FERD: FAIRNESS-ENHANCED DATA-FREE ADVERSARIAL ROBUSTNESS DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Data-Free Robustness Distillation (DFRD) aims to transfer the robustness from the teacher to the student without accessing the training data. While existing methods focus on overall robustness, they overlook the robust fairness issues, leading to severe disparity of robustness across different categories. In this paper, we find two key problems: (1) student model distilled with equal class proportion data behaves significantly differently across distinct categories; and (2) the robustness of student model is not stable across different attacks targets. To bridge these gaps, we present the first Fairness-Enhanced data-free adversarial Robustness Distillation (FERD) framework to adjust the proportion and distribution of adversarial examples. For the proportion, FERD adopts a robustness-guided class reweighting strategy to synthesize more samples for the less robust categories, thereby improving robustness of them. For the distribution, FERD generates complementary data samples for advanced robustness distillation. It generates Fairness-Aware Examples (FAEs) by enforcing a uniformity constraint on feature-level predictions, which suppress the dominance of class-specific non-robust features, providing a more balanced representation across all categories. Then, FERD constructs Uniform-Target Adversarial Examples (UTAEs) from FAEs by applying a uniform target class constraint to avoid biased attack directions, which distributes the attack targets across all categories and prevents overfitting to specific vulnerable categories. Extensive experiments on three public datasets demonstrate that FERD achieves state-of-the-art worst-class robustness and NSD under all adversarial attacks. For instance, FERD improves worst-class robustness by up to 19.6% and reduces NSD by 0.092 compared to the optimal baseline on CIFAR-10 with MobileNet-V2. Our code is available at: `https://anonymous.4open.science/r/FERD-2A48/`.

1 INTRODUCTION

With the widespread use of Deep Neural Networks (DNNs) (Goswami et al., 2018; Gongye et al., 2024; Lim et al., 2024), the deployment of lightweight models on edge devices has become increasingly important (Mittal, 2024; Min et al., 2024; Liu et al., 2024). However, a large number of studies have shown that these lightweight models are weakly robust in the face of adversarial attacks (Ma et al., 2021; Li et al., 2022; Croce & Hein, 2020; Bai et al., 2023). While traditional adversarial training methods (Jia et al., 2022; Hsiung et al., 2023; Jia et al., 2024b;a), though showing significant advantages on large models, are difficult to achieve desirable results in lightweight models (Wang et al., 2024b; Ye et al., 2019; Huang et al., 2021). To enhance the defense capability of lightweight models, researchers have proposed the concept of adversarial robustness distillation (Zhang et al., 2019; Goldblum et al., 2020; Zi et al., 2021; Zhu et al., 2021; Huang et al., 2023; Yue et al., 2024; Zhu et al., 2023), which aims to migrate the defense capability of the robust teacher to the student, thereby enhancing the latter's performance in an adversarial setting.

However, in practice, raw training data for the teacher model are often unavailable, making it difficult to apply traditional distillation methods directly. To break through this limitation, researchers propose the Data-Free Knowledge Distillation (DFKD) method (Micaelli & Storkey, 2019; Fang et al., 2021; Yin et al., 2020; Fang et al., 2022), which synthesizes alternative samples through a training generator to simulate the original data distribution and achieve knowledge transfer. Based on this, the Data-Free Robustness Distillation (DFRD) method (Yuan et al., 2024; Wang et al., 2024a; Zhou et al., 2024) is further developed to combine the generation and distillation mechanisms to effectively deliver robustness without the need for real data, providing a novel solution for resource-constrained and data-unavailable environments for model defense.

Despite their effectiveness, existing DFRD methods mainly aim to enhance the overall robustness of the student, while overlooking a critical issue: robust fairness (Sun et al., 2023; Yue et al., 2023; Zhao et al., 2024). The robust model may exhibit strong resistance to adversarial attacks on specific categories and remain vulnerable in others, leading to inconsistent robustness performance across different categories. It impacts the reliability and fairness of the model in practical applications.

In this paper, we make the first attempt at investigating the robust fairness problem in the context of DFRD. We find that although the student tends to inherit the teacher's class-wise robustness pattern, the inter-class robustness gap is significantly amplified in the distillation process. We find two phenomena affecting robust fairness: (1) students distilled with equally distributed synthetic data still show class-wise robustness discrepancies; and (2) the success rate of adversarial attacks on students varies significantly depending on the target class. Based on these findings, we propose a Fairness-Enhanced data-free adversarial Robustness Distillation (FERD) framework by adjusting the proportion and distribution of the synthetic samples to mitigate these problems.

Specifically, for the proportion, we introduce a robustness-guided class reweighting strategy that encourages the generator to synthesize more samples from weakly robust categories, thereby compensating for their vulnerability and promoting fairness in robustness. For the distribution, we propose to generate complementary data samples for advanced robustness and fairness distillation. Firstly, we generate Fairness-Aware Examples (FAEs) by enforcing a uniformity constraint on feature-level predictions that are closely associated with non-robust representations. This helps suppress the dominance of class-specific non-robust features, ensuring that the benign samples provide a more balanced representation across all classes. To avoid biased attack directions, we further construct Uniform-Target Adversarial Examples (UTAEs) from FAEs by applying a uniform target class constraint during adversarial generation. It distributes the attack targets evenly across all categories and prevent overfitting to specific vulnerable categories. We conjecture robust distillation on adversarial examples with uniformly distributed targets can defend against attacks from different targets.

Experiments on the three datasets (CIFAR-10, CIFAR-100, and Tiny-ImageNet) show that compared with baseline method, FERD improves +19.6%, +8.9%, +10.3% and +6.6% on the worst-class robustness against FGSM (Goodfellow et al., 2014), PGD-20 (Madry et al., 2017), CW_∞ (Carlini & Wagner, 2017), and AutoAttack (AA) (Croce & Hein, 2020), respectively, alleviating the robustness bias problem in the DFRD to a certain extent.

Our contributions can be summarized as follows: (I) We make the first attempt at investigating the problem of robust fairness in DFRD, revealing that uniform distribution among original data categories and attack target bias are the two key factors affecting fairness. (II) We propose the FERD framework, which enhances robust fairness at both the proportional and distributional levels through robustness-guided category reweighting and distribution-aware sample generation mechanisms. (III) Experiments have demonstrated that FERD significantly enhances the robustness and fairness of the student model, and the robust accuracy in the weakest class has been improved by +19.6% compared with the existing optimal methods, effectively alleviating the robust bias phenomenon in DFRD.

2 Related Work

2.1 Data-Free Robustness Distillation

DFRD aims to transfer the robustness from teacher to student using synthetic samples instead of teacher's original training data. DFARD (Wang et al., 2024a) first defines the concept of DFRD and proves that the difficulty lies in the lower upper bound of knowledge transfer information. They propose an interactive temperature adjustment strategy and an adaptive generator to solve the problem. DERD (Zhou et al., 2024) takes a homogenized expert guidance strategy. Both clean and robust knowledge are distilled from clean and robust teachers respectively, using the same synthetic data. To coordinate the gradients of the clean and robust distillation tasks, DERD also introduces a stochastic gradient aggregation module, thereby optimizing the trade-off between robustness and

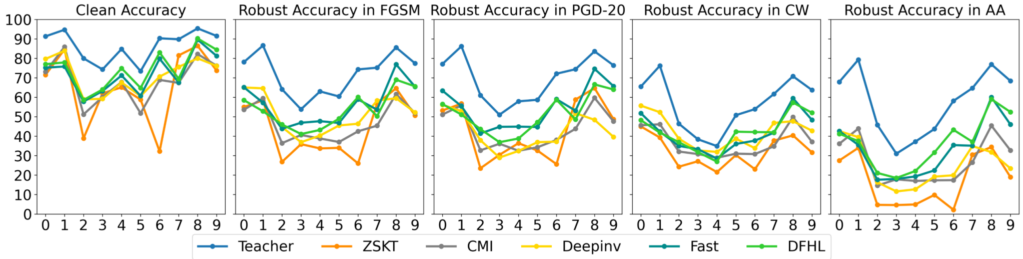


Figure 1: Comparison of the accuracy of the teacher and the students distilled from different DFRD methods under benign and adversarial examples. The blue line represents the teacher and the other lines correspond to the students. The horizontal axis indicates the category, and the vertical axis indicates the accuracy.

accuracy. DFHL (Yuan et al., 2024) proposes the concept of High-Entropy Examples (HEEs), which can characterize a more complete shape of the classification boundary. Distillation on HEEs achieves the best balance between clean accuracy and robustness. It is worth mentioning that although DFKD does not involve robustness, we can transform it into DFRD by adding adversarial noise to synthetic samples during its distillation stage. For example, Fast (Fang et al., 2022) proposes a fast DFKD, which reuses the common features in the training data to synthesize different data instances. By generating adversarial examples on synthetic samples, a distilled robust student can be obtained. In this paper, we make the first attempt at solving the problem of robust fairness transfer in DFRD.

2.2 Fairness in Robustness

While enhancing overall model robustness is a common goal, it inevitably leads to significant class-wise performance discrepancies: models become highly robust for some categories while others remain vulnerable to adversarial attacks. FRL (Xu et al., 2021) is a pioneer work in highlighting this issue and introduces the concept of "robust fairness" to assess such class-wise robustness disparities. To address this issue, FRL proposes fairness constraints, adjusting decision margins and sample weights when these constraints are violated. BAT (Sun et al., 2023) identify distinct "source-class" and "target-class" unfairness within adversarial training, tackling these by adjusting per-class attack intensities and applying a uniform distribution constraint. Further methods include those by Fair-ARD (Yue et al., 2023), who improves student model robust fairness by increasing the weights of difficult classes, and ABSLD (Zhao et al., 2024), who focuses on adaptively reducing inter-class error risk gaps by modulating the class-wise smoothness of samples' soft labels during training. In this paper, we introduce FERD to simultaneously enhance model robustness and alleviate robust unfairness problems.

3 Observation of Fairness in DFRD

In this section, we investigate whether robust fairness of the teacher is transferred to the student distilled from DFRD methods. We research the robust fairness performance of the models and defense effects against adversarial attacks from different target categories.

