviate the fairness issue in the robustness. Xu *et al*. [43] firstly revealed that the issue of robust fairness occurs in conventional adversarial training, which can introduce severe disparity of accuracy and robustness between different groups of data when boosts the average robustness. To mitigate this problem, they proposed a Fair-Robust-Learning (FRL) framework, by employing reweight and remargin strategies to finetune the pre-trained model, it is able to reduce the significant boundary error in a certain margin. Ma *et al*. [28] empirically discovered that the trade-off between robustness and robustness fairness exists and AT with a larger perturbation radius will result in a larger variance. To mitigate the trade-off between robustness and fairness, they add a variance regularization term into the objective function, named FAT, which relieves the trade-off between average robustness and robust fairness. Sun *et al*. [35] proposed a method called Balance Adversarial Training (BAT), which

adjusts the attack strengths and difficulties of each class to generate samples near the decision boundary for easier and fairer model learning. Wei *et al*. [40] presented a framework named CFA, which customizes specific training configurations for each class automatically according to which customizes specific training configurations, such that improving the worst-class robustness while maintaining the average performance. More recently, Li and Liu [26] considered the worst-class robust risk, where they proposed a framework named WAT (worst-class adversarial training) and leverage no-regret dynamics to solve this problem.

2.2. Distributional Robust Optimization

The origins of DRO are found in the early studies on robust optimization [2, 4, 7], which eventually led to the development of DRO as a tool for handling distributional uncertainties. The application of DRO to machine learning problems has garnered significant attention like domain generalization [32], data distribution shift [27, 33], adversarial robustness [5, 8, 34] and traditional fairness in machine learning [14, 16, 24, 36]. While the intersection of DRO with robustness and fairness has begun to receive attention, there remain gaps in the literature, particularly in understanding how DRO can address this fairness in an adversarial setting.

3. Methodology

3.1. Preliminaries

Given the training data drawn from a distribution P, when it comes to predicting labels $y \in \mathcal{Y}$ based on input features $x \in \mathcal{X}$, within a model family denoted as Θ, and utilizing a loss function $\ell := \Theta \times (\mathcal{X} \times \mathcal{Y}) \to \mathbb{R}$, the conventional training approach for achieving this objective is precisely what's known as empirical risk minimization (ERM):

$$\min_{\theta} \mathbb{E}_{(x,y)\sim P} \ \ell(f_\theta(x), y) \tag{1}$$

In traditional adversarial training, the focus is on identifying the worst-case perturbation for each input. This is formulated as a min-max problem, as defined in [29]. Mathematically, it can be expressed as follows:

$$\min_{\theta} \mathbb{E}_{(x,y)\sim P} \max_{\delta \in B_\epsilon} \ \ell(f_\theta(x + \delta), y) \tag{2}$$

When accounting for class fairness, specifically the performance across different classes, Eq. (2) can be rewritten as:

$$\min_{\theta} \frac{1}{C} \sum_{c=1}^{C} \mathbb{E}_{(x,y)\sim P_c} \max_{\delta \in B_\epsilon} \ \ell(f_\theta(x + \delta), y) \tag{3}$$

It is noted that when a batch contains an equal number of data points for each class, Eq. (3) is technically identical to Eq. (2). According to the definition of distributional robust optimization [3, 13], we now consider minimizing the expected loss in the worst-case scenario over a set of uncertain distributions. This can be mathematically expressed as:

$$\min_{\theta} \sup_{Q \in \mathcal{Q}} \mathbb{E}_{(x,y)\sim Q} \max_{\delta \in B_\epsilon} \ \ell(f_\theta(x + \delta), y) \tag{4}$$

where $\mathcal{Q}$ represents the uncertainty set, encompassing the range of potential test distributions for which we seek the model to exhibit commendable performance aligned with the data distribution P.

To establish the connection between robust fairness and DRO, we can naturally delineate the classes as distinct groups within the training data. Subsequently, the uncertainty set $\mathcal{Q}$ can be defined with respect to these groups. Specifically, the setting of group DRO are borrowed [18, 30, 32], where the training distribution P is assumed to be a mixture of C groups (classes) P_C indexed by $c = \{1, 2, ..., C\}$. Thus the uncertainty set $\mathcal{Q}$ is defined as any mixture of these classes, *i.e.* $\mathcal{Q} := \{\sum_{c=1}^{C} q_c P_c : \boldsymbol{q} \in \Delta_C\}$, where Δ_C is the probability simplex. Hence, the worst-case risk can be reformulated as the most detrimental combination across different groups, taking into account the expected loss for each class:

$$\min_{\theta} \sup_{\boldsymbol{q} \in \Delta_C} \sum_{c=1}^{C} q_c \cdot \mathbb{E}_{(x,y)\sim P_c} \max_{\delta \in B_\epsilon} \ \ell(f_\theta(x + \delta), y) \tag{5}$$

However, as proven in [18], applying DRO directly to robust learning training is overly pessimistic, which often yields results that do not surpass those achieved by a classifier adversarially trained using ERM. This outcome can be attributed to the specific classification loss function and the distributions that DRO seeks to encompass for the purpose of robustness are notably extensive. A similar pattern of failure is also encountered in the context of group DRO [32], and they advocate that sufficient regularization is required for over-parameterized neural networks to enhance worst-group generalization.

