Supplementary Material

Tejaswini Medi[1], Steffen Jung[1,2], Margret Keuper[1,2]
[1]University of Mannheim, [2]MPI for Informatics, Saarland Informatics Campus
tejaswini.medi@uni-mannheim.de

Appendix

In this document, we provide the additional details and experimental results of our approach. The supplementary material is structured as follows:

(A) Ablation Study

- Fairness in adversarially trained models.
- Ablations on uniform sampling of targets.
- Ablations on constant perturbation margin during AT.
- Ablations on epsilon scaling.

(B) Additional Experimental Results

- Experimental results with PRN-18 on CIFAR-10 over mutliple seeds.
- Experimental results with XCiT-S12 on CIFAR-10.
- Overall accuracy results on common corruptions.

A. Ablation Study

A.1. Fairness in Adversarially Trained Models.

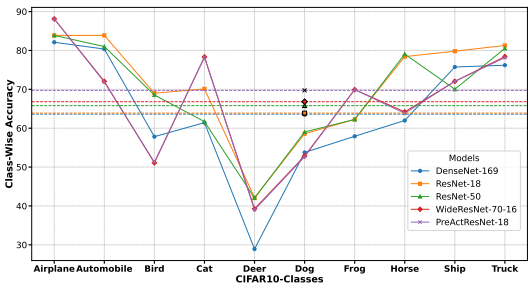

Figure 1. Class-wise accuracy C_{acc} on the CIFAR-10 dataset using adversarially trained models.

We conducted an experiment on the CIFAR-10 dataset using robust models trained adversarially, sourced from a standard adversarial benchmark [1]. Figure 1 shows the class-wise robustness accuracies (C_{acc}) on the CIFAR-10 dataset. It illustrates the class-wise accuracy of different adversarially trained models on clean validation samples from the CIFAR-10 dataset. The disparity in class-wise accuracies are clearly visible, which means the adversarially trained models are not fair. The dotted lines in the Figure 1 represent the overall accuracies of the respective models. The overall accuracy of each model allows us to categorize classes into two groups: Easy classes and Hard classes. This categorization is based on whether a class has a class-wise accuracy above or below the model's average overall accuracy. Easy classes are easy to predict and hard classes are often hard to distinguish. This pattern remains consistent across different model architectures. Therefore training a fair robust model is very important. Our approach aims to work on this robust fairness issue.

A.2. Ablation on Uniform Sampling of Targets

In this section, we compare our approach, FAIR-TAT, which samples targets based on the class-wise false positive prior distribution, to a uniform target sampling method during targeted adversarial training. This comparison demonstrates the effectiveness of using class-wise false positive information in target sampling, highlighting its benefits in preserving overall accuracy and improving the worst-class accuracy.

We refer to our approach, which employs class-wise false positive target selection, as FAIR-TAT(CTP). Conversely, we denote the approach that uses uniform target sampling during training as FAIR-TAT(UTP) in Table 1. The results in the table present PGD evaluations using PRN-18 on the CIFAR-10 dataset. It should also be noted that the perturbation margins for the classes are customized as described in [3] during the training of these approaches. From Table 1, it is clear that sampling targets based on class-wise false positive scores improves model robustness and fairness compared to uniform sampling of targets. This improvement arises from assigning more weight to classes that are more prone to misclassification during training. Thus, our results demonstrate the effectiveness of incorporating class-wise false positive scores when sampling targets for targeted adversarial training scenario of our approach. We observe the same trend with different weight averaging techniques EMA and CFA as well.

Table 1. Comparison of PGD evaluations of FAIR-TAT framework with its variants.