3.1 Robust Fairness of the Student

To evaluate the robust fairness of the student, we compare the classification performance of the teacher and the students distilled by five methods (ZSKT (Micaelli & Storkey, 2019), CMI (Fang et al., 2021), DeepInv (Yin et al., 2020), Fast (Fang et al., 2022), and DFHL (Yuan et al., 2024)) on different classes. We measure the accuracy of each category under benign samples and four adversarial attacks (FGSM, PGD-20, CW_{∞}, and AA). The results in Fig.1 show that the performance of the student on different categories follows a consistent trend with that of the teacher. Note that the sample labels all adopt a uniform sampling strategy in the aforementioned DFRD methods, in which case the student conducts robust distillation on equally distributed synthetic data. However, the robustness among different categories varies significantly. Therefore, we argue that sample number

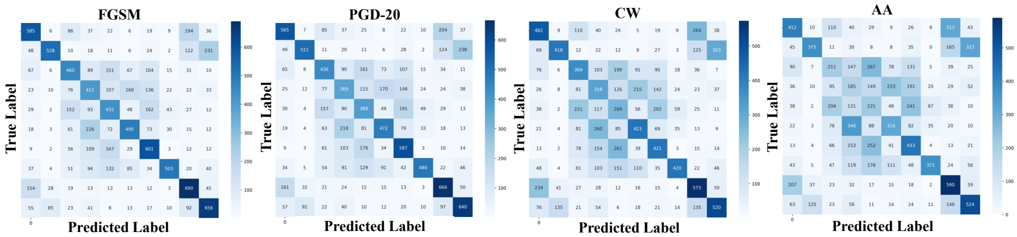


Figure 2: Confusion matrices under different adversarial attacks. The horizontal axis denotes predicted labels, and the vertical axis denotes true labels. Darker colors indicate a higher number of samples predicted as the correct class.

imbalance is the key to achieving class balance, because categories behave differently in robust distillation. Some categories are more difficult to achieve robustness, while others quickly reach a high level of robustness. We conjecture that by adjusting the weight of class sampling and increasing the number of samples from weakly robust classes, the problem of robust unfairness can be effectively alleviated. The proof of this conjecture is shown in Appendix A.1.

3.2 Target Fairness of Adversarial Examples

We make confusion matrices to visualize the classification results, as shown in Fig.2. We observe that the students' robustness against adversarial attacks on different targets varies. For samples with the original label of class 0, when the attack target is class 9, the student is more prone to misclassification, showing poor robustness; while for adversarial attacks on other classes, its robustness is relatively strong. This indicates that the student exhibits significant differences in defending against adversarial attacks on different targets and are more vulnerable to specific class adversarial attacks. We conjecture that robust distillation on adversarial examples with uniformly distributed targets can defend against attacks from different targets. The proof of this conjecture is shown in Appendix A.2.

4 Methodology

4.1 Overall Framework

In the above section, we find two factors affecting robust fairness: (1) the student distilled on data with equal class proportions shows class-wise robustness discrepancies; and (2) the success rate of adversarial attacks on the student varies significantly depending on the target class.

In this section, we introduce our FERD framework and the overall framework is in Fig.3. The algorithm is shown in the Appendix A.3. For the the first proportion problem, we introduce a robustness-guided class reweighting strategy that encourages the generator to synthesize more samples from weakly robust classes. For the second distribution problem, we impose uniform constraints on the feature space of synthetic samples and the generation direction of adversarial examples, respectively, making the targets of adversarial examples more uniformly distributed. Specifically, we firstly generate Fairness-Aware Examples(FAEs) by enforcing constraint on predictions from non-robust feature, which are highly related to adversarial targets. Then, to avoid biased attack directions, we further construct Uniform-Target Adversarial Examples (UTAEs) from FAEs by applying constraint during adversarial generation, preventing attacks focus on "vulnerable category".

4.2 Class-Aware Sample Reweighting for Fair Distillation

In traditional DFRD, the labels y_i of synthetic samples x_i are sampled from a uniform distribution, as $y_i \sim \mathcal{U}(0, C-1)$, where $\mathcal{U}$ means uniform distribution and C means class number. However, robust distillation on equally distributed data infers a significant robustness unfairness problem.

To mitigate this problem, we introduce the robustness-guided class reweighting strategy based on adversarial margin, aiming to guide the generator to synthesize more samples of categories with poorer robustness. Specifically , we first generate adversarial examples x_i^{adv} from the synthetic

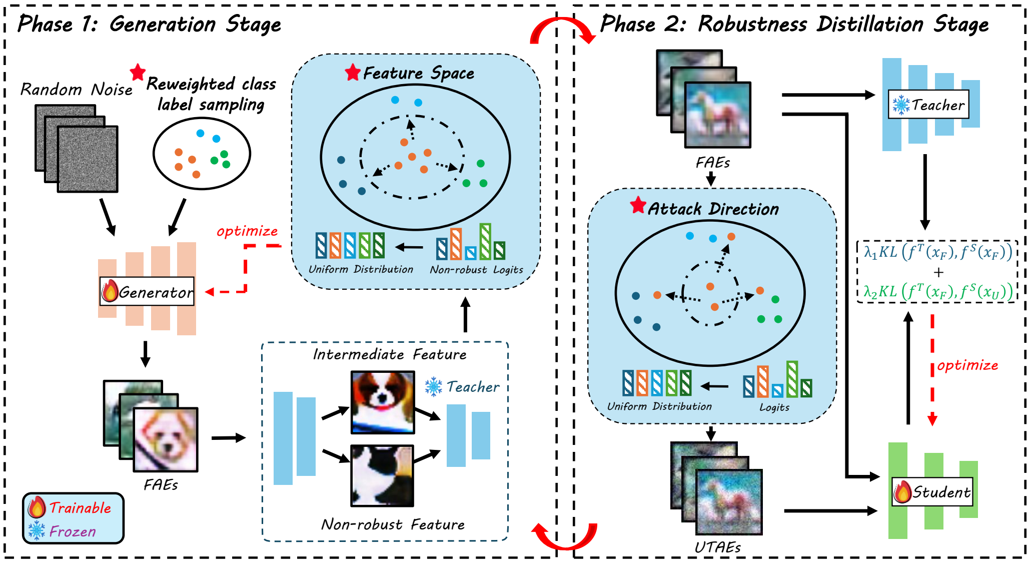


Figure 3: Framework of our FERD. In the generation stage, we use a robustness-guided class reweighting strategy to synthesize more weak-class samples and apply a uniformity constraint to non-robust feature predictions to generate FAEs. During robust distillation, we construct UTAEs from FAEs, using them respectively as benign samples and adversarial examples.

samples x_i using PGD-20 attack. Then, we calculate the adversarial margin m_i of each adversarial sample x_i^{adv} under the teacher model f^T:

$$m_i = \left(f^T(x_i^{adv})\right)_{y_i} - \max_{j \neq y_i} \left(f^T(x_i^{adv})\right)_j . \tag{1}$$

The adversarial margin measures the gap between the model's confidence in the correct category and the strongest confusable category. The smaller the value, the more probably x_i^{adv} is to be misclassified. A negative margin indicates that the attack has been successful.

Therefore, we calculate the average negative adversarial margin for each category:

$$\mathcal{D}c = \frac{1}{N_c} \sum_{i:y_i=c} (-m_i), \tag{2}$$

where N_c means the number of samples in category c. This value is used to measure the robustness vulnerability of the category. A higher value indicates that the category c is more likely to be misclassified when facing adversarial attack.

Finally, we apply softmax function to transform $\mathcal{D}c$ into sampling probability distribution p_c, which is used as the sampling weight for categories in the subsequent sample generation stage, enabling to adaptively generate more samples of less robust categories. Furthermore, we compare several other reweighting strategies. The experimental results are shown in Appendix A.6.

4.3 Non-Robust Feature Suppression for Balanced Representations

To achieve the fair tendency to each target when generating adversarial examples, we design a FAEs generation method and use them as benign samples, which ensures that FAEs' non-robust feature predictions are not concentrated in a few categories; otherwise, adversarial perturbations would be more likely to attack such categories.

We adopt a modified information bottleneck approach to achieve this. The standard information bottleneck seeks to get a compressed representation Z of the input X that is maximally informative about the target label Y. The objective function can be written as follows:

$$\mathcal{L}_{IB} = \mathcal{I}(Z;Y) - \beta\mathcal{I}(Z;X), \tag{3}$$

where $\mathcal{I}$ denotes the mutual information and β balances the trade-off between prediction accuracy and compression. We attempt to distill non-robust features Z_{nr} from the intermediate feature representation $Z = f_l^T(x)$, where $f_l^T(\cdot)$ describes l-th layer outputs of the teacher.

When the synthetic samples x_i is input into f^T, we inject a learnable noise scale λ_r into Z and define informative features Z_I as follows:

$$Z_I = f_l^T(x_i) + \text{softplus}(\lambda_r) \cdot \epsilon, \epsilon \sim \mathcal{N}(0, I). \tag{4}$$

After obtaining Z_I added by λ_r, we propagate Z_I to the subsequent output layer f_{l+}^T and evaluate the influence of each unit's feature to the teacher prediction. When distilling Z_{nr} from Z_I, we must ensure Z_I remains predictive while encouraging robustness to noise. Since directly optimizing mutual information is intractable, we use a variational approximation (Kim et al., 2021). The optimization objective is formulated as:

$$\min \mathcal{L}(\lambda_r) = CE(f_{l+}^T(Z_I), y_i) + \beta \cdot \sum_{c=1}^{Channel} \left(\frac{v_c}{\lambda_c^2} + \log\left(\frac{\lambda_c^2}{v_c}\right) - 1 \right), \tag{5}$$

where v_c and λ_c denote the c-th channel of Z_I and λ_r respectively. The first term ensures that Z_I correctly predicts the target label y_i, while the second term acts as a regularizer to control the amount of information passing through the bottleneck.