3.2. Problem Definition

By disentangling the Eq. (5), it becomes evident that it can be interpreted as a *reweighting* objective of the ERM framework within the context of adversarial settings, incorporating with the weighted factor $\boldsymbol{q}$. Empirical evidence has demonstrated the efficacy of several reweighting methods [6, 40, 43] in improving robust fairness and all of them fall into the same paradigm of Eq. (5). This aligns with the assumption stated in the introduction section, while we seek to leverage some optimization techniques for promoting fairness directly, rather than a heuristic assignment.

As previously mentioned, the extensive range of distributions encompassed by the uncertainty set $\mathcal{Q}$ could present difficulties for DRO in sustaining its robustness. Nevertheless, the pivotal factor in addressing the group DRO

lies in configuring the uncertainty set. To tackle this issue, we advocate an alternative solution: instead of relying solely on substantial regularization [32], we propose to use a straightforward yet effective ambiguity set with extra constraint. This is achieved by defining the ambiguity set as $\mathcal{Q}' := \{\sum_{c=1}^{C} q_c P_c : d(\mathcal{U}, \boldsymbol{q}) \le \tau, \boldsymbol{q} \in \Delta_C\}$, where $d(\cdot, \cdot)$ represents some distance metrics measuring the difference between two distributions, τ is the constraint parameter and $\mathcal{U}$ is the uniform distribution. This choice of $\mathcal{Q}'$ shrinks the width of $\mathcal{Q}$ and allows us to learn models that are robust to some group shifts, rather than identically uniform distribution among different classes. Hence, our final objective in addressing the challenge posed by potential group distribution shifts within an adversarial setting can be denoted as:

$$\min_{\theta} \max_{d(\mathcal{U}, \boldsymbol{q}) \le \tau, \boldsymbol{q} \in \Delta_C} \sum_{c=1}^{C} q_c \cdot \mathbb{E}_{(x,y) \sim P_c} \max_{\delta \in B_\epsilon} \ell(f_\theta(x + \delta), y)$$

$$(6)$$

Since the robust fairness issue occurs in general adversarial training, we also do not know the real class distribution shift may occur at test time, especially under adversarial training. The uncertainty set $\mathcal{Q}'$ encodes the possible test distributions that we want our model to perform well on. Therefore, a suitable divergence ball around the class distribution confers robustness to a set of distributional shifts. In our settings, we use KL divergence as d in our following experiments, as we will see the KL-DRO [5] has its unique property which provides the optimal solution for handling the fairness issue. In other words, such an overall objective will optimize the worse distribution of the neighborhood around the uniform distribution for different classes by learning those adversarial examples.

3.3. Fairness-Aware Adversarial Learning

Based on the above objective, we propose a novel adversarial learning paradigm, named **F**airness-**A**ware **A**dversarial **L**earning (FAAL)[1], to improve the robust fairness via distributional robust optimization. Specifically, within the intermediary stage of the conventional adversarial training, *i.e.* between inner maximization and outer minimization, we introduce a class-wise distributionally adversarial weight for orientating the learning directions among different categories, which can be optimally solved by leveraging distributional robust optimization. By incorporating this weight into the outer minimization process to update the model's parameters, the class (group) distribution shift can be protected to alleviate the robust fairness issue.

To provide a clearer illustration of the whole learning problem, we break down it into three distinct stages:
- *Phase 1*: Inner maximization for finding adversarial examples;

---

[1]Our code is available at https://github.com/TrustAI/FAAL

Algorithm 1 Fairness-Aware Adversarial Learning
___
Input: Training set $\{X, Y\}$, total epochs T, adversarial radius ϵ, step size α, the number of adversarial iteration K, model f parameterized by θ, the number of mini-batches M, batch size B, distribution shift constraint τ
Output: A robust and fair model
___
1: **for** $t = 1 \dots T$ **do**
2:    **for** $i = 1 \dots M$ **do**
3:      # *Phase 1: Inner maximization*
       $\delta = 0$
4:      **for** $j = 1 \dots K$ **do**
5:        $\delta = \delta + \alpha \cdot \text{sign}(\nabla_\delta \ell_{\text{CE}}(f_\theta(x_i + \delta), y_i))$
6:        $\delta = \max(\min(\delta, \epsilon), -\epsilon)$
7:      **end for**
8:      $x_i^{adv} = \text{clip}(x_i + \delta, 0, 1)$
9:      # *Phase 2: Intermediate maximization*
       $\ell_i = \ell_{\text{CE}}(f_\theta(x_i^{adv}), y_i, \text{reduction} = \text{'none'})$
       # Calculate the cross-entropy loss for each instance
10:      **for** $c = 1 \dots C$ **do**
11:        $\ell'_c = \ell_{\text{CW}}(f_\theta(x_i^{adv}, y_i)[y_i = c]$
       # Calculate the average margin for each class c
12:      **end for**
13:      $\boldsymbol{w}_*^{cda} = \text{solve\_kl\_dro}(\boldsymbol{\ell}', \tau)$
       # Solve the optimal class-wise weights for the current batch under the worst distribution via DRO
14:      $\mathcal{L}_{\text{FAAL}} = \frac{1}{B} \sum_{i=1}^{B} \boldsymbol{w}_*^{cda}[y] \cdot \ell_i \cdot C$
15:      # *Phase 3: Outer Minimization*
       $\theta = \theta - \nabla_\theta \mathcal{L}_{\text{FAAL}}$
16:    **end for**
17: **end for**
18: **return** Robust model f_θ with high fairness
___

- *Phase 2*: Intermediate maximization for finding the distributionally adversarial weight (worst-case distribution around the uniform distribution);
- *Phase 3*: Outer minimization for updating model's parameters.

