# Class-wise Training Dynamics and Fairness in Adversarially Robust Models

**Hyuk Ahn**

## Abstract

Despite the remarkable progress of deep neural networks, they remain vulnerable to adversarial examples, which can drastically degrade performance with imperceptible perturbations. Adversarial training has emerged as a leading defense, yet existing methods often exacerbate class-wise fairness issues, unevenly benefiting different classes. In this work, we systematically investigate the class-wise effects of adversarial training across varying perturbation strengths and defense strategies on CIFAR-100. Our analysis reveals that fairness disparities stem primarily from intrinsic differences in class-specific clean learning difficulty, further magnified by adversarial training. We show that stronger perturbations offer limited robustness gains while increasingly harming clean accuracy and fairness. We also evaluate fair adversarial training methods, such as Fair Robust Learning (FRL) and Class-wise Calibrated Fair Adversarial Training (CFA), and highlight their strengths and limitations. Our findings demonstrate that clean training convergence is a strong predictor of robust performance, and that addressing clean training convergence and robust overfitting is critical for achieving fair adversarial robustness.

## 1. Introduction

Despite the remarkable successes of deep neural networks (DNNs) across a variety of domains, they have been shown to be vulnerable to adversarial examples that are maliciously perturbed to alter the model's predictions while preserving their semantic meaning. To mitigate these vulnerabilities, adversarial training (AT) has emerged as a leading defense mechanism, wherein adversarial examples are incorporated into the training process to enhance model robustness. Among AT methods, Projected Gradient Descent (PGD) (Madry et al., 2018) and TRADES (Zhang et al., 2019) have demonstrated strong effectiveness in improving adversarial robustness.

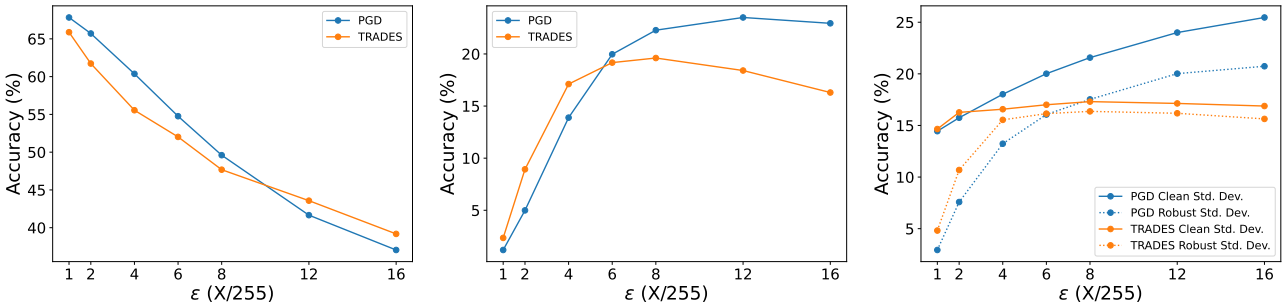
Beyond these foundational methods, a variety of approaches have been proposed to further enhance adversarial robustness. One notable direction is data augmentation. Improved Diversity and Balanced Hardness (IDBH) (Li & Spratling, 2023) introduces a new augmentation technique, Cropshift, and combines it with existing augmentations to strengthen adversarial training. Similarly, Adversarial Vertex Mixup (AVmixup) (Lee et al., 2020) improves robust generalization by incorporating soft-labeled data augmentations through a mixup strategy.

Despite these advances, adversarial training is often accompanied by significant trade-offs, particularly in terms of class-wise performance disparities. While AT improves overall model robustness, it tends to unevenly benefit different classes, leading to robustness fairness issues. Some classes achieve substantial gains in robustness, whereas others suffer notable performance degradation. Recent works have attempted to address these challenges. Fair Robust Learning (FRL) (Xu et al., 2021) proposes dynamically adjusting class margins and sample weights when class-wise performance drops below a threshold. Similarly, Class-wise Calibrated Fair Adversarial Training (CFA) (Wei et al., 2023) introduces class-specific adjustments to attack perturbation margins and regularization strengths, along with a modified weight averaging scheme to stabilize class-wise robustness.

While prior studies have primarily focused on fairness within the context of adversarial robustness, a comprehensive analysis that jointly considers class-wise clean and robust fairness can offer deeper insights into the underlying causes of the fairness issues. We argue that understanding the class-wise impacts of AT methods is critical for developing truly fair and robust models. Therefore, in this work, we empirically investigate classification fairness by analyzing the class-wise effects of different AT strategies from multiple perspectives. Our analysis reveals that fundamental fairness issues arise from a combination of intrinsic class-specific clean training difficulties and the exacerbation of robust overfitting under adversarial training.