After optimizing λ_r, we identify non-robust channels index i_{nr_k} by comparing λ_r^2 with the maximum variance across all channels:

$$i_{nr_k} = \mathbb{1}\left[\lambda_k^2 < \max_{c' \in \{1,\ldots,C\}} \left\{ \text{Var}\left(Z_I^{c'} \right) \right\} \right]. \tag{6}$$

The right-hand side represents the upper limit of perturbation. Correspondingly, non-robust features Z_{nr} are obtained via channel-wise masking: $Z_{nr} = i_{nr} \cdot Z$. The prediction results of Z_{nr} are vulnerable to perturbations, so that they are highly correlated with adversarial predictions. To enable the generator to synthesize FAEs, we minimize the KL divergence between the predictions of non-robust features and the uniform distribution:

$$\mathcal{L}_{uni} = KL(\mathcal{U}, f_{l+}^T(Z_{nr})). \tag{7}$$

To further enhance the quality and diversity of the FAEs, we introduce additional loss functions during the training process of the generator:

$$\begin{cases} \mathcal{L}_{adv} = KL\left(f^T(x_i), f^S(x_i)\right) \\ \mathcal{L}_{bn} = \sum_l \left(\|\mu_l(x_i) - \mu_l\|_2 + \left\|\sigma_l^2(x_i) - \sigma_l^2\right\|_2 \right), \\ \mathcal{L}_{oh} = CE\left(f^T(x_i), y_i\right) \end{cases} \tag{8}$$

where $\mathcal{L}_{adv}$ encourages divergence between the student and teacher, promoting the diversity of FAEs; $\mathcal{L}_{bn}$ improves the visualization of the FAEs by matching the statistical information (mean μ_l and variance σ_l^2) stored in BatchNorm layers of the teacher and the student; $\mathcal{L}_{oh}$ ensures that the FAEs are correctly predicted by the teacher.

Therefore, the overall loss function for the generator in the generation stage are summarized as:

$$\mathcal{L}_{gen} = \lambda_{adv} \cdot \mathcal{L}_{adv} + \lambda_{bn} \cdot \mathcal{L}_{bn} + \lambda_{oh} \cdot \mathcal{L}_{oh} + \lambda_{uni} \cdot \mathcal{L}_{uni}, \tag{9}$$

where these hyperparameters λ_{adv}, λ_{bn}, λ_{oh} and λ_{uni} are adjusted empirically to balance the trade-offs between robustness, fairness, and accuracy. The hyper-parameter selection experiments are shown in Appendix A.7. By training with the above loss function, the generator synthesizes FAEs x_F, which not only convey effective knowledge, but also be fairer in terms of the tendency towards different categories when generating adversarial examples.

4.4 Label-Space Attack Diversification for Fairness Optimization

To address the problem that the student shows significant differences in defending against adversarial attacks with different targets and is more vulnerable to them from specific categories, we propose a novel adversarial examples generation method, named UTAEs generation.

Table 1: Result in average robustness(%) (Avg. ↑), worst robustness(%) (Worst ↑), and normalized standard deviation (NSD ↓) on CIFAR-10. RN-18 and MN-V2 are abbreviations of ResNet-18 and MobileNet-V2 respectively. The best results are **bolded**, and the second best results are underlined.

Student	Method	Clean			FGSM			PGD			CW_{∞}			AA		
		Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
RN-18	ZSKT	65.48	32.20	0.266	29.80	5.00	0.563	29.35	4.50	0.591	15.84	1.80	0.714	17.16	2.20	0.740
	CMI	68.29	51.20	0.161	33.34	19.50	0.337	28.97	16.00	0.392	21.74	11.00	0.453	26.85	14.60	0.419
	DeepInv	71.22	58.90	0.124	38.60	22.50	0.303	27.83	14.30	0.381	27.92	13.80	0.406	25.19	11.60	0.422
	Fast	72.20	57.80	0.132	43.96	26.90	0.295	39.60	22.40	0.328	28.03	14.00	0.434	33.24	17.70	0.401
	DFHL	74.42	58.60	0.129	41.67	24.30	0.298	36.91	19.30	0.344	29.45	13.40	0.408	36.39	18.50	0.351
	FERD(Ours)	**79.26**	**65.90**	**0.105**	**51.09**	**31.30**	**0.237**	**42.68**	**23.50**	**0.298**	**36.17**	**19.00**	**0.338**	**40.12**	**20.80**	**0.325**
MN-V2	ZSKT	54.69	16.10	0.394	23.62	3.90	0.663	22.76	3.80	0.678	12.78	1.00	0.733	14.85	1.60	0.749
	CMI	60.70	38.60	0.205	27.90	12.50	0.393	19.61	8.50	0.480	20.10	7.90	0.448	18.09	7.60	0.506
	DeepInv	62.77	46.00	0.170	31.66	14.50	0.306	19.08	7.20	0.404	23.60	9.10	0.372	15.39	5.40	0.471
	Fast	58.60	43.40	0.198	29.91	15.70	0.374	24.26	11.60	0.441	18.20	6.00	0.556	18.35	6.80	0.563
	DFHL	70.83	50.60	0.153	37.22	16.70	0.320	33.22	14.20	0.358	25.28	10.40	0.407	32.58	13.70	0.368
	FERD(Ours)	**78.34**	**64.10**	**0.108**	**52.19**	**36.30**	**0.214**	**40.87**	**23.10**	**0.301**	**37.13**	**20.70**	**0.321**	**38.06**	**20.30**	**0.349**

We aim to construct a more evenly distributed type of adversarial perturbation, so that the adversarial targets are not limited to certain "easily misclassified" classes, but uniformly cover the entire class space. To achieve this, we apply a uniform target class constraint during adversarial generation, avoiding to attacks from a single direction. The generation formula is as follows:

$$x_U^{t+1} = \Pi_{x_U+\mathcal{S}}\Big(x_U^t + \alpha \cdot \text{sign}\Big(\nabla_{x_U^t}\left[KL\left(f^T(x_i), f^T(x_U^t)\right) - \gamma \cdot KL\left(\mathcal{U}, f^T(x_U^t)\right)\right]\Big)\Big), \quad (10)$$

where $\nabla_{x_U^t}$ denotes the gradient of the entropy loss function w.r.t. the UTAE in step t and α is the step size. By introducing the KL divergence between $f^T(x_U)$ and $\mathcal{U}$ in the adversarial example generation, we make the target distribution of adversarial examples more extensive.

After synthesizing FAEs x_F and UTAEs x_U, we employ them as benign samples and adversarial examples respectively for robust distillation. Here, the robust distillation framework is as follows:

$$\mathcal{L}_{stu} = \lambda_1 KL\left(f^T(x_F), f^S(x_F)\right) + \lambda_2 KL\left(f^T(x_F), f^S(x_U)\right). \quad (11)$$

Through robust distillation of diversified adversarial targets, we force the student to inherit robustness across a far broader range of categories, thereby enhancing overall defensive capability against attacks from all directions.

5 Experiments

5.1 Experimental Settings

Datasets & Models. We conduct our experiments on three datasets: CIFAR-10 (Krizhevsky et al., 2009), CIFAR-100, and Tiny-ImageNet (Le & Yang, 2015). For the teacher model, we select WideResNet-34-10 (Zagoruyko & Komodakis, 2016) for both CIFAR-10 and CIFAR-100, while PreActResNet-34 is used for Tiny-ImageNet. For the student model, we select ResNet-18 (He et al., 2016a) and MoileNet-v2 (Sandler et al., 2018) for CIFAR-10 and CIFAR-100. For Tiny-ImageNet, we use PreActResNet-18. The performance of the teacher and the experiments on Tiny-ImageNet are shown in Appendix A.4 and A.5.

Baselines. We compare our method with five different DFRD methods, including ZSKT (Micaelli & Storkey, 2019), CMI (Fang et al., 2021), DeepInv (Yin et al., 2020), Fast (Fang et al., 2022), and DFHL (Yuan et al., 2024). Note that the first four methods belong to DFKD initially and do not involve robustness. We use PGD to generate adversarial examples and then apply the same distillation training loss as RSLAD (Zi et al., 2021) to transform them into DFRD.

Implementation Details. Our proposed method and all baselines are implemented on NVIDIA A800 GPU. The generator is trained via Adam optimizer with a learning rate of 2e-3, β_1 of 0.5, β_2 of 0.999. The student is trained via SGD optimizer with an initial value of 0.1, momentum of 0.9, and weight decay of 5e-4. The distillation epochs is set to 220 and the training iterations of generator and student are 200 and 400 respectively. The batch size is set to 256 for CIFAR-10 and 512 for both CIFAR-100 and Tiny-ImageNet. In Eq.4, we extract the intermediate features from the

Table 2: Result in average robustness(%) (Avg. ↑), worst-10% robustness(%) (Worst ↑), and normalized standard deviation (NSD ↓) on CIFAR-100. RN-18 and MN-V2 are abbreviations of ResNet-18 and MobileNet-V2. The best results are **bolded**, and the second best results are underlined.

Student	Method	Clean			FGSM			PGD			CW_∞			AA		
		Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
RN-18	ZSKT	36.44	6.40	0.522	12.69	0.10	0.985	12.90	0.00	0.960	6.91	0.00	1.307	6.01	0.00	1.361
	CMI	51.60	27.20	0.295	25.40	5.60	0.624	22.62	3.70	**0.685**	15.66	1.10	0.877	17.72	1.30	0.824
	DeepInv	53.44	27.20	0.298	26.14	5.60	0.630	23.10	3.50	0.699	15.99	1.40	0.899	17.49	1.60	0.857
	Fast	50.69	23.50	0.323	26.45	5.50	0.634	24.12	4.00	0.701	15.85	1.30	0.907	18.24	**1.70**	0.859
	DFHL	46.27	18.70	0.370	18.73	2.60	0.799	15.71	1.30	0.929	13.00	0.80	1.008	15.23	1.10	0.954
	FERD(Ours)	**55.91**	**29.00**	**0.289**	**29.23**	**5.70**	**0.622**	**25.90**	**4.10**	0.696	**19.10**	**1.50**	**0.861**	**21.79**	**1.70**	**0.805**
MN-V2	ZSKT	30.79	2.70	0.650	13.66	0.00	0.975	13.55	0.00	0.971	8.61	0.00	1.209	8.62	0.00	1.208
	CMI	35.80	6.00	0.425	13.82	0.60	0.828	9.89	0.50	0.964	9.59	0.30	0.953	8.04	0.00	1.042
	DeepInv	38.10	8.00	0.447	13.97	1.00	0.920	10.03	0.10	1.122	8.73	0.00	1.205	6.67	0.00	1.416
	Fast	39.00	6.00	0.441	16.42	1.40	0.821	14.07	0.80	0.912	10.26	0.00	1.082	9.53	0.00	1.139
	DFHL	42.08	11.30	0.436	16.25	0.90	0.914	13.48	0.60	1.020	10.67	0.20	1.148	12.99	0.40	1.051
	FERD(Ours)	**56.14**	**29.20**	**0.285**	**29.63**	**6.10**	**0.611**	**24.39**	**3.30**	**0.718**	**19.77**	**2.00**	**0.821**	**21.20**	**2.30**	**0.793**

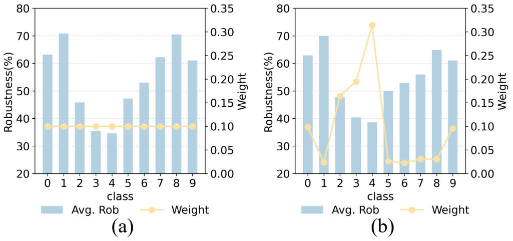


Figure 4: (a): The robustness of the student for each category under equal weight. (b): The robustness of the student for each category under reweighting situation.