Method	PRN-18 (Clean Accuracy)		PRN-18 (Robust Accuracy)	
	Overall	Worst Class	Overall	Worst Class
FAIR-TAT (UTS)	85.0 ± 0.2	70.9 ± 2.3	46.2 ± 0.5	18.4 ± 0.9
FAIR-TAT (CTS) ◆	$\mathbf{85.5 \pm 0.3}$	$\mathbf{71.3 \pm 1.8}$	$\mathbf{46.6 \pm 0.6}$	$\mathbf{19.7 \pm 1.1}$
FAIR-TAT (UTS) + EMA	85.1 ± 0.6	72.2 ± 0.8	47.4 ± 0.3	21.2 ± 0.5
FAIR-TAT (CTS) + EMA ◆	$\underline{\mathbf{86.0 \pm 0.4}}$	$\underline{\mathbf{73.0 \pm 0.7}}$	$\mathbf{47.7 \pm 0.3}$	$\mathbf{22.1 \pm 0.9}$
FAIR-TAT (UTS) + CFA	83.0 ± 0.4	70.2 ± 1.3	47.8 ± 0.2	23.3 ± 0.8
FAIR-TAT (CTS) + CFA ◆	$\mathbf{84.8 \pm 0.3}$	$\mathbf{72.0 \pm 0.9}$	$\underline{\mathbf{48.3 \pm 0.5}}$	$\underline{\mathbf{24.6 \pm 1.0}}$

Table 2. Comparison of PGD evaluations of FAIR-TAT framework with constant perturbation margin during training.

Method	PRN-18 (Clean Accuracy)		PRN-18 (Robust Accuracy)	
	Overall	Worst Class	Overall	Worst Class
AT	84.0 ± 0.2	66.4 ± 1.3	$\mathbf{45.6 \pm 0.2}$	16.9 ± 1.1
FAIR-TAT (UTS)	87.3 ± 0.1	$\mathbf{73.8 \pm 2.9}$	43.6 ± 0.7	$\mathbf{19.0 \pm 1.9}$
FAIR-TAT (CTS)	$\mathbf{87.3 \pm 0.3}$	73.2 ± 1.7	43.5 ± 0.6	17.7 ± 2.7
AT + EMA	84.7 ± 0.3	68.5 ± 0.9	$\mathbf{47.6 \pm 0.1}$	18.6 ± 0.6
FAIR-TAT (UTS) + EMA	$\underline{\mathbf{88.0 \pm 1.5}}$	$\underline{\mathbf{75.9 \pm 1.0}}$	44.5 ± 0.5	$\mathbf{20.1 \pm 0.4}$
FAIR-TAT (CTS) + EMA	87.8 ± 0.2	74.9 ± 0.9	44.6 ± 0.3	19.0 ± 1.2
AT + CFA	84.6 ± 0.3	69.5 ± 1.5	$\underline{\mathbf{48.3 \pm 0.2}}$	21.8 ± 0.6
FAIR-TAT (UTS) + CFA	$\mathbf{87.3 \pm 0.4}$	$\mathbf{74.8 \pm 1.1}$	45.9 ± 0.5	$\underline{\mathbf{23.0 \pm 2.4}}$
FAIR-TAT (CTS) + CFA	86.0 ± 3.2	73.3 ± 7.4	44.8 ± 0.3	22.3 ± 1.4

A.3. Ablations on Constant Perturbation Margin during AT.

In this section, we assess the efficacy of our approach when the perturbation margin ϵ is kept constant during the training process. For a fair comparison, we use vanilla adversarial training (AT) as a baseline, which also employs a constant perturbation margin on the CIFAR-10 dataset using the PRN-18 architecture. Additionally, we evaluate our method under different target sampling schemes, as well as conventional AT, using various weight averaging schemes.

Table 2 presents the results of our approach under two conditions: Uniform Target Sampling (UTS) and Class-wise False Positive Target Sampling (CTS), compared to the baseline AT. The results indicate that FAIR-TAT (UTS) with a constant perturbation margin ϵ performs well in terms of fairness, showing improved worst-class accuracies compared to the baseline. Furthermore, FAIR-TAT (UTS) and FAIR-TAT (CTS) increases the overall clean accuracy, although the overall robustness decreases significantly when compared to the baseline. This observation holds across different weight averaging schemes.

For results using a customized perturbation margin, refer to Table 1. From Table 1, it is evident that adjusting the perturbation margin during targeted adversarial training allows for adjusting attack strengths across different classes. Harder classes are assigned a smaller margin, while easier classes receive a larger margin, as described in [3]. Interestingly, FAIR-TAT with uniform target sampling performs slightly better with a constant ϵ than class-wise false positive target sampling with a constant ϵ. FAIR-TAT (UTS) also demonstrates superior fairness in terms of clean sample accuracy when considering all versions of our method.