Phases 1 and *3* are the classic processes of conventional adversarial training, and *Phase 2* is the core element of our proposed learning paradigm, as we assume that tackling the unknown class distributional shift can contribute to enhancing robust fairness. The whole procedure is summarized in Algorithm 1. In the following content, we replace the notion of $\boldsymbol{q}$ in Eq. (6) with $\boldsymbol{w}^{cda}$ for convenience and define it as Class-wise Distributionally Adversarial Weight.

Definition 1 (CDAW: Class-wise Distributionally Adversarial Weight). *Given a class-wise objective loss $\ell'_c \in \mathbb{R}$ on the adversarial examples, for all classes $c \in C$, the optimal Class-wise Distributionally Adversarial Weight vector $\boldsymbol{w}_*^{cda}$ aims to maximize the overall loss:*

$$\mathcal{L}_{\text{FAAL}} := \max \sum_{c=1}^{C} w_c^{cda} \ell_c' \tag{7}$$

$$s.t. \quad d(\mathcal{U}, \boldsymbol{w}^{cda}) \leq \tau, \boldsymbol{w}^{cda} \in \Delta_C$$

$$\boldsymbol{w}_*^{cda} := \arg \max \mathcal{L}_{\text{FAAL}} \tag{8}$$

In the case of a reweighting strategy employed to address robust fairness, it dictates the learning trajectory for each individual class. By optimizing the model using this optimized weight, the model is exposed to learning from the worst-case distribution under the pre-defined constraint τ, such that fairness among different classes will be encouraged. When $\boldsymbol{w}^{cda}$ is identical to $\mathcal{U}$, *i.e.* $\tau = 0$, it reduces the regular mean calculation for the overall loss, making the entire learning paradigm regress to conventional adversarial training that contains *Phases 1* and *3* only. We use KL divergence as d in the intermediate maximization, such that it can be solved via the *conic convex optimization*, and another elegant property of it on the generalization can be obtained, as demonstrated in Theorem 1 below.

Theorem 1. *Given the loss* $\mathcal{L}_{\text{FAAL}}$ *in Eq.* (7) *on the observed distribution, and suppose the regular loss* $\mathcal{L} =: \frac{1}{C}\sum_{c=1}^{C} \ell_c'$ *on the test distribution with unknown group distribution shift, then the following holds for all* $\boldsymbol{w}^{cda} \in \Delta_C$:

$$\Pr(\mathcal{L}_{\text{FAAL}} > \mathcal{L}) \geq 1 - e^{-\tau n + O(n)} \tag{9}$$

Where $\text{KL}(\mathcal{U}, \boldsymbol{w}^{cda}) \leq \tau$, $\mathcal{U}$ *is the uniform distribution.*

Theorem 1 tells that the $\mathcal{L}_{\text{FAAL}}$ is guaranteed to be the upper bound of $\mathcal{L}$ with high probability given the large number of observed sample n. In line with this, it enjoys a strong generalization where the performance on the test distribution with some unknown group distribution shift is at least as good as the estimated performance with high probability. So the solution of the class-wise distributionally adversarial weight solving by the convex optimization is optimal and will provide protection on the unknown class distribution shift. As cross-entropy loss cannot well-represents how good the discrepancy between classes [9], we instead use the CW margin loss [10] as $\mathcal{L}$ for calculating the class-wise distributionally adversarial weight:

$$\ell_c' := \mathbb{E}_{(x,y) \sim P_c} \left(\max_{j != y} z_j - z_y \right) \tag{10}$$

where z_j is the probability of the class j, *i.e.* the softmax output of the network. It is noted that the objective functions in *Phases 1-3* of our learning paradigm are not necessarily consistent, so the proposed learning mechanism is flexible and can be combined with any min-max adversarial approaches. In the following experiments, we will mainly solve the distributional robust optimization on the bounded margin loss among classes, which provides better performance than using cross-entropy loss in the intermediate maximization. More details can be found in the Appendix.

4. Experimental Results

Given our method's focus on the robust fairness challenge, it is reasonable to assume that the model already possesses a certain degree of average robustness. Otherwise, considering the robust fairness issue might not yield meaningful results. Hence, the question arises: *Is it imperative to initiate the training of a model from the beginning for achieving fairness with a certain robustness level?* In the next section, we first test our approach through adversarial fine-tuning, and then explore if training from scratch with our method can offer additional benefits.