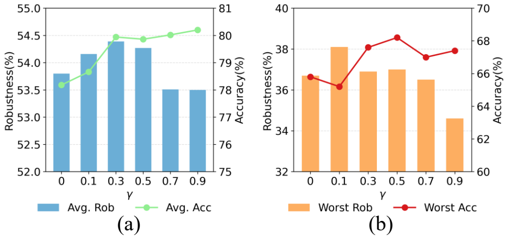


Figure 5: Ablation study. (a): The average robustness and accuracy of the student under different γ. (b):The worst-class robustness and accuracy of the student under different γ.

last convolutional layer l. In Eq.9, Eq.10, and Eq.11, we set the hyperparameter λ_{adv}=1, λ_{bn}=5, λ_{oh}=1, λ_{uni}=5, γ=0.5, λ_1=5/6, and λ_2=1/6.

Evaluation Metrics. We evaluate the robustness of the student against four adversarial attacks: FGSM, PGD, CW_∞, and AA. Following (Yue et al., 2023; Zhao et al., 2024), we employ the worst-class robustness and Normalized Standard Deviation (NSD) to quantify the robust fairness across categories. NSD is a normalized metric of the adversarial robustness with respect to the standard deviation across different classes. The smaller the value of NSD, the better. Notably, for CIFAR-100 and Tiny-ImageNet, we adopt worst-10% robustness in place of worst-class robustness, due to the limited size of test set per category and poor performance in the worst class robustness.

5.2 Experimental Results

Overall performance. Tab. 1 and Tab. 2 show the performance of ResNet-18 and MobileNet-V2 distilled on CIFAR-10 and CIFAR-100 by our method and baselines. The results demonstrate that the student distilled from FERD has a improvement in the robustness and fairness. Our method achieves state-of-the-art worst-class robustness under all attacks. When distilled on CIFAR-10 with MobileNet-V2, FERD improves the worst class robustness by 19.6%, 8.9%, 10.3%, and 6.6% compared with the best baseline against four adversarial attack respectively. In addition, there is an improvement in the average accuracy and NSD in most situation. Specifically, as shown in Tab. 1, FERD achieves the highest average accuracy under all attacks with an improvement of up to 14.97%, which is attributed to our high-quality synthetic samples.

Effectiveness of reweighting. Fig. 4 illustrates the impact of reweighting sampling strategy on the robustness of the student model. The robustness-guided reweighting strategy we proposed effectively identifies the categories with poor robustness and increases their sampling weights. Specifically, the weight of the fourth class which has the lowest robustness reaches 0.314, exceeding other classes substantially. Correspondingly, the student's robustness in these categories are improved, thereby enhancing robust fairness.

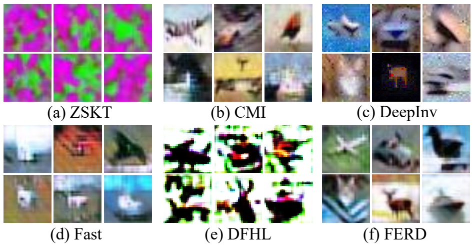


Figure 6: 32×32 images generated by inverting a WideResNet-34-10 trained on CIFAR-10 with different methods. Clockwise: airplane, car, bird, ship.

Table 3: Ablation study of different components in the framework.

Method	Clean		PGD	
	Avg.	Worst	Avg.	Worst
FERD	79.86	**65.90**	**42.68**	**23.50**
w/o reweighting	**80.13**	63.20	42.08	19.30
w/o FAEs	79.02	65.30	42.21	21.90
w/o UTAEs	78.17	65.10	41.96	22.60
w/o FAEs UTAEs	78.08	62.50	41.24	21.00

Synthetic data visualization. In Fig. 7, we visualize the synthetic data generated by FERD and baselines. The results indicate that our generator is able to synthesize more visible data. In such scenarios, CMI and Fast even suffer from model collapse problem, where the visual quality of the synthetic samples is extremely low. This further demonstrates the superiority of our method, which recover high-quality samples from the robust teacher.

5.3 Ablation Study

In this section, we provide ablation studies on FERD. We keep the same settings with experiments and use WideResNet-34-10 as the teacher, ResNet18 as the student and CIFAR-10 as dataset.

Effectiveness of all components. To verify the effectiveness of our FERD, we conduct ablation studies for each component, and the experimental results are shown in Tab. 3. We first examine the impact of the reweighting strategy. Compared with the absence of it, the student's accuracy on worst-class significantly improves when reweighting strategy is applied. However, it slightly compromises the overall accuracy, which is consistent with the findings of previous researches (Yue et al., 2023; Zhao et al., 2024). Further, we analyze the contributions of FAEs and UTAEs. Removing either component individually results in a consistent performance decline across all metrics. When both FAEs and UTAEs are removed simultaneously, the performance further deteriorates. This confirms that FAEs and UTAEs are complementary and their joint effect is crucial for achieving the best robust fairness performance.

Hyperparameter γ. In Fig.5 (a) and (b), we illustrate the average and worst-class performance of the student distilled with with varying γ during UTAEs generation. γ controls the strength of the uniform target class constraint during adversarial examples generation. Note that when γ=0, the adversarial examples are the same as the standard PGD method. For average robustness and average accuracy, we observe that a low to medium γ (e.g., 0.1 to 0.5) positively enhances the robustness and fairness of the student. This indicates the uniform constraint enhances the robustness of the model without significantly affecting the intensity of adversarial perturbations. However, A high γ (e.g., 0.7 or 0.9) enforces strong uniform constraint, which overly suppress the optimization of the adversarial loss, leading to weaker perturbations that reduce attack strength of the adversarial examples. This negatively affects the robust distillation and leads to a reduction in overall robustness.

6 Conclusion

In this paper, we made the first attempt to investigate the robust fairness in DFRD. We summarized two key factors affecting robust fairness and propose a FERD framework to mitigate these problems by adjusting the proportion and distribution of adversarial examples. For the proportion, we introduced a robustness-guided class reweighting strategy to encourage the generator to synthesize more samples from weakly robust classes. For the distribution, we designed FAEs and UTAEs, taking them as benign samples and adversarial examples respectively for robust distillation. Extensive experiments show that FERD significantly improves the robust and fairness performance of the student model. Our work is more applicable and can be effectively applied in practical scenarios.

7 Ethics Statement

This paper aims to address the issue of robust fairness in data-free robustness distillation. The FERD framework we propose is designed to enhance the fairness of the model's robust performance across different categories, thereby improving its reliability and security in real-world applications. All the experiments are conducted on publicly available and standard datasets, which are widely used in the machine learning. These datasets do not involve any sensitive or private personal information, avoiding the risks of data privacy and abuse.

8 Reproducibility Statement

To ensure the reproducibility of our work, we have provided comprehensive details of our experiments. The experimental setting of FERD is detailed in Section 5.1, including the models, baselines, evaluation metrics, and implementation details. All datasets (CIFAR-10, CIFAR-100, and Tiny-ImageNet) are publicly available. Final hyperparameter values for all experiments are explicitly listed, with further analysis on their selection provided in Appendices A.7. To facilitate direct replication of our results, we make our source code publicly available upon publication.

References

Yang Bai, Yisen Wang, Yuyuan Zeng, Yong Jiang, and Shu-Tao Xia. Query efficient black-box adversarial attack on deep neural networks. *Pattern Recognition*, 133:109037, 2023.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM review*, 60(2):223–311, 2018.

Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy (SP)*, pp. 39–57. Ieee, 2017.

Erh-Chung Chen and Che-Rung Lee. Data filtering for efficient adversarial training. *Pattern Recognition*, 151:110394, 2024.

Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020.

Gongfan Fang, Jie Song, Xinchao Wang, Chengchao Shen, Xingen Wang, and Mingli Song. Contrastive model inversion for data-free knowledge distillation. *arXiv preprint arXiv:2105.08584*, 2021.

Gongfan Fang, Kanya Mo, Xinchao Wang, Jie Song, Shitao Bei, Haofei Zhang, and Mingli Song. Up to 100x faster data-free knowledge distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 6597–6604, 2022.

Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pp. 3996–4003, 2020.

Cheng Gongye, Yukui Luo, Xiaolin Xu, and Yunsi Fei. Side-channel-assisted reverse-engineering of encrypted dnn hardware accelerator ip and attack surface exploration. In *2024 IEEE Symposium on Security and Privacy (SP)*, pp. 4678–4695. IEEE, 2024.

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

Gaurav Goswami, Nalini Ratha, Akshay Agarwal, Richa Singh, and Mayank Vatsa. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

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*, pp. 770–778, 2016a.

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

Lei Hsiung, Yun-Yun Tsai, Pin-Yu Chen, and Tsung-Yi Ho. Towards compositional adversarial robustness: Generalizing adversarial training to composite semantic perturbations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24658–24667, June 2023.

Bo Huang, Mingyang Chen, Yi Wang, Junda Lu, Minhao Cheng, and Wei Wang. Boosting accuracy and robustness of student models via adaptive adversarial distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24668–24677, 2023.