However, the combination of customized margins and class-wise false positive target sampling (CTP) in FAIR-TAT offers better trade-offs between robustness and fairness, considering both robust and clean samples. Thus, comparing Table 1 and Table 2, we conclude that the combination of FAIR-TAT (CTP), customized perturbation margins, and effective weight averaging leads to a robust and fair classifier.

A.4. Ablations on Epsilon Scaling.

It is well known that targeted adversaries are weaker than untargeted adversaries. Thus, we adjust the perturbation margin for each class during adversarial training (AT) using the update equation:

$$\epsilon_k \leftarrow (\lambda_1 + r_k)\epsilon.$$

This perturbation margin update for each class, ϵ_k, depends on the robust class-wise accuracy performance, r_k, so that the perturbation margin for the individual class adjusts according to the model's performance on this class. This ensures the strength of the adversaries for each class during

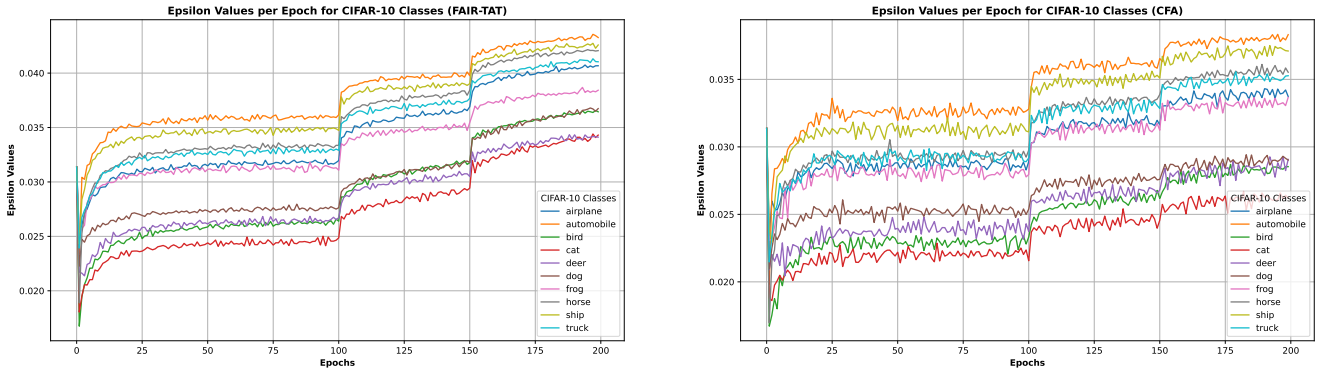

Figure 2. Epsilon scaling during AT for FAIR-TAT and CFA method respectively on 10 classes of CIFAR10 in order.

AT. The Figure 2 showcases the class-wise perturbation margin for both our approach FAIR-TAT and CFA (which uses an adaptable class-wise perturbation margin during untargeted adversarial training) on the CIFAR10 dataset using the PRN-18 architecture. This provides intuition regarding the strength of adversaries for both targeted and untargeted setups during the dynamic training scheme. The perturbation margin, ϵ, for hard classes is lower than for easier classes, and the ϵ values for the targeted scenario are greater than those for the untargeted scenario during the training. This balance ensures that the attack strength is balanced considering the targeted and untargeted adversaries.

Furthermore, the variance introduced by including targeted adversaries in the adversarial training of our approach is enhanced when we sample the targets from the prior distribution of class-wise false positive scores. This additional degree of freedom in the training process is intentional and helps to guide the sampling towards vulnerable classes during adversarial training, ultimately improving model fairness.

B. Additional Experimental Results

B.1. Experimental Results with PRN-18 on CIFAR-10 over Mutliple Seeds.

In this section, we present the evaluations of our approach using PRN-18 on CIFAR-10 over 5 random seeds. Table 4 highlights the PGD evaluations, comparing our method, adapted to other approaches, to baseline methods on CIFAR-10. From our observations, it is clear that our approach, utilizing the weight averaging technique from CFA [3], provides enhanced model fairness, as it incorporates a fairness-aware weight averaging scheme.