4.1. Fine-tuning for Enhancing Robust Fairness

Baselines and experiment settings: We first conducted experiments on CIFAR-10 dataset [25], which is popularly used for adversarial training evaluation. We use the average & worst-class accuracy under different adversarial attacks (Clean / PGD [29] / CW [10] / AutoAttack [12]) as the evaluation metrics. The perturbation budget is set to $\epsilon = 8/255$ on CIFAR-10 dataset. FRL [43] is the only existing state-of-the-art technique from the recent literature which performs fine-tuning to a pre-trained model for improving robust fairness. FRL proposed two strategies based on TRADES [47] for enhancing robust fairness, including reweight (RW) and remargin (RM). Hence we apply the best versions of FRL from their paper: FRL-RWRM with $\tau_1 = \tau_2 = 0.05$ and FRL-RWRM with $\tau_1 = \tau_2 = 0.07$, where τ_1 and τ_2 are the fairness constraint parameters for reweight and remargin of FRL, we name them FRL-RWRM$_{0.05}$ and FRL-RWRM$_{0.07}$ for short. The results of FRL are reproduced using their public code, where the target models are fine-tuned for 80 epochs and the best results are presented.

In terms of the proposed method, although we utilize the PGD-AT adversary method by default (named FAAL$_{\text{AT}}$ for short), it is completely compatible with other AT approaches like TRADES or MART. To achieve this, one just needs to keep the original implementation for both inner maximization and outer minimization unchanged and add the intermediate maximization independently. We found that 2 epochs of fine-tuning are enough to improve the robust fairness greatly, without sacrificing too much average clean/robust accuracy. We set the value of τ in our method as 0.5, and the learning rate is configured from 0.01 in the first epoch and drops to 0.001 in the second epoch.

Table 1 demonstrates our main results of finetuning Wide-Resnet34-10 (WRN-34-10) models [45] on CIFAR-10 dataset, where different state-of-the-art adversarial defensed methods are adopted, including PGD-AT [29], TRADES [47], MART [39] and AWP [42]. We can see that FAAL$_{\text{AT}}$ outperforms the two FRL methods with respect to both average and worst-class robustness. Notably, in the majority of cases, FAAL$_{\text{AT}}$ achieves this without significant compromises in clean accuracy, unlike FRL meth-

Table 1. Evaluation of different fine-tuning methods on CIFAR-10 dataset using WRN-34-10 model. The best result is highlighted in **Bold**.

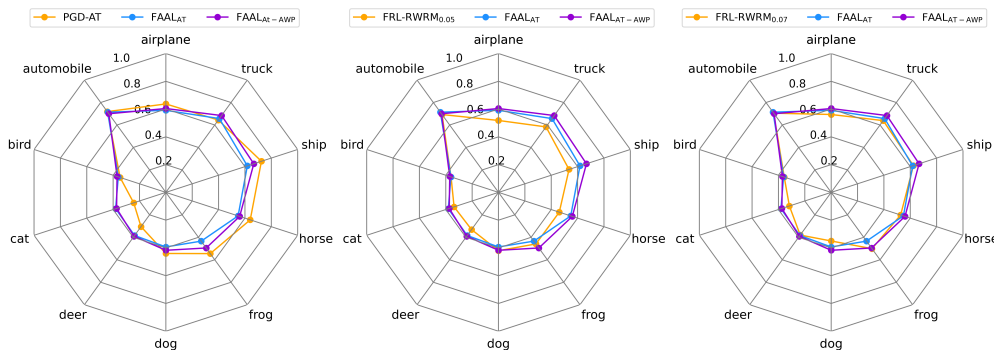| Adversarially Trained WRN-34-10 Model	Fine-Tuning Epochs	Average Accuracy (Worst-class Accuracy) (%)			
		Clean	PGD-20	CW-20	AutoAttack
PGD-AT	-	86.07 (69.70)	55.90 (29.90)	54.29 (28.30)	52.46 (24.40)
+ Fine-tune with FRL-RWRM$_{0.05}$	80	83.25 (**74.80**)	50.37 (38.10)	49.77 (36.60)	46.97 (33.10)
+ Fine-tune with FRL-RWRM$_{0.07}$	80	85.12 (71.60)	52.56 (37.10)	51.92 (35.50)	49.60 (31.70)
+ Fine-tune with FAAL$_{AT}$	2	**86.23** (69.70)	54.00 (37.60)	53.11 (36.90)	50.81 (35.70)
+ Fine-tune with FAAL$_{AT-AWP}$	2	85.47 (69.40)	**56.46** (**39.20**)	**54.50** (**38.10**)	**52.47** (**36.90**)
TRADES	-	84.92 (67.00)	55.32 (27.10)	53.92 (24.80)	**52.51** (23.20)
+ Fine-tune with FRL-RWRM$_{0.05}$	80	82.90 (72.70)	53.16 (40.60)	51.39 (36.30)	49.97 (**35.40**)
+ Fine-tune with FRL-RWRM$_{0.07}$	80	85.19 (70.90)	53.76 (39.20)	52.92 (36.80)	51.30 (34.60)
+ Fine-tune with FAAL$_{AT}$	2	**85.96** (**75.00**)	53.46 (39.80)	52.72 (38.20)	50.91 (35.30)
+ Fine-tune with FAAL$_{AT-AWP}$	2	85.39 (72.90)	**56.07** (**43.30**)	**54.16** (**38.60**)	52.45 (**35.40**)
MART	-	83.62 (67.90)	**56.22** (32.50)	**52.79** (25.70)	**50.95** (22.00)
+ Fine-tune with FRL-RWRM$_{0.05}$	80	**83.72** (**71.80**)	52.16 (37.50)	50.73 (35.00)	49.19 (31.70)
+ Fine-tune with FRL-RWRM$_{0.07}$	80	82.09 (**71.80**)	50.86 (36.00)	49.78 (33.00)	47.78 (30.30)
+ Fine-tune with FAAL$_{AT}$	2	83.49 (68.00)	51.65 (37.80)	50.36 (37.10)	48.63 (34.00)
+ Fine-tune with FAAL$_{AT-AWP}$	2	82.17 (64.00)	54.31 (**39.50**)	51.72 (**37.70**)	50.31 (**36.40**)
TRADES-AWP	-	85.35 (67.90)	59.20 (28.80)	**57.14** (26.50)	**56.18** (25.80)
+ Fine-tune with FRL-RWRM$_{0.05}$	80	82.31 (65.90)	49.90 (31.70)	49.68 (34.00)	46.50 (27.70)
+ Fine-tune with FRL-RWRM$_{0.07}$	80	84.24 (65.70)	48.63 (30.90)	49.77 (31.50)	46.53 (28.60)
+ Fine-tune with FAAL$_{AT}$	2	87.02 (**76.30**)	52.54 (35.00)	51.70 (34.40)	49.87 (30.60)
+ Fine-tune with FAAL$_{AT-AWP}$	2	**86.75** (74.80)	57.14 (**43.40**)	55.34 (**40.10**)	53.93 (**37.00**)