## 2. Experimental Setup

In this work, we evaluate the performance of various adversarial training (AT) methods on the CIFAR-100 dataset (Krizhevsky, 2009) using ResNet-9 (He et al., 2016). For all experiments, models are trained using the SGD optimizer




(a) Clean Accuracy (b) Robust Accuracy (c) Class Accuracy Standard Deviation

*Figure 1.* Comparison of overall CIFAR-100 performance of ResNet9 model trained on PGD and TRADES with different perturbation strength

with an initial learning rate of 0.1, which is decayed by a factor of 0.1 at the 100th and 150th epochs. We employ a weight decay of $5 \times 10^{-4}$ and train each model for a total of 200 epochs with a batch size of 128.

For AT methods based on Projected Gradient Descent (PGD) (Madry et al., 2018), we adopt a step size of 2, 10 attack iterations, and random initialization. For TRADES (Zhang et al., 2019), we used the same step size and number of attack iterations, with the trade-off parameter $\beta$ set to 6. Unless otherwise specified, we used perturbation strength of $\epsilon = 8/255$ for all AT methods.

Model robustness is evaluated using 20-step PGD attack (PGD-20) with a perturbation strength of $\epsilon = 8/255$, and random initialization.

## 3. Fairness vs. Perturbation Strength

In this section, we investigate how varying perturbation strengths influence the performance of adversarial training (AT). We provide a comprehensive analysis by first examining the impact of different perturbation strengths on the overall model performance, and then exploring the class-wise effects on both clean and robust accuracies.

### 3.1. Overall Performance

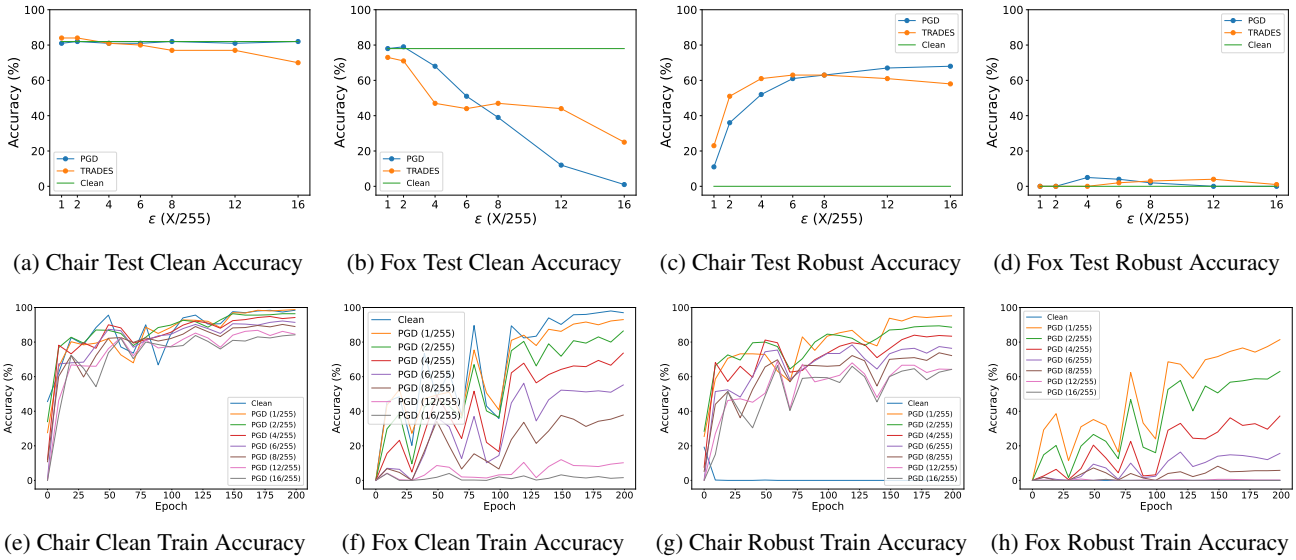
To study the effect of perturbation strength on adversarial training, we train multiple models using two representative AT methods: Projected Gradient Descent (PGD) (Madry et al., 2018) and TRADES (Zhang et al., 2019). For each method, we vary the $l_\infty$-norm perturbation strength from 1/255 to 16/255. Figure 1 summarizes the results, comparing the clean and robust accuracies across perturbation strengths for both training methods.

Figure 1a demonstrates that the overall clean accuracy of the models consistently decreases as the perturbation strength increases. In contrast, as shown in Figure 1b, the robust accuracy for both adversarial training (AT) methods improves with increasing perturbation strength from 1/255 up to 8/255. However, further increases beyond 8/255 do not yield significant additional robustness; notably, in the case of TRADES, robust accuracy declines when perturbation strength exceeds this threshold. Figure 1c reports the standard deviation of per-class accuracies on CIFAR-100, where a clear trend of increasing variability is observed with stronger perturbations for both PGD and TRADES. These results suggest that while larger perturbation strengths substantially impair clean accuracy, they provide only marginal robustness gains beyond a certain point. Furthermore, the rising standard deviation implies heterogeneous effects of adversarial training across classes: some classes exhibit greater degradation in clean accuracy or greater improvement in robustness compared to others, highlighting a non-uniform impact of increased perturbation strength.