Similarly, Table 5 presents the evaluations using standard AutoAttack and a black box attack named Square Attack, comparing our method against baseline approaches on CIFAR-10 over 5 seeds. The results follow the same trend, with our approach demonstrating improved fairness through the appropriate weight averaging scheme, while maintaining overall robust accuracy. Also, our approach provides better performance considering robustness and fairness on clean samples.

B.2. Experimental Results on XCiT-S12 on CIFAR10

Most of the works related to fairness consider CNN based models, as they are easy to train adversarially compared to transformer based models. Therefore the fairness is often validated on CNN based models. To further validate the efficacy of FAIR-TAT, we use XCiT-S12, a transformer based model sourced from robustness benchmark. In Table 3, we compare our results to the model trained using vanilla adversarial training. We train XCiT following the recipe of adversarial fine-tuning as mentioned in [2]. Table 3 shows the PGD evaluations of FAIR-TAT vs baseline adversarial training on CIFAR-10. Evaluations in the table indicate that FAIR-TAT is fair on transformer based models as well.

Table 3. Comparison of PGD evaluations of FAIR-TAT framework using XCiT on CIFAR10. Our method is marked with ◆.

Method	XCiT-S12 (Clean Accuracy)		XCiT-S12 (Robust Accuracy)	
	Overall	Worst Class	Overall	Worst Class
AT	89.6	79.8	**62.1**	30.2
FAIR-TAT ◆	**92.7**	**80.3**	61.7	**31.7**

Table 4. Comparison of PGD evaluations of FAIR-TAT framework with other methods focused on model fairness using PRN-18 on CIFAR-10 over 5 seeds. Our method is marked with ◆.

Method	PRN-18 (Clean Accuracy)		PRN-18 (Robust Accuracy)	
	Overall	Worst Class	Overall	Worst Class
AT	83.5 ± 0.3	66.8 ± 2.7	$\mathbf{48.2 \pm 0.3}$	19.6 ± 2.0
FAIR-TAT ◆	$\mathbf{85.5 \pm 0.2}$	$\mathbf{71.3 \pm 2.1}$	46.6 ± 0.2	$\mathbf{19.7 \pm 2.0}$
AT + EMA	83.8 ± 0.2	66.6 ± 1.3	$\mathbf{49.6 \pm 0.2}$	21.0 ± 0.5
FAIR-TAT + EMA ◆	$\mathbf{86.0 \pm 0.2}$	$\mathbf{73.0 \pm 0.5}$	47.7 ± 0.2	$\mathbf{21.7 \pm 0.5}$
AT + CFA	83.8 ± 0.2	68.1 ± 1.1	$\mathbf{50.1 \pm 0.2}$	22.8 ± 1.4
FAIR-TAT + CFA◆	$\mathbf{84.8 \pm 1.1}$	$\mathbf{72.0 \pm 3.4}$	48.3 ± 1.1	$\mathbf{24.6 \pm 2.4}$
TRADES	$\mathbf{82.0 \pm 0.4}$	64.7 ± 1.4	52.9 ± 0.4	$\mathbf{25.9 \pm 1.7}$
FAIR-TAT + TRADES ◆	81.7 ± 0.1	$\mathbf{67.7 \pm 1.1}$	$\mathbf{52.3 \pm 1.1}$	25.5 ± 0.7
TRADES + EMA	$\mathbf{82.4 \pm 0.1}$	65.3 ± 0.7	$\underline{\mathbf{53.8 \pm 0.1}}$	25.3 ± 0.6
FAIR-TAT + TRADES + EMA ◆	82.1 ± 0.2	$\mathbf{68.4 \pm 0.8}$	53.1 ± 0.2	$\mathbf{26.8 \pm 0.8}$
TRADES + CFA	$\mathbf{82.3 \pm 0.2}$	65.4 ± 0.7	$\underline{\mathbf{53.7 \pm 0.2}}$	25.2 ± 0.2
FAIR-TAT + TRADES + CFA ◆	82.1 ± 0.2	$\mathbf{68.6 \pm 1.0}$	53.1 ± 0.2	$\underline{\mathbf{26.9 \pm 0.8}}$
FAT	84.8 ± 0.3	69.3 ± 1.1	$\mathbf{48.0 \pm 0.3}$	18.7 ± 0.8
FAIR-TAT + FAT ◆	$\mathbf{86.6 \pm 0.3}$	$\underline{\mathbf{77.2 \pm 0.9}}$	46.1 ± 0.3	$\mathbf{22.7 \pm 0.8}$
FAT + EMA	85.1 ± 0.2	69.5 ± 1.4	$\mathbf{49.0 \pm 0.2}$	19.5 ± 0.9
FAIR-TAT + FAT+ EMA ◆	$\underline{\mathbf{86.8 \pm 0.2}}$	$\mathbf{75.4 \pm 1.4}$	47.7 ± 0.2	$\mathbf{22.3 \pm 0.9}$
FAT + CFA	85.0 ± 0.6	71.0 ± 1.2	$\mathbf{51.0 \pm 0.6}$	23.4 ± 0.3
FAIR-TAT + FAT + CFA ◆	$\mathbf{86.1 \pm 0.2}$	$\mathbf{74.9 \pm 1.1}$	48.2 ± 0.2	$\mathbf{24.4 \pm 5.2}$
FRL	82.8 ± 0.1	71.4 ± 2.4	45.7 ± 0.3	24.4 ± 1.0
FRL + EMA	83.6 ± 0.3	69.5 ± 0.7	46.3 ± 0.2	24.8 ± 0.4
BAT + TRADES	86.5 ± 0.1	73.4 ± 1.4	49.8 ± 0.2	22.1 ± 0.9
WAT + TRADES	81.2 ± 0.3	65.9 ± 2.0	47.1 ± 0.3	26.3 ± 0.9
Clean Training	94.0 ± 0.3	79.4 ± 2.4	2.7 ± 0.7	0