Figure 2. Class-wise robust accuracy against AutoAttack after fine-tuning the PGD adversarially trained WRN model

ods which tend to trade off the clean accuracy for improving robustness. FAAL$_{AT}$ promotes the worst-class AutoAttack (AA) accuracy by approximately 2.6% than FRL for fine-tuning PGD-AT, MART, and AWP. Except for fine-tuning TRADES, both methods yield comparable performance, this is partially due to that FRL is a TRADES-based method and it takes advantage of knowing the source method. Most importantly, FRL requires many epochs (80 epochs) to obtain the best results, while our method, is able to achieve better results within **only 2 epochs**. As adversarially training a large model with high robustness is already time-consuming, to circumvent the need for retraining the model from the beginning, FAAL offers a solution for saving time and computational resources. It demonstrates the capability to quickly fine-tune a robust model that initially lacks fairness, resulting in a model that is both robust and fair. Due to the space limit, similar improvements on the Preact-Resnet model can be found in the Appendix.

Strong adversarial attacks can help? The remargin of FRL [43] claims that increasing the perturbation margin can help for obtaining better robust fairness, while this may hurt the average clean accuracy/robustness, as indicated in Tab. 1. Certainly, there is a *trade-off* existing between the average robustness and worst-class robustness, but is it necessary to increase the perturbation margin ϵ for improving the class-wise robustness? We question whether this is a mandatory requirement for improvements, and we assume the benefits come from the stronger strength

of adversarial perturbation. Hence, instead of enlarging the perturbation margin, we capitalize on the flexibility of our learning framework and integrate our method with AWP [42], a well-regarded model weight perturbation technique, to strengthen the attacks. As shown in Tab. 1, when combining with AWP, $FAAL_{AT-AWP}$ further enhances the worst-class robust accuracy on WRN34-10 models, especially for the original one adversarially trained with AWP. $FAAL_{AT-AWP}$ is almost unharmed on the improvement to the original unfair models most of the time. Therefore, it is not compulsory to enlarge the perturbation margin to gain better results, where applying a stronger adversary indeed benefits robust fairness without enlarging the perturbation margin. Figure 2 visualizes the results of class-wise AA accuracy for the comparison of the proposed method $FAAL_{AT}$ and $FAAL_{AT-AWP}$, and two FRL baselines, respectively. It can be seen that FAAL boost the worst-class robust accuracy, presenting outstanding capacity to improve robust fairness with high effectiveness and efficiency, respectively, where it outperforms FRL not only for the average/worst-class robustness but also for the very rare fine-tuning steps.

Table 2. Training from scratch with different methods on CIFAR-10 dataset using Preact-ResNet18 model.

Adversarially Trained PRN-18 Model	Average Acc (Worst-class Acc) (%)	
	Clean	AutoAttack
PGD-AT	**82.72** (55.80)	47.38 (12.90)
TRADES	82.54 (66.10)	49.05 (20.70)
CFA_{AT}	80.82 (64.60)	**50.10** (24.40)
CFA_{TRADES}	80.36 (66.20)	**50.10** (26.50)
WAT_{TRADES}	80.37 (66.00)	46.16 (30.70)
$FAAL_{AT}$	82.20 (62.90)	49.10 (**33.70**)
$FAAL_{TRADES}$	81.62 (**68.90**)	48.48 (33.60)