### 3.2. Class-wise Performance

We now examine how varying perturbation strengths influence class-wise performance. Figure 2a and Figure 2b depict the clean accuracies for the classes *chair* and *fox* across models trained with clean data, PGD, and TRADES at different perturbation strengths. For the *chair* class, models trained with both adversarial training (AT) methods maintain clean accuracies comparable to the clean model, even at the highest perturbation strength of $\epsilon$=16/255. In contrast, the *fox* class exhibits a substantial decline in clean accuracy under adversarial training, with the PGD-trained model's accuracy falling below 20% as perturbation strength increases. This pronounced disparity between classes highlights the necessity of examining class-wise performance, rather than relying solely on overall metrics, to better understand and address failure modes.

Turning to robust accuracies in Figure 2c and Figure 2d, we observe that for the *chair* class, both PGD and TRADES models experience improved robustness with increasing per-




(a) Chair Test Clean Accuracy (b) Fox Test Clean Accuracy (c) Chair Test Robust Accuracy (d) Fox Test Robust Accuracy

(e) Chair Clean Train Accuracy (f) Fox Clean Train Accuracy (g) Chair Robust Train Accuracy (h) Fox Robust Train Accuracy

*Figure 2.* Clean and adversarial accuracies on CIFAR-100 classes (chair and fox) with clean and adversarial training methods on ResNet9 and different perturbation strengths.

turbation strength up to $\epsilon = 6/255$. Compared to its clean accuracy in Figure 2a, the robust accuracy for *chair* sees less than 15% difference even at the largest perturbation strength. In contrast, robust accuracies for the *fox* class remain below 5% across all perturbation strengths, as shown in Figure 2d, and stronger perturbations offer no meaningful improvement. These findings suggest that certain classes, such as *fox*, may be intrinsically hard to learn under adversarial settings, whereas others, like *chair*, are relatively easy to learn. We further explore the characteristics that differentiate easy and hard classes in the subsequent sections.

### 3.3. Class-wise Training Dynamics

As observed in Section 3.2, the easy class *chair* displayed significantly different performance compared to the hard class *fox* under varying perturbation strengths. Next, we extend the analysis to the training performance. We compare the training accuracies of the models trained on different perturbation strengths at various training checkpoints in the bottom row of Figure 2.

We observe a distinct contrast in training behavior between easy and hard classes under both clean and adversarial training. As shown in Figure 2e and Figure 2g, all models for the *chair* class exhibit stable convergence toward high training accuracies across epochs. In contrast, Figure 2f and Figure 2h show that models for the *fox* class experience significant training instability, characterized by fluctuating training accuracies throughout training. Additionally, we observe pronounced instances of robust overfitting for the *fox* class; notably, the PGD model trained with $\epsilon = 1/255$ achieves above 80% final training robust accuracy but 0% test robust accuracy. Additional results for other classes are provided in the Appendix A.1.

## 4. Fair Adversarial Training

Building on the comparison between *chair* and *fox* in Section 3, our findings are consistent with prior works showing that certain classes are intrinsically harder to classify (e.g., *dog* vs. *car*) (Xu et al., 2021). To address fairness challenges in adversarial training (AT), several methods have been proposed, such as FRL (Fair Robust Learning) which works to balance class performances (Xu et al., 2021) and CFA (Class-wise Calibrated Fair Adversarial Training) that focuses on improving the worst-performing classes (Wei et al., 2023). In this section, we thoroughly evaluate these fair AT methods alongside other AT methods, including IDBH (Li & Spratling, 2023) and AVmixup (Lee et al., 2020), to compare their impact on class-wise performance.

### 4.1. Fairness Comparison

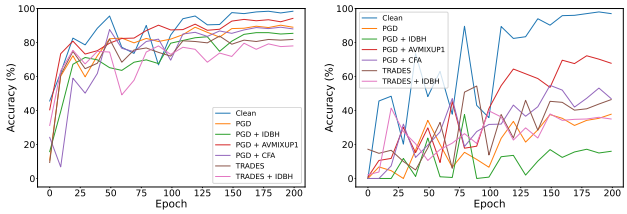
Table 1 presents the clean and robust accuracies of various AT methods, evaluated against PGD-20 adversarial attacks. As expected, the clean model achieves the highest performance on clean examples but performs the worst against adversarial examples. Interestingly, we observe that PGD+AVmixup and methods based on TRADES demonstrate improved fairness, despite some methods not explicitly addressing fairness. We attribute this effect to their training strategies, which incorporate both clean and adversarial examples, thereby encouraging more diverse class representations and potentially mitigating the effect of ro-