Table 5. Comparison of AutoAttack and Squares evaluations of FAIR-TAT framework with other methods focused on model fairness using PRN-18 on CIFAR-10 over 5 seeds. Our method is marked with ◆.

Method	PRN-18 (AutoAttack)		PRN-18 (Squares)	
	Overall	Worst Class	Overall	Worst Class
AT	$\mathbf{45.7 \pm 0.3}$	15.4 ± 1.6	51.3 ± 0.2	19.4 ± 1.3
FAIR-TAT ◆	45.0 ± 0.2	$\mathbf{18.7 \pm 1.7}$	$\mathbf{51.8 \pm 0.1}$	$\mathbf{26.7 \pm 0.7}$
AT + EMA	$\mathbf{45.6 \pm 0.2}$	15.4 ± 1.8	51.7 ± 0.2	19.7 ± 0.4
FAIR-TAT + EMA ◆	45.0 ± 0.3	$\mathbf{18.8 \pm 0.4}$	$\mathbf{51.8 \pm 0.3}$	$\mathbf{26.7 \pm 0.7}$
AT + CFA	47.4 ± 0.8	19.3 ± 0.3	51.7 ± 0.2	19.9 ± 0.4
FAIR-TAT+CFA ◆	$\mathbf{47.0 \pm 1.3}$	$\underline{\mathbf{24.8 \pm 2.1}}$	$\mathbf{52.6 \pm 0.8}$	$\underline{\mathbf{31.3 \pm 1.2}}$
TRADES	$\mathbf{49.8 \pm 0.2}$	18.7 ± 1.1	$\mathbf{54.0 \pm 0.3}$	22.4 ± 0.2
FAIR-TAT + TRADES ◆	48.6 ± 0.1	$\mathbf{20.8 \pm 1.3}$	52.6 ± 0.2	$\mathbf{26.1 \pm 0.4}$
TRADES + EMA	$\mathbf{49.8 \pm 0.2}$	18.6 ± 0.4	$\mathbf{54.0 \pm 0.3}$	22.4 ± 0.2
FAIR-TAT + TRADES + EMA ◆	48.6 ± 0.3	$\mathbf{21.1 \pm 0.6}$	52.6 ± 0.2	$\mathbf{26.1 \pm 0.4}$
TRADES + CFA	$\underline{\mathbf{50.3 \pm 0.1}}$	21.3 ± 0.3	$\underline{\mathbf{54.2 \pm 0.3}}$	22.6 ± 0.4
FAIR-TAT + TRADES + CFA ◆	49.8 ± 0.3	$\mathbf{24.0 \pm 0.8}$	53.0 ± 0.5	$\mathbf{25.8 \pm 0.9}$
FAT	44.2 ± 0.3	16.0 ± 1.1	50.8 ± 0.2	20.8 ± 0.1
FAIR-TAT + FAT ◆	$\mathbf{44.6 \pm 0.2}$	$\mathbf{17.9 \pm 0.8}$	$\mathbf{51.5 \pm 0.4}$	$\mathbf{26.3 \pm 0.3}$
FAT + EMA	44.2 ± 0.1	15.9 ± 0.9	50.8 ± 0.2	20.8 ± 0.1
FAIR-TAT + FAT+ EMA ◆	$\mathbf{44.6 \pm 0.2}$	$\mathbf{18.1 \pm 0.7}$	$\mathbf{51.5 \pm 0.4}$	$\mathbf{26.3 \pm 0.3}$