Hanxun Huang, Yisen Wang, Sarah Erfani, Quanquan Gu, James Bailey, and Xingjun Ma. Exploring architectural ingredients of adversarially robust deep neural networks. *Advances in neural information processing systems*, 34:5545–5559, 2021.

Xiaojun Jia, Yong Zhang, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Las-at: adversarial training with learnable attack strategy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13398–13408, 2022.

Xiaojun Jia, Jianshu Li, Jindong Gu, Yang Bai, and Xiaochun Cao. Fast propagation is better: Accelerating single-step adversarial training via sampling subnetworks. *IEEE Transactions on Information Forensics and Security*, 19:4547–4559, 2024a. doi: 10.1109/TIFS.2024.3377004.

Xiaojun Jia, Yong Zhang, Xingxing Wei, Baoyuan Wu, Ke Ma, Jue Wang, and Xiaochun Cao. Improving fast adversarial training with prior-guided knowledge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(9):6367–6383, 2024b. doi: 10.1109/TPAMI.2024.3381180.

Junho Kim, Byung-Kwan Lee, and Yong Man Ro. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159, 2021.

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

Ya Le and Xuan Yang. Tiny imagenet visual recognition challenge. *CS 231N*, 7(7):3, 2015.

Yao Li, Minhao Cheng, Cho-Jui Hsieh, and Thomas CM Lee. A review of adversarial attack and defense for classification methods. *The American Statistician*, 76(4):329–345, 2022.

Jeong-A Lim, Joohyun Lee, Jeongho Kwak, and Yeongjin Kim. Cutting-edge inference: Dynamic dnn model partitioning and resource scaling for mobile ai. *IEEE Transactions on Services Computing*, 2024.

Hou-I Liu, Marco Galindo, Hongxia Xie, Lai-Kuan Wong, Hong-Han Shuai, Yung-Hui Li, and Wen-Huang Cheng. Lightweight deep learning for resource-constrained environments: A survey. *ACM Computing Surveys*, 56(10):1–42, 2024.

Xingjun Ma, Yuhao Niu, Lin Gu, Yisen Wang, Yitian Zhao, James Bailey, and Feng Lu. Understanding adversarial attacks on deep learning based medical image analysis systems. *Pattern Recognition*, 110:107332, 2021.

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.

Paul Micaelli and Amos J Storkey. Zero-shot knowledge transfer via adversarial belief matching. *Advances in Neural Information Processing Systems*, 32, 2019.

Xuanlin Min, Wei Zhou, Rui Hu, Yinyue Wu, Yiran Pang, and Jun Yi. Lwuavdet: A lightweight uav object detection network on edge devices. *IEEE Internet of Things Journal*, 2024.

Payal Mittal. A comprehensive survey of deep learning-based lightweight object detection models for edge devices. *Artificial Intelligence Review*, 57(9):242, 2024.

Mark Sandler, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4510–4520, 2018.

Chunyu Sun, Chenye Xu, Chengyuan Yao, Siyuan Liang, Yichao Wu, Ding Liang, Xianglong Liu, and Aishan Liu. Improving robust fariness via balance adversarial training. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pp. 15161–15169, 2023.

Yuzheng Wang, Zhaoyu Chen, Dingkang Yang, Pinxue Guo, Kaixun Jiang, Wenqiang Zhang, and Lizhe Qi. Out of thin air: Exploring data-free adversarial robustness distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 5776–5784, 2024a.

Zeyu Wang, Xianhang Li, Hongru Zhu, and Cihang Xie. Revisiting adversarial training at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24675–24685, 2024b.

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*, pp. 11492–11501. PMLR, 2021.

Shaokai Ye, Kaidi Xu, Sijia Liu, Hao Cheng, Jan-Henrik Lambrechts, Huan Zhang, Aojun Zhou, Kaisheng Ma, Yanzhi Wang, and Xue Lin. Adversarial robustness vs. model compression, or both? In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 111–120, 2019.

Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8715–8724, 2020.

Xiaojian Yuan, Kejiang Chen, Wen Huang, Jie Zhang, Weiming Zhang, and Nenghai Yu. Data-free hard-label robustness stealing attack. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 6853–6861, 2024.

Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36:30390–30401, 2023.

Xinli Yue, Mou Ningping, Qian Wang, and Lingchen Zhao. Revisiting adversarial robustness distillation from the perspective of robust fairness. *Advances in Neural Information Processing Systems*, 36, 2024.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. *arXiv preprint arXiv:1605.07146*, 2016.

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*, pp. 7472–7482. PMLR, 2019.

Shiji Zhao, Ranjie Duan, Xizhe Wang, and Xingxing Wei. Improving adversarial robust fairness via anti-bias soft label distillation. *Advances in Neural Information Processing Systems*, 37:89125–89149, 2024.

Yuhang Zhou, Yushu Zhang, Leo Yu Zhang, and Zhongyun Hua. Derd: data-free adversarial robustness distillation through self-adversarial teacher group. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pp. 10055–10064, 2024.

Jianing Zhu, Jiangchao Yao, Bo Han, Jingfeng Zhang, Tongliang Liu, Gang Niu, Jingren Zhou, Jianliang Xu, and Hongxia Yang. Reliable adversarial distillation with unreliable teachers. *arXiv preprint arXiv:2106.04928*, 2021.

Kaijie Zhu, Xixu Hu, Jindong Wang, Xing Xie, and Ge Yang. Improving generalization of adversarial training via robust critical fine-tuning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4424–4434, 2023.

Bojia Zi, Shihao Zhao, Xingjun Ma, and Yu-Gang Jiang. Revisiting adversarial robustness distillation: Robust soft labels make student better. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 16443–16452, 2021.

A APPENDIX

A.1 THE PROOF OF CONJECTURE 1

Conjecture 1 is: by adjusting the weight of class sampling, the problem of robust unfairness is effectively alleviated.

Assumption 1. The robust risk function $\mathcal{R}_c(\theta)$ for each class c is differentiable and L-smooth (Bottou et al., 2018). That is, for any θ_1, θ_2, there exists a constant $L > 0$ such that:

$$\mathcal{R}_c(\theta_2) \leq \mathcal{R}_c(\theta_1) + \langle \nabla \mathcal{R}_c(\theta_1), \theta_2 - \theta_1 \rangle + \frac{L}{2}\|\theta_2 - \theta_1\|^2. \tag{12}$$

Theorem 1. *If the sampling weight p_k for the weak-robust class k satisfies $p_k > \frac{1}{C}$ at step t, then the expected risk reduction for the weak-robust class k is strictly greater under reweighting sampling than under uniform sampling, provided that the gradient alignment condition holds:*

$$\mathbb{E}[\mathcal{R}_k(\theta_{t+1}^{rew})] < \mathbb{E}[\mathcal{R}_k(\theta_{t+1}^{uni})]. \tag{13}$$

Proof. Consider the SGD update rule at step t. For uniform sampling, the expected update direction is $\mathbb{E}[\Delta\theta_{uni}] = -\eta\frac{1}{C}\sum_c \nabla\mathcal{R}_c(\theta_t)$. For reweighting sampling, the expected update direction is $\mathbb{E}[\Delta\theta_{rew}] = -\eta\sum_c p_c \nabla\mathcal{R}_c(\theta_t)$.

Using the L-smoothness on the weak-robust class k:

$$\mathcal{R}_k(\theta_{t+1}) \leq \mathcal{R}_k(\theta_t) + \langle \nabla\mathcal{R}_k(\theta_t), \Delta\theta \rangle + \frac{L}{2}\|\Delta\theta\|^2. \tag{14}$$

We compare the linear decrement term $\mathcal{D} = \langle \nabla\mathcal{R}_k(\theta_t), \Delta\theta \rangle$ for both strategies, neglecting the second-order term for sufficiently small η.

For uniform sampling:

$$\mathcal{D}_{uni} = -\eta\langle \nabla\mathcal{R}_k, \frac{1}{C}\sum_{c=1}^{C} \nabla\mathcal{R}_c \rangle = -\eta\left(\frac{1}{C}\|\nabla\mathcal{R}_k\|^2 + \frac{1}{C}\sum_{c\neq k}\langle \nabla\mathcal{R}_k, \nabla\mathcal{R}_c \rangle\right). \tag{15}$$

For reweighting sampling:

$$\mathcal{D}_{rew} = -\eta\langle \nabla\mathcal{R}_k, \sum_{c=1}^{C} p_c\nabla\mathcal{R}_c \rangle = -\eta\left(p_k\|\nabla\mathcal{R}_k\|^2 + \sum_{c\neq k} p_c\langle \nabla\mathcal{R}_k, \nabla\mathcal{R}_c \rangle\right). \tag{16}$$

We aim to prove $\mathcal{D}_{rew} < \mathcal{D}_{uni}$. Let $\phi_{k,c} = \langle \nabla\mathcal{R}_k, \nabla\mathcal{R}_c \rangle$ and we analyze the difference:

$$\Delta_{diff} = \mathcal{D}_{uni} - \mathcal{D}_{rew} = \eta\left((p_k - \frac{1}{C})\|\nabla\mathcal{R}_k\|^2 + \sum_{c\neq k}(p_c - \frac{1}{C})\phi_{k,c}\right). \tag{17}$$

Reweighting strategy assigns $p_k \gg \frac{1}{C}$ for the weak-robust class k, and $p_c < \frac{1}{C}$ for robust classes. Since $\|\nabla\mathcal{R}_k\|^2 > 0$ is the dominant gradient term for the weak-robust class, the term $(p_k - \frac{1}{C})\|\nabla\mathcal{R}_k\|^2$ is positive and large.

While gradient conflicts may exist ($\phi_{k,c} < 0$), the gradients of vulnerable classes typically exhibit larger magnitudes compared to robust classes due to their higher loss values in adversarial training(Madry et al., 2017). Specifically, the term $(p_k - \frac{1}{C})\|\nabla\mathcal{R}_k\|^2$ outweighs the residual terms. Thus, $\Delta_{diff} > 0$, implying $\mathcal{D}_{fer} < \mathcal{D}_{uni}$.