4.2. Training from Scratch for Enhancing Fairness

Previous sections demonstrate the effectiveness and efficiency of the proposed approaches. Here we also investigate the advancements by training the model from the ground up using our method. We compare our methods with two common adversarial training methods (PGD-AT [29] and TRADES [47]) and two recent state-of-the-art techniques: CFA [40] and WAT [26], which have been proposed to mitigate the robust fairness issues recently. We adversarially trained Preact-ResNet-18 models [17] for 200 epochs with a learning rate of 0.1, which will be decayed by a factor of 0.1 at 100 and 150 epochs, successively. We start to facilitate the proposed intermediate maximization (see Algorithm 1 lines 9-14) after the 100-th epoch with the only hyperparameter τ from 0.25 and enlarge it to 0.5 after the 150-th epoch. In addition, similar to CFA using weight average,

we also applied EMA [21, 38], to gain a more stable performance, however, we only applied it after the 100-th epoch, where we start to apply the intermediate maximization. We report the best results under AutoAttack on the average accuracy and worst-class accuracy in Tab. 2. Besides, Fig. 3 visualizes the results of different training approaches including the proposed $FAAL_{TRADES}$ with other 3 TRADES-based models: TRADES, CFA_{TRADES} and WAT_{TRADES} respectively. We can observe that FAAL outperforms other approaches on the worst-class clean/robust accuracy, with less sacrifice on the average robustness.

Table 3. Result comparison of different methods on CIFAR-100 dataset using ResNet18 model.

Adversarially Trained RN-18 Model	Average Acc (Worst-class Acc) (%)	
	Clean	AutoAttack
TRADES	54.57 (19.00)	23.57 (1.00)
+ Fine-tune with $FRL\text{-}RWRM_{0.05}$	52.55 (22.00)	21.11 (2.00)
+ Fine-tune with $FAAL_{AT}$	**58.50** (21.00)	21.91 (2.00)
+ Fine-tune with $FAAL_{AT-AWP}$	58.41 (19.00)	23.44 (2.00)
+ Fine-tune with $FAAL_{TRADES}$	54.96 (18.00)	22.71 (2.00)
+ Fine-tune with $FAAL_{TRADES-AWP}$	54.90 (18.00)	23.25 (2.00)
CFA_{TRADES}	55.57 (**23.00**)	**24.56** (2.00)
WAT_{TRADES}	53.99 (19.00)	22.89 (**3.00**)
$FAAL_{AT}$	56.84 (16.00)	21.85 (**3.00**)
$FAAL_{TRADES}$	55.87 (21.00)	23.57 (**3.00**)

4.3. Additional Experiments on CIFAR-100 dataset

The experiments above mainly focused on CIFAR-10 dataset, which only has 10 classes in the dataset. In this section, we explore the proposed FAAL into a more challenging dataset, i.e. CIFAR-100 with 100 categories. Similarly, we reported the results of the average/worst clean accuracy and AutoAttack accuracy. The value of τ in our method is set to 0.05. For fine-tuning, we compare our proposed method FAAL with $FRL\text{-}RWRM_{0.05}$, it can be seen that FAAL is able to achieve comparable to FRL-RWRM while reducing the amount of learning epoch up to 40 times (2 epochs vs. 80 epochs). For full adversarial training, following the experimental settings in WAT [26], we train the ResNet-18 models for 100 epochs via different adversarial training approaches, where the learning rate is decayed from 0.1 to 0.01 and 0.001 at the 75-th epoch and the 90-th epoch, respectively. We compare the results of FAAL compared to three baselines, i.e. TRADES, CFA and WAT. It can be seen in Tab. 3 that $FAAT_{TRADES}$ achieves the highest worst-class robust accuracy (same as WAT), but it remains comparable results on the average robustness without sacrificing the average/worst-class clean accuracy. More details of the training settings can be found in the Appendix.
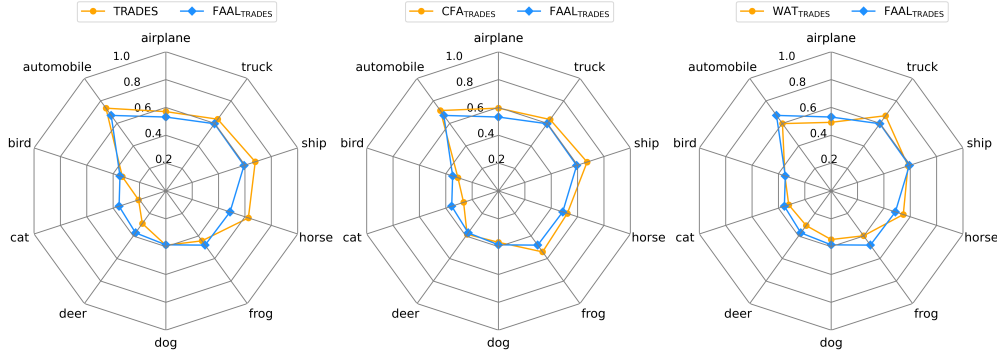
Figure 3. Class-wise robust accuracy against AutoAttack after adversarially trained PRN-18 model from scratch

Table 4. Comparison among different SOTA methods, all models are trained with the same number of samples under a single NVIDIA 3090Ti GPU in the same conda environment.