FAT + CFA	$\mathbf{49.4 \pm 0.1}$	22.6 ± 0.8	$\mathbf{53.4 \pm 0.1}$	24.1 ± 0.5
FAIR-TAT + FAT + CFA ◆	44.3 ± 0.3	$\mathbf{22.1 \pm 2.1}$	49.0 ± 0.6	$\mathbf{24.4 \pm 0.3}$
FRL	44.0 ± 0.2	23.2 ± 1.2	49.7 ± 0.3	24.6 ± 0.9
FRL + EMA	44.2 ± 0.3	23.9 ± 0.4	50.8 ± 0.6	24.9 ± 0.3
BAT + TRADES	45.9 ± 0.4	18.7 ± 1.2	52.3 ± 0.4	23.8 ± 0.7
WAT + TRADES	47.1 ± 0.3	24.5 ± 1.1	51.7 ± 0.3	25.6 ± 0.9

Table 6. Overall accuracies of FAIR-TAT method on common corruptions along with combination of TRADES (T) and FAT(F) approaches using PreActResNet-18 on CIFAR-10C dataset.

Corruption Type	FAIR-TAT	FAIR-TAT (EMA)	FAIR-TAT (CFA)	FAIR-TAT (T)	FAIR-TAT (T+EMA)	FAIR-TAT (T+CFA)	FAIR-TAT (F)	FAIR-TAT (F+EMA)	FAIR-TAT (F+CFA)
gaussian_noise	80.6	81.6	79.0	78.5	76.3	76.3	82.2	82.5	73.3
shot_noise	81.7	82.6	80.1	79.3	77.5	77.3	83.1	83.6	74.5
speckle_noise	81.5	82.6	80.2	79.3	77.3	77.3	82.9	83.6	74.1
impulse_noise	71.5	72.7	72.3	74.9	72.2	72.2	73.0	74.1	66.5
defocus_blur	80.3	80.6	76.5	76.8	76.6	76.6	81.0	81.2	68.1
gaussian_blur	78.0	78.3	73.7	74.6	74.5	74.4	78.5	78.7	64.4
glass_blur	76.3	76.4	71.9	73.2	73.1	73.0	76.9	77.2	62.4
motion_blur	79.1	79.5	75.2	75.6	75.7	75.7	80.2	80.5	65.5
zoom_blur	78.8	79.5	77.2	75.2	76.3	76.2	80.1	80.3	72.0
snow	56.1	56.4	50.6	53.4	53.0	53.0	58.3	57.7	39.6
frost	82.6	82.7	79.9	77.0	78.2	78.2	83.1	83.0	74.3
fog	41.0	40.5	35.6	38.4	38.1	38.0	42.5	41.6	28.7
brightness	79.2	79.8	75.9	76.0	75.6	75.5	80.4	80.6	67.1
contrast	83.4	83.9	80.8	80.0	79.6	79.5	84.3	84.5	74.1
elastic_transform	82.9	83.6	80.6	80.2	79.6	79.5	84.1	84.4	74.7
pixelate	80.0	81.0	79.0	78.3	76.8	76.8	81.7	81.9	73.6
jpeg_compression	81.6	82.3	80.6	78.6	77.6	77.6	82.6	83.0	76.0
defocus_blur	75.8	75.8	71.3	68.1	70.8	70.8	76.2	75.8	63.7