Consequently, the upper bound of the worst-class risk decreases more significantly under reweighting sampling:

$$\mathcal{R}_k(\theta_{t+1}^{rew}) \leq \mathcal{R}_k(\theta_t) + \mathcal{D}_{rew} + O(\eta^2) < \mathcal{R}_k(\theta_t) + \mathcal{D}_{uni} + O(\eta^2). \tag{18}$$

Then the Theorem 1 is proved.

A.2 The Proof of Conjecture 2

Conjecture 2 is: robust distillation on adversarial examples with uniformly distributed targets defend against attacks from different targets.

Definition 1. The standard adversarial risk makes the worst perturbation within a ball $\mathcal{B}_\epsilon(x) = \{x' : \|x' - x\|_p \leq \epsilon\}$:

$$\mathcal{R}_{std}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{\delta \in \mathcal{B}_\epsilon(0)} \mathcal{L}(f_\theta(x+\delta), y) \right]. \tag{19}$$

Adversarial attacks implicitly targets the decision boundary of the nearest incorrect class.

Definition 2. Adversarial examples with uniformly distributed targets enforce attacks towards a target class t sampled uniformly from $\mathcal{Y} \setminus \{y\}$. The risk objective is:

$$\mathcal{R}_{uni}(\theta) = \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\mathbb{E}_{t\sim\mathcal{U}(\mathcal{Y}\setminus\{y\})} \left[\max_{\delta \in \mathcal{B}_\epsilon(0)} \mathcal{L}_{target}(f_\theta(x+\delta), t) \right] \right], \tag{20}$$

where $\mathcal{L}_{target}$ encourages the model to classify $x + \delta$ as t.

Assumption 2. The event of a successful untargeted adversarial attack is the union of successful targeted attacks towards all other classes $k \neq y$. Specifically, let E be the error region $E = \{x' \in \mathcal{B}_\epsilon(x) : \arg\max f_\theta(x') \neq y\}$. We assume $E = \bigcup_{k\neq y} E_k$, where E_k is the region where the input is misclassified specifically as class k (Carlini & Wagner, 2017).

Theorem 2. *Minimizing the Uniform-Target Risk $\mathcal{R}_{uni}(\theta)$ minimizes an upper bound of the Standard Adversarial Risk $\mathcal{R}_{std}(\theta)$.*

Proof. The standard adversarial loss seeks the perturbation that maximizes the loss, corresponding to the model predicting an incorrect class k:

$$\max_\delta \mathcal{L}(f_\theta(x+\delta), y) \approx \max_{k\neq y} \left(\max_\delta \mathcal{L}_{target}(f_\theta(x+\delta), k) \right). \tag{21}$$

Let $\ell_k(\theta, x) = \max_{\delta\in\mathcal{B}_\epsilon} \mathcal{L}_{target}(f_\theta(x+\delta), k)$ be the worst loss towards specific target k. The standard robust risk is determined by the maximum component:

$$\mathcal{L}_{std}(\theta, x) = \max_{k\neq y} \ell_k(\theta, x). \tag{22}$$

The training on adversarial examples with uniformly distributed targets minimizes the expectation as follows:

$$\mathcal{L}_{uni}(\theta, x) = \frac{1}{C-1} \sum_{k\neq y} \ell_k(\theta, x). \tag{23}$$

Therefore, we have the following inequality:

$$\max_{k\neq y} \ell_k(\theta, x) \leq \sum_{k\neq y} \ell_k(\theta, x) = (C-1) \cdot \mathcal{L}_{uni}(\theta, x). \tag{24}$$

Taking the expectation over the data distribution $\mathcal{D}$:

$$\mathcal{R}_{std}(\theta) \leq (C-1) \cdot \mathcal{R}_{uni}(\theta). \tag{25}$$

This inequality proves that training on adversarial examples with uniformly distributed targets minimizes a upper bound of the standard adversarial risk. Then the Theorem 2 is proved.

A.3 The Whole Framework of FERD

This section provides the whole framework of our Fairness-Enhanced data-free adversarial Robust Distillation (FERD). The FERD framework is composed of three key components: a robustness-guided class reweighting strategy, the generation of Fairness-Aware Examples (FAEs), and the construction of Uniform-Target Adversarial Examples (UTAEs). The detailed algorithm description is outlined in Algorithm 1.

Algorithm 1: The Whole Framework of FERD

```
Input: A pre-trained robust teacher $f^T$, a student $f^S$ with parameter $\theta_s$, a generator $g$ with
       parameter $\theta_g$, distillation epochs $T$, the iterations of generator $g$ in each epoch $T_g$, the
       iterations of student $f^S$ in each epoch $T_s$, class sampling probability distribution $p$,
       learning rate for generator $\eta_g$ and student $\eta_s$.
Randomly initialize parameters $\theta_g$ and $\theta_s$ ;
Uniformly initialize $p$ ;
Initialize an empty dataset $D = \{\}$ ;
for number of iterations $T$ do
    //* Generation stage *//
    $z \sim \mathcal{N}(0, 1), y \sim p$ ;
    for number of iterations $T_g$ do
        $x_F \leftarrow g(z, y)$ ;                                   // Synthesize FAEs
        $\theta_g \leftarrow \theta_g - \eta_g \cdot \nabla_{\theta_g} \mathcal{L}_{gen}(f^T, f^S, x_F, y)$ ;
        $x_F^* \leftarrow select(x_F)$ ;
    $D \leftarrow D \cup \{x_F^*\}$ ;
    $p \leftarrow reweight(\mathcal{L}_{class})$ ;                          // Reweighting
    //* Robustness distillation stage *//
    for number of iterations $T_s$ do
        $x_F \leftarrow sample(D)$ ;
        $x_U \leftarrow UTAEs(x_F)$ ;                                 // Create UTAEs
        $\theta_s \leftarrow \theta_s - \eta_s \cdot \nabla_{\theta_s} \mathcal{L}_{stu}(f^T, f^S, x_F, x_U)$ ;
Output: A student $f^S$ with fair robustness
```

A.4 The Performance of the Teacher

In this section, we show the performance of the teacher models under various attacks. We select WideResNet-34-10 (Zagoruyko & Komodakis, 2016) trained by (Chen & Lee, 2024) on CIFAR-10 and CIFAR-100 (Krizhevsky et al., 2009), respectively. Meanwhile, we select PreActResNet-34 (He et al., 2016b) trained by TRADES (Zhang et al., 2019) on Tiny-ImageNet (Le & Yang, 2015). The performance is shown in Tab. 4.

Table 4: Robustness (%) of the teacher models.

Dataset	Model	Clean	FGSM	PGD	CW_∞	AA
CIFAR-10	WRN-34-10	86.55	64.96	62.97	59.72	57.29
CIFAR-100	WRN-34-10	64.32	39.51	37.89	35.17	31.16
Tiny-ImageNet	PARN-34	48.96	25.69	24.37	21.12	18.56

A.5 The Results on Tiny-ImageNet

Table 5: Result in average robustness(%) (Avg. ↑), worst-10(%) (Worst ↑), and normalized standard deviation (NSD ↓) on Tiny-ImageNet. The best results are **bolded** and the second best results are underlined respectively.

Method	Clean			FGSM			PGD			CW_∞		AA	
	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	NSD	Avg.	NSD
CMI	38.56	11.90	0.445	16.46	1.10	0.785	15.07	0.90	0.827	7.90	1.223	9.23	1.152
DeepInv	36.25	9.70	0.465	13.46	0.50	0.861	12.91	0.70	0.874	6.33	1.287	6.58	1.288
Fast	33.76	7.80	0.473	12.94	0.40	0.804	12.38	0.30	0.812	5.86	1.193	7.11	1.138
DFHL	23.13	1.10	0.719	6.87	0.00	1.368	5.78	0.00	1.535	4.04	1.778	4.95	1.654
FERD(Ours)	**39.64**	**12.30**	**0.439**	**18.00**	**2.10**	**0.722**	**17.13**	**1.40**	**0.743**	**9.39**	**1.046**	**11.24**	**1.0071**

Table 6: The performance of different reweighting strategies. The best results are **bolded**.

Method	Clean			FGSM			PGD		
	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
power	**80.54**	65.20	0.111	52.04	28.70	0.280	43.23	21.00	0.333
sig	80.16	64.10	0.111	51.29	27.70	0.290	**43.02**	21.20	0.350
linear	79.89	63.70	0.113	**52.06**	29.10	0.275	42.67	20.20	0.344
tanh	80.40	64.60	0.115	51.47	28.90	0.284	42.74	21.30	0.345
FERD	79.26	**65.90**	**0.105**	51.09	**31.30**	**0.237**	42.68	**23.50**	**0.298**

We conduct experiments on Tiny-ImageNet. We select PreActResNet-34 as the teacher and PreActResNet-18 as the student. For the settings, we keep consistent with the experiments on CIFAR-100. The results are shown in Tab. 5. The results prove the effectiveness of FERD, achieving state-of-the-art accuracy and worst-10% robustness on Tiny-ImageNet. Note that ZSKT is not applicable for large datasets and the distilled student performs extremely poorly on Tiny-ImageNet, so we do not compare with it on this dataset. The worst-10% robustness under both CW$_\infty$ and AA attacks is all 0, so we also do not compare on it.

From the results, it is observed that our proposed FERD achieves optimal performance under most adversarial attacks. Specifically, FERD achieves state-of-the-art average and worst-10% robustness under all attacks. For the average robustness, FERD is 1.54%, 2.06%, 1.49%, and 2.01% higher than the suboptimal method under the four adversarial attacks. For the worst-10% robustness, FERD also comprehensively improves 1.00% and 0.50% respectively. Meanwhile, FERD achieves the minimum on NSD under all attacks.

A.6 Comparison of Reweighting Strategies

To further demonstrate the superiority of our reweighting strategy, we conduct experimental comparisons with different reweighting functions (Yue et al., 2023). We use WideResNet-34-10 as the teacher, ResNet18 as the student and CIFAR-10 as dataset, while other settings keep consistent with main paper.

The reweighting strategies we compare are all based on a important metric: the least PGD steps (LPS), which represents the steps needed for perturbations of a benign example to become a adversarial example. The average LPS within a class κ_c serve as a metric of class-wise robustness , which is calculated as follows:

$$\kappa_c = \frac{1}{N_c}\sum_{j=1}^{N_c} \arg\min_{t\in[0,K]} \left(f^T\left(x_{adv}^{(t)}\right) \neq y\right), \tag{26}$$

where K is the PGD steps. The higher the value, the more robust the model is to this class, in which case they reduce the weight during training.