Methods	Training time per epoch (*min*)		Reweighting level	Adversary free	Validation set
	CIFAR-10 (PRN-18)	CIFAR-100 (RN-18)			
TRADES	2.63	2.68	fixed	×	×
FRL-RWRM	2.73	2.80	epoch	×	✓
WAT	2.88	3.00	epoch	×	✓
CFA	2.75	2.78	epoch	✓	✓
FAAL	2.69	2.73	batch	✓	×

5. Essential Differences to SOTAs

In this section, we highlight the essential differences of FAAL with existing state-of-the-art works, including FRL [43], WAT [26] and CFA [40]. Both FRL and WAT are *TRADES-based* approaches, which require a *separate* validation set for performing the reweight strategies. For example, FRL updates the lagrangian multiplier according to the performance of the validation set to meet the fairness constraints, so it requires many epochs for fine-tuning since it needs to search the whole space to achieve the optimal equilibrium without fairness constraint violation. Also, the remargin strategy in FRL essentially sacrifices some average clean accuracy. We argue that it is not necessary to enlarge the margin for improvement, which can be achieved by combining stronger perturbations instead. As another TRADES-based variant, WAT leverages no-regret dynamics and also relies on the validation set to tune the class weights for the current epoch training, Similarly, CFA proposed to apply the weight averaging only if the performance on the extra validation set meets a certain threshold, and relies on empirically adjusting the class margins and class regularization based on the performance of the previous epoch.

Different form those methods that rely on historical performance or an extra validation set for manual or heuristic weight adjustment per class in each epoch, our method *by-passes* these requirements. FAAL introduces an additional conic convex optimization problem after the adversarial example generation, based solely on the current batch's objective loss, the bringing solving cost is negligible. The comparison of training computation time and other key properties is illustrated in Tab. 4. As model training can be unpredictable due to random mini-batch sampling, causing quick shifts in class distribution and bias that may differ from previous epochs or validations. More importantly, FAAL can generalize to *any* adversarial training methods, as our intermediate maximization is a completely independent component plugged into the popular min-max framework, so it is not limited to any adversaries, unlike some methods FRL and WAT that are restricted to TRADES variants. Our data-driven component enhances flexibility in managing the balance between average robustness and robust fairness during adversarial training, and demonstrates its potential in handling various distribution shifts for the current batch.

6. Conclusion

In conclusion, we establish a connection between robust fairness and potential overfitting issues caused by the unknown group distribution shift, and present a new fairness-aware adversarial learning paradigm to address robust fairness via distributional robust optimization. Compared to state-of-the-art methods, extensive experiments on CIFAR-10 and CIFAR-100 datasets demonstrate the effectiveness and superior efficiency of the proposed approach. Notably, by just two epochs of fine-tuning, our training strategy can transform a biased robust model into one with high fairness with little cost on average accuracy. We believe our research provides a meaningful contribution to the discourse on robustness and fairness in machine learning, deepening our insight into the model's behaviors under adversarial settings.

Acknowledgement

References

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pages 60–69. PMLR, 2018. 2

[2] Aharon Ben-Tal, Laurent El Ghaoui, and Arkadi Nemirovski. *Robust optimization*. Princeton university press, 2009. 3

[3] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013. 2, 3

[4] Amine Bennouna and Bart Van Parys. Holistic robust data-driven decisions. *arXiv preprint arXiv:2207.09560*, 2022. 3

[5] Amine Bennouna, Ryan Lucas, and Bart Van Parys. Certified robust neural networks: Generalization and corruption resistance. *arXiv preprint arXiv:2303.02251*, 2023. 3, 4

[6] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. In *NeurIPS 2020 Workshop on Pre-registration in Machine Learning*, pages 325–342. PMLR, 2021. 3

[7] Dimitris Bertsimas, David B Brown, and Constantine Caramanis. Theory and applications of robust optimization. *SIAM review*, 53(3):464–501, 2011. 3

[8] Tuan Anh Bui, Trung Le, Quan Tran, He Zhao, and Dinh Phung. A unified wasserstein distributional robustness framework for adversarial training. *arXiv preprint arXiv:2202.13437*, 2022. 3

[9] Kaidi Cao, Colin Wei, Adrien Gaidon, Nikos Arechiga, and Tengyu Ma. Learning imbalanced datasets with label-distribution-aware margin loss. *Advances in neural information processing systems*, 32, 2019. 5

[10] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 ieee symposium on security and privacy (sp)*, pages 39–57. Ieee, 2017. 5

[11] Zhen Chen, Fu Wang, Ronghui Mu, Peipei Xu, Xiaowei Huang, and Wenjie Ruan. Nrat: towards adversarial training with inherent label noise. *Machine Learning*, pages 1–22, 2024. 1

[12] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International conference on machine learning*, pages 2206–2216. PMLR, 2020. 1, 5

[13] John C Duchi, Peter W Glynn, and Hongseok Namkoong. Statistics of robust optimization: A generalized empirical likelihood approach. *Mathematics of Operations Research*, 46(3):946–969, 2021. 2, 3

[14] Julien Ferry, Ulrich Aivodji, Sébastien Gambs, Marie-José Huguet, and Mohamed Siala. Improving fairness generalization through a sample-robust optimization method. *Machine Learning*, pages 1–62, 2022. 3

[15] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1

[16] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pages 1929–1938. PMLR, 2018. 2, 3