Table 7. Overall accuracies of baselines on common corruptions along with combination of TRADES (T) and FAT(F) approaches using PreActResNet-18 on CIFAR-10C dataset.

Corruption Type	AT	AT(EMA)	AT(CFA)	AT(T)	AT(T+EMA)	AT(T+CFA)	AT(F)	AT(F+EMA)	AT(F+CFA)
gaussian_noise	79.0	80.0	79.8	77.6	78.2	76.3	80.2	80.8	78.9
shot_noise	80.0	80.9	80.6	78.5	79.4	77.5	81.3	81.9	80.1
speckle_noise	79.9	80.7	80.5	78.4	79.2	77.3	81.4	81.9	80.0
impulse_noise	74.5	75.1	75.0	73.7	74.5	72.2	75.0	75.2	74.0
defocus_blur	78.6	79.0	78.9	77.9	78.2	76.6	79.9	80.1	77.6
gaussian_blur	76.6	76.8	76.6	76.1	76.3	74.5	77.7	77.6	75.0
motion_blur	74.7	74.9	75.0	74.7	75.0	73.1	75.9	76.0	72.8
zoom_blur	77.6	77.9	77.8	77.2	77.5	75.7	79.4	79.3	76.5
snow	75.6	77.3	77.8	76.3	76.2	77.5	76.7	78.3	77.5
fog	58.1	56.3	55.8	57.8	58.0	53.0	59.1	57.5	52.8
brightness	78.4	79.7	80.1	78.3	78.1	78.2	79.7	80.9	79.8
contrast	42.7	40.5	40.2	41.4	42.0	37.3	43.8	41.7	37.3
elastic_transform	77.4	78.1	78.0	76.9	77.2	75.6	79.1	79.4	77.0
pixelate	81.2	82.0	82.0	80.3	80.6	79.6	82.7	83.1	81.3
jpeg_compression	81.1	81.8	81.8	80.1	80.3	79.6	82.6	83.0	81.4
spatter	77.9	79.1	79.1	78.9	78.3	76.8	79.2	80.0	79.2
saturate	78.7	79.9	79.9	78.1	78.4	77.6	80.3	81.0	80.4
frost	69.3	71.6	72.5	70.7	70.2	70.8	70.9	73.3	70.7

B.3. Overall accuracy results on common corruptions.

We evaluate the robustness and fairness of the adversarial model trained using our approach on common corruptions, and compare it to baseline methods. For this evaluation, we utilize the CIFAR-10C dataset, which contains 18 types of corruptions applied to CIFAR-10 images, using the PreActResNet-18 (PRN-18) architecture.

Table 6 and Table 7 present the overall accuracies for our method and the baselines, respectively. Table 6 shows the overall accuracies of the model using FAIR-TAT, adapted to various approaches on CIFAR-10C, while Table 7 lists the overall accuracies of the baseline methods. The best results on common corruptions for both our method and the baselines are underlined in the tables.

Our approach, FAIR-TAT, particularly when adapted to the baselines, notably FAT, consistently outperforms the baselines in terms of overall model performance on common corruptions. In summary, our findings indicate that FAIR-TAT enhances both robustness and fairness when compared to existing methods on common corruptions. We conclude that FAIR-TAT achieves a better-balanced trade-off between robustness and fairness, as demonstrated by our evaluations on both adversaries and common corruptions.

References

[1] Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. RobustBench: a standardized adversarial robustness benchmark. In Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2021. 1

[2] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In 2023 IEEE Conference on Secure and Trustworthy Machine Learning (SaTML), pages 225–253. IEEE, 2023. 3

[3] 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. 1, 2, 3