Based on the metric, we compare four reweighting strategies. The first is power-type function:

$$\omega_c = \frac{1}{\kappa_c^{\beta}}, \tag{27}$$

where β controls the smoothness of the reweighting function. The second strategy is linear-type function:

$$\omega_c = 1 - \frac{\kappa_c}{K+1}. \tag{28}$$

The third strategy is sigmoid-type function:

$$\omega_c = sigmoid\left(\lambda + 5 \times \left(1 - 2 \times \frac{\kappa_c}{K}\right)\right). \tag{29}$$

The last strategy is tanh-type function:

$$\omega_c = \frac{1 + \tanh\left(\lambda + 5 \times \left(1 - 2 \times \frac{\kappa_c}{K}\right)\right)}{2}, \tag{30}$$

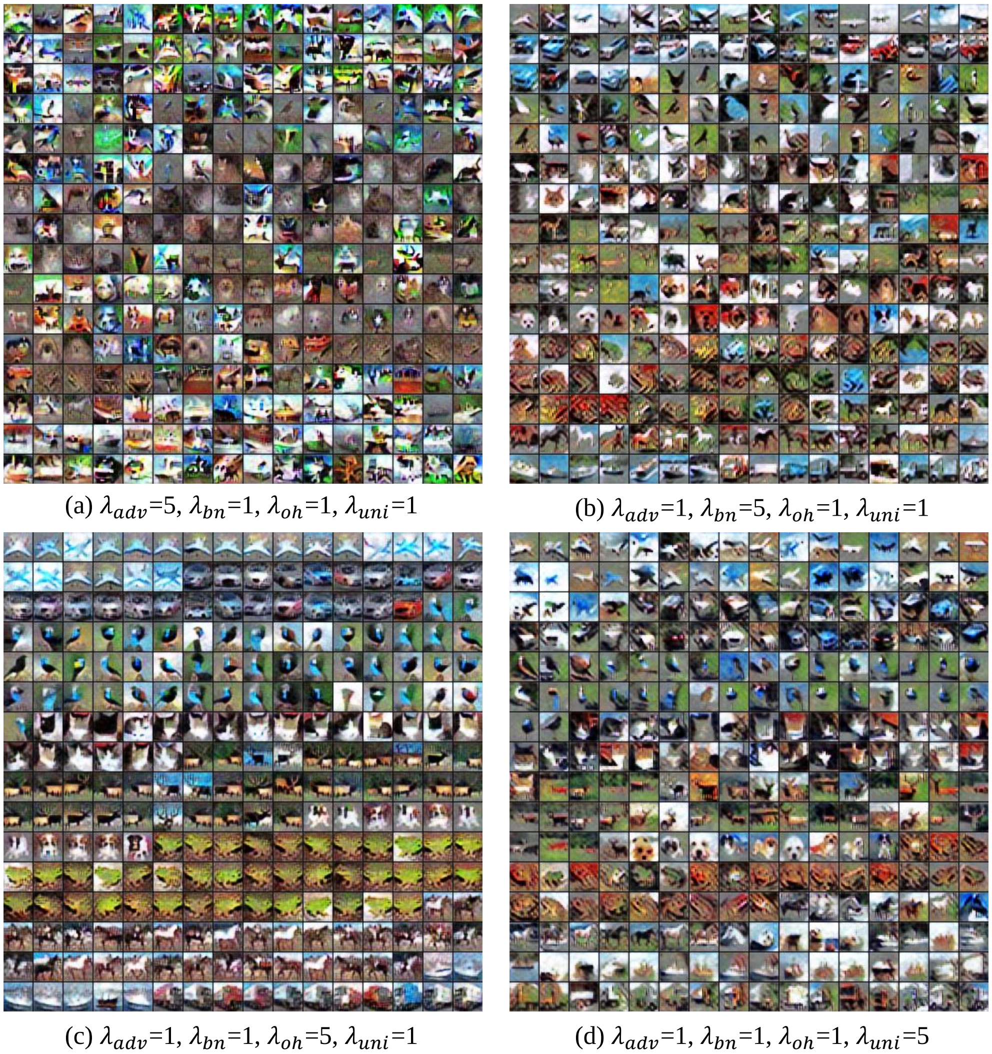
where λ is the parameter. For the above parameters, we set $K = 20$, $\beta = 2$ and $\lambda = 0$ respectively following (Yue et al., 2023). The experimental results are shown in Tab. 6.

We observe that the our reweighting strategy achieves the best results in the aspect of worst-class robustness, which demonstrates that it identifies the categories with poor robustness more effectively.

A.7 Hyper-Parameter Selection

In this section, we show the effect of parameters on the distillation results and illustrate how we determine the choice of parameters. We select WideResNet-34-10 as the teacher and ResNet-18 as the student to experiment on CIFAR-10.

λ_{adv}, λ_{bn}, λ_{oh} and λ_{uni} represent the weights of different loss terms in the loss function $\mathcal{L}_{gen}$, respectively. In order to explore the influence of different weights on the experimental results, we change the weights, respectively, for the experiments. The results are shown in Tab. 7. The corresponding synthetic samples are shown in Fig. 7.



(a) λ_{adv}=5, λ_{bn}=1, λ_{oh}=1, λ_{uni}=1

(b) λ_{adv}=1, λ_{bn}=5, λ_{oh}=1, λ_{uni}=1

(c) λ_{adv}=1, λ_{bn}=1, λ_{oh}=5, λ_{uni}=1

(d) λ_{adv}=1, λ_{bn}=1, λ_{oh}=1, λ_{uni}=5

Figure 7: Synthetic samples under different paramters.

From the results, we find that $\mathcal{L}_{adv}$ significantly improves the average robustness under all adversarial attacks, especially against FGSM and PGD. From the synthetic samples, we find that $\mathcal{L}_{bn}$ helps to improve the sample quality, as shown in Fig. 7 (b), making the student more applicable to the actual scenario. We also observe that appropriate $\mathcal{L}_{oh}$ supervision is necessary, but too high will

Table 7: Result in average robustness(%) (Avg. ↑), worst-class(%) (Worst ↑), and normalized standard deviation (NSD ↓) on different λ_{adv}, λ_{bn}, λ_{oh} and λ_{uni}. The best results are **bolded**, and the second best results are underlined.

Parameters				Clean			FGSM			PGD			CW$_\infty$			AA		
λ_{adv}	λ_{bn}	λ_{oh}	λ_{uni}	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD	Avg.	Worst	NSD
1	1	1	1	81.23	67.60	0.100	50.59	31.40	0.258	41.99	20.00	0.326	39.98	18.60	0.336	39.73	16.90	0.350
5	1	1	1	**81.32**	**70.00**	**0.089**	**53.26**	**34.40**	0.229	**44.67**	**23.90**	0.283	**42.30**	19.90	0.300	39.64	18.20	0.328
1	5	1	1	80.10	66.40	**0.089**	50.75	31.20	**0.204**	41.60	21.60	**0.282**	40.30	20.20	**0.290**	39.06	**19.30**	**0.296**
1	1	5	1	72.97	52.20	0.153	40.65	18.30	0.318	35.51	14.40	0.352	34.42	12.30	0.362	33.67	11.80	0.383
1	1	1	5	80.55	66.70	0.107	52.55	30.10	0.265	43.91	21.60	0.333	41.10	**20.80**	0.334	**40.74**	18.50	0.364
1	3	1	3	**80.51**	65.20	0.110	50.97	30.00	**0.235**	42.28	22.60	0.305	40.14	17.10	0.348	39.97	19.20	0.359
1	5	1	5	79.26	**65.90**	**0.105**	**51.09**	31.30	0.237	**42.68**	23.50	**0.298**	36.17	**19.00**	**0.338**	**40.12**	**20.80**	**0.325**
1	7	1	7	79.16	64.30	0.124	50.45	**31.90**	0.246	42.21	**24.55**	0.313	**41.07**	18.20	0.347	39.73	20.30	0.353
1	9	1	9	78.81	65.40	0.125	50.12	30.20	0.246	41.56	21.90	0.314	39.25	18.10	0.341	38.9	18.90	0.344

Table 8: Robustness of DFKD and corresponding DFRD methods. RN-18 and MN-V2 are abbreviations of ResNet-18 and MobileNet-V2 respectively. The value in parentheses is the change value after transforming into DFRD.

Student	Method	Clean	FGSM	PGD-20	CW$_\infty$	AA
RN-18	ZSKT	**66.45**	29.60	**29.50**	14.82	16.43
	ZSKT+RSLAD	65.48 (–0.97)	**29.80 (+0.02)**	29.35 (–0.15)	**15.84 (+1.02)**	**17.16 (+0.73)**
	CMI	**76.20**	32.34	28.30	20.35	26.53
	CMI+RSLAD	68.29 (–7.91)	**33.34 (+1.00)**	**28.97 (+0.67)**	**21.74 (+1.39)**	**26.85 (+0.34)**
	DeepInv	**74.20**	34.98	27.38	21.08	22.84
	DeepInv+RSLAD	71.22 (–2.98)	**38.60 (+3.62)**	**27.83 (+0.45)**	**27.92 (+6.84)**	**25.19 (+2.35)**
	Fast	**78.51**	41.04	37.19	24.38	27.95
	Fast+RSLAD	72.20 (–6.31)	**43.96 (+1.98)**	**39.60 (+2.41)**	**28.03 (+3.65)**	**33.24 (+5.29)**
MN-V2	ZSKT	52.64	22.35	22.69	12.70	14.41
	ZSKT+RSLAD	**54.69 (+2.05)**	**23.62 (+1.27)**	**22.76 (+0.7)**	**12.78 (+0.08)**	**14.85 (+0.44)**
	CMI	**67.95**	**35.43**	17.46	**23.22**	6.87
	CMI+RSLAD	60.70 (–7.25)	27.90 (–7.53)	**19.61 (+2.15)**	20.10 (–3.12)	**18.09 (+11.22)**
	DeepInv	**66.94**	**34.98**	15.93	23.20	6.19
	DeepInv+RSLAD	62.77 (–4.15)	31.66 (–3.32)	**19.80 (+3.87)**	**23.60 (+0.4)**	**15.39 (+9.2)**
	Fast	**69.81**	**30.17**	23.01	15.43	14.59
	Fast+RSLAD	58.60 (–11.21)	29.91 (–0.26)	**23.26 (+0.25)**	**18.20 (+2.77)**	**18.35 (+3.76)**

suppress sample diversity and lead to performance degradation. For $\mathcal{L}_{uni}$, it helps to improve the robustness performance under the strongest attack (such as AA) and alleviate the robust unfairness problem. Here, considering the practicality and robust fairness of the model, we focus on λ_{bn} and λ_{uni} as parameters tuning. From the bottom four rows of the Tab. 7, we find that it has the best performance in most cases when λ_{bn} and λ_{uni} are both equal to 5, so we set the hyperparameter λ_{adv}=1, λ_{bn}=5, λ_{oh}=1, λ_{uni}=5 finally.