[17] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 7

[18] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. Does distributionally robust supervised learning give robust classifiers? In *International Conference on Machine Learning*, pages 2029–2037. PMLR, 2018. 3

[19] Xiaowei Huang, Daniel Kroening, Wenjie Ruan, James Sharp, Youcheng Sun, Emese Thamo, Min Wu, and Xinping Yi. A survey of safety and trustworthiness of deep neural networks: Verification, testing, adversarial attack and defence, and interpretability. *Computer Science Review*, 37:100270, 2020. 1

[20] Xiaowei Huang, Wenjie Ruan, Wei Huang, Gaojie Jin, Yi Dong, Changshun Wu, Saddek Bensalem, Ronghui Mu, Yi Qi, Xingyu Zhao, et al. A survey of safety and trustworthiness of large language models through the lens of verification and validation. *arXiv preprint arXiv:2305.11391*, 2023. 1

[21] Pavel Izmailov, Dmitrii Podoprikhin, Timur Garipov, Dmitry Vetrov, and Andrew Gordon Wilson. Averaging weights leads to wider optima and better generalization. *arXiv preprint arXiv:1803.05407*, 2018. 7

[22] Gaojie Jin, Xinping Yi, Wei Huang, Sven Schewe, and Xiaowei Huang. Enhancing adversarial training with second-order statistics of weights. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15273–15283, 2022. 1

[23] Gaojie Jin, Xinping Yi, Dengyu Wu, Ronghui Mu, and Xiaowei Huang. Randomized adversarial training via taylor expansion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16447–16457, 2023. 1

[24] Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. Re-weighting based group fairness regularization via classwise robust optimization. In *The Eleventh International Conference on Learning Representations*, 2022. 3

[25] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 5

[26] Boqi Li and Weiwei Liu. Wat: improve the worst-class robustness in adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 14982–14990, 2023. 2, 3, 7, 8

[27] Jiashuo Liu, Zheyan Shen, Peng Cui, Linjun Zhou, Kun Kuang, and Bo Li. Distributionally robust learning with stable adversarial training. *IEEE Transactions on Knowledge and Data Engineering*, 2022. 3

[28] Xinsong Ma, Zekai Wang, and Weiwei Liu. On the trade-off between robustness and fairness. In *Advances in Neural Information Processing Systems*, 2022. 2

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 3, 5, 7

[30] Yonatan Oren, Shiori Sagawa, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust language modeling. *arXiv preprint arXiv:1909.02060*, 2019. 3

[31] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[32] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019. 3, 4

[33] Aman Sinha, Hongseok Namkoong, and John Duchi. Certifying some distributional robustness with principled adversarial training. In *International Conference on Learning Representations*, 2018. 3

[34] Matthew Staib and Stefanie Jegelka. Distributionally robust deep learning as a generalization of adversarial training. In *NIPS workshop on Machine Learning and Computer Security*, page 4, 2017. 3

[35] Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, XiangLong Liu, and Aishan Liu. Improving robust fairness via balance adversarial training. *arXiv preprint arXiv:2209.07534*, 2022. 2

[36] Hieu Vu, Toan Tran, Man-Chung Yue, and Viet Anh Nguyen. Distributionally robust fair principal components via geodesic descents. In *International Conference on Learning Representations*, 2021. 3

[37] Fu Wang, Yanghao Zhang, Yanbin Zheng, and Wenjie Ruan. Dynamic efficient adversarial training guided by gradient magnitude. In *Progress and Challenges in Building Trustworthy Embodied AI*, 2022. 1

[38] Hongjun Wang and Yisen Wang. Self-ensemble adversarial training for improved robustness. *arXiv preprint arXiv:2203.09678*, 2022. 7

[39] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International conference on learning representations*, 2019. 5

[40] Zeming Wei, Yifei Wang, Yiwen Guo, and Yisen Wang. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8193–8201, 2023. 2, 3, 7, 8

[41] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations*, pages 38–45, 2020. 1

[42] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33:2958–2969, 2020. 5, 7

[43] Han Xu, Xiaorui Liu, Yaxin Li, Anil Jain, and Jiliang Tang. To be robust or to be fair: Towards fairness in adversarial training. In *International Conference on Machine Learning*, pages 11492–11501. PMLR, 2021. 1, 2, 3, 5, 6, 8

[44] Xiangyu Yin, Wenjie Ruan, and Jonathan Fieldsend. Dimba: discretely masked black-box attack in single object tracking. *Machine Learning*, pages 1–19, 2022. 1

[45] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016. 1, 5

[46] Shaoning Zeng, Bob Zhang, Yanghao Zhang, and Jianping Gou. Collaboratively weighting deep and classic representation via l_2 regularization for image classification. In *Asian conference on machine learning*, pages 502–517. PMLR, 2018. 1

[47] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019. 5, 7

[48] Yanghao Zhang, Fu Wang, and Wenjie Ruan. Fooling object detectors: Adversarial attacks by half-neighbor masks. *arXiv preprint arXiv:2101.00989*, 2021. 1

[49] Yanghao Zhang, Wenjie Ruan, Fu Wang, and Xiaowei Huang. Generalizing universal adversarial perturbations for deep neural networks. *Machine Learning*, 112(5):1597–1626, 2023. 1