A.8 Transforming DFKD into DFRD

For most Data-Free Knowledge Distillation (DFKD), they do not involve robustness, in which case we transform them into Data-Free Robust Distillation (DFRD) by adding adversarial noise to synthetic samples and then take synthetic samples and adversarial examples for robust distillation. Here, we use PGD to generate adversarial examples and then apply the same distillation training loss as RSLAD (Zi et al., 2021) to transform them into DFRD. Here, we set step size to 2/255 and attack iteration to 10. The results are shown in Tab. 8.

We observe that incorporating RSLAD into the distillation leads to a slight decrease in clean accuracy, but it achieves a notable improvement in robustness, particularly under the stronger attacks. This trade-off is consistent with findings in prior adversarial training (Zhang et al., 2019). Therefore, we think that this transformation approach is reasonable and effective.

A.9 Additional Ablation Study

A.9.1 Intermediate Layer l

Table 9: Ablation study of different Intermediate Layer l.

Layer	Clean		FGSM		PGD-20	
	Avg.	Worst	Avg.	Worst	Avg.	Worst
block1	78.50	64.80	48.83	30.50	40.48	23.10
block2	78.91	64.30	49.37	29.50	41.16	21.90
block3	**79.26**	**65.90**	**51.09**	**31.30**	**42.68**	**23.50**

In FAEs generation process, we distill non-robust features from the output in the intermediate layer l of the teacher and impose uniform constraints, ensuring that the synthetic samples have an equal tendency to different targets when generating adversarial samples. In this section, we research the sensitivity of FERD's robustness performance to different layers. We distill non-robust features from the outputs of the penultimate, second and third residual modules, respectively. The results are shown in Tab. 9.

We observe that the penultimate residual module performs best in all aspects. Both the robustness and the fairness have achieved the best results. This further demonstrates that the higher layer contains high-level feature information, which is effective for distilling non-robust feature.

A.9.2 Distillation Epoch e

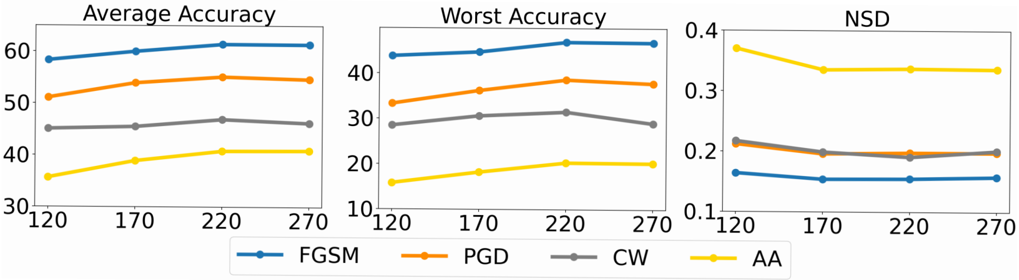


Figure 8: Metric values under different epochs. The horizontal axis is epoch and the vertical axis is metric value.

To explore the impact of the epoch on the student performance, we evaluate the distilled student under different epoch settings. The results are shown in Fig. 8.

With the increase of the epoch, both the average accuracy and the worst-class robustness of the student exhibit a generally increasing trend, reaching a peak at round 220 and then floating slightly, indicating that the robustness is optimal at this time. At the same time, the NSD remains almost constant after 170 epochs. Therefore, it is reasonable to select 220 as the final distillation epochs.

A.10 Fairness of FAEs and UTAEs

To further validate the rationality and effectiveness of our proposed FAEs and UTAEs, we conduct a comparative analysis by visualizing the confusion matrix corresponding to different types of adversarial examples. Specifically, we investigate the impact of UTAEs generated by FAEs, adversarial samples generated by FAEs, and adversarial samples generated by the original dataset, respectively. The results are shown in Fig. 9.

By comparing Fig. 9 (a) and (b), we find that the adversarial targets of our designed FAEs is more uniform than the original data when constructing adversarial examples. This suggests that FAEs contribute to a more balanced adversarial target distribution when crafting perturbations. Such uniformity is particularly beneficial in the context of adversarial training and knowledge distillation,

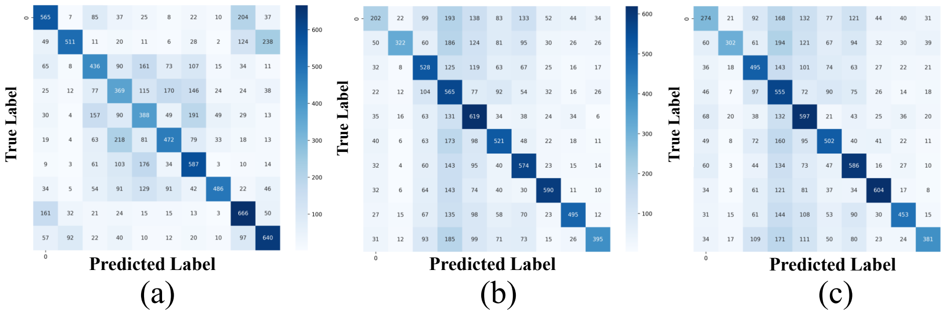


Figure 9: Confusion matrix of different samples and corresponding adversarial examples. The horizontal axis denotes predicted labels, and the vertical axis denotes true labels. Darker colors indicate a higher number of samples predicted as the corresponding class. (a): Adversarial examples generated by PGD using data from CIFAR-10. (b): Adversarial examples generated by PGD using FAEs. (c): UTAEs generated by FAEs.

as it facilitates the exposure of the student to a broader class-wise perturbations. By comparing Fig. 9 (b) and (c), we observe that the accuracy of the model is weaker on UTAEs. The increased misclassification rate suggests that UTAEs are capable of identifying and exploiting a wider range of class-specific vulnerabilities, including those of weakly robust classes that may not be adequately targeted by conventional adversarial examples. Therefore, incorporating UTAEs into the distillation leads to a more comprehensive robustness enhancement, as the model is encouraged to improve its resilience across all classes rather than overfitting to a subset of dominant or easily perturbed classes.

A.11 Synthetic Samples from different DFRD

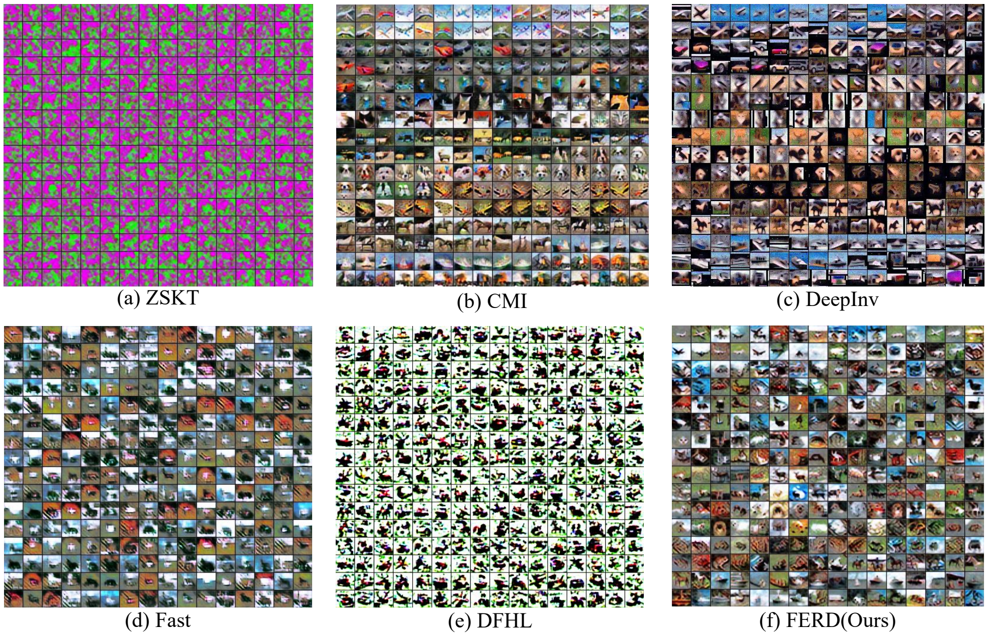

(a) ZSKT
(b) CMI
(c) DeepInv
(d) Fast
(e) DFHL
(f) FERD(Ours)

Figure 10: Inverted samples from different DFRD.

This section shows samples synthesized by different DFRD methods. From the visual results, it is evident that our FERD generates samples that are both more realistic and more diverse. This means that the method is suitable for practical scenarios.

A.12 LIMITATIONS

Although FERD mitigates the robust fairness by introducing a robust-guided class reweighting strategy, the class imbalance may still affect the final results in some extreme cases. For example, if some classes are very rare in the original data, the problem of unfairness of robustness may not be completely solved by just adjusting the sample proportion. Moreover, the generator needs to optimize to multiple loss functions to synthesize high-quality samples, while the student also needs to perform robust distillation. These processes involve complex optimizations and a large number of iterations, which lead to high computational cost.

A.13 THE USE OF LARGE LANGUAGE MODELS

In the writing process for this paper, we employed a large language model for text refinement and grammatical corrections to enhance overall readability and precision. We clarify that the model's role was strictly advisory and confined to language aspects. The background of the study, the design of the experiments, and the analysis of all results represent the independent work of the authors.