*Table 1.* Overall performance comparison on ResNet9 and CIFAR-100 with adversarial training methods. All adversarial models are trained with 10 attack iterations and 8/255 perturbation strength.

| | CLEAN ACCURACY / ROBUST ACCURACY | | | |
|---|---|---|---|---|
| METHOD | AVERAGE | WORST CLASS | BEST CLASS | CLASS STD. DEV. |
| CLEAN | 69.4 / 0.0 | 39.0 / 0.0 | 91.0 / 0.0 | 12.7 / 0.0 |
| PGD | 49.6 / 22.3 | 1.0 / 0.0 | 87.0 / 65.0 | 21.6 / 17.5 |
| PGD + IDBH | 47.8 / 24.3 | 1.0 / 0.0 | 89.0 / 70.0 | 23.2 / 19.1 |
| PGD + CFA | 50.8 / 17.9 | 7.0 / 0.0 | 88.0 / 63.0 | 19.1 / 16.1 |
| PGD + AVMIXUP | 60.4 / 16.3 | 12.0 / 0.0 | 91.0 / 52.0 | 18.4 / 14.0 |
| TRADES | 47.7 / 19.6 | 11.0 / 0.0 | 87.0 / 67.0 | 17.3 / 16.36 |
| TRADES + IDBH | 49.2 / 20.3 | 10.0 / 0.0 | 86.0 / 67.0 | 18.0 / 16.9 |
| TRADES + CFA | 54.1 / 20.8 | 6.0 / 0.0 | 89.0 / 68.0 | 19.3 / 17.6 |
| TRADES + FRL | 43.7 / 15.9 | 18.0 / 1.0 | 81.0 / 59.0 | 13.7 / 13.9 |

bust overfitting.

Among the fairness-oriented AT methods, CFA improves fairness relative to the standard PGD baseline, increasing the worst-class clean accuracy to 7% and reducing the clean and robust class standard deviations by more than 1%. FRL exhibits improvement in the worst-class clean accuracy up by 7% and significant reduction in the clean and robust standard deviation by more than 2%. However, these fairness gains come at many costs. Specifically, FRL suffers a decrease in the average clean and robust accuracies beyond 3%, while CFA exhibits average robust accuracy drop of over 4% but a more than 1% increase in the average clean accuracy.




(a) Chair Clean Accuracy (b) Fox Clean Accuracy

*Figure 3.* Clean training accuracies on CIFAR-100 classes (*chair* and *fox*) with ResNet9 and different adversarial training methods.

### 4.2. Class-wise Training Dynamics

Here, we present the class-wise performance results on the CIFAR-100 classes *chair* and *fox*, following the methodology described in Section 3.3. As shown in Figure 3a, all adversarial training (AT) models exhibit stable convergence when training on the *chair* class, consistent with the trends previously observed in Figure 2e. Similarly, in Figure 3b, all AT models display unstable training dynamics with fluctuating accuracies on the *fox* class, resembling the behavior observed in Figure 2f. These results demonstrate that, despite

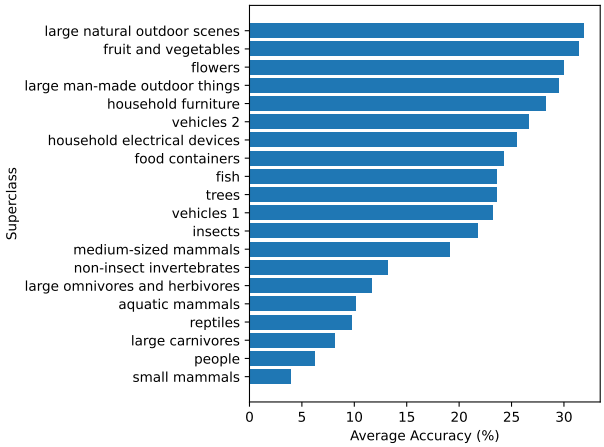



*Figure 4.* Superclasses ordered by their average robust accuracies from different adversarial training methods.

employing strong data augmentation techniques and fair adversarial training methods, the models were unable to overcome the inherent class-specific difficulties. This supports the notion that easy and hard classes persist, and that such class-level challenges remain difficult to mitigate through the approaches investigated. Additional experiments on other CIFAR-100 classes are provided in Appendix A.2.

## 5. Hard Classes

Thus far, our analyses show that all AT methods exhibit similar behaviors with respect to easy and hard classes. CIFAR-100 consists of 100 classes, grouped into 20 superclasses, with each superclass containing five classes. In this section, we examine the superclass-level performance trends. Figure 4 ranks superclasses by their average robust accuracies across all AT methods from Table 1. A clear performance gap emerges: superclasses above and including

*medium-sized mammals* achieve an average robust accuracy beyond 19%, whereas those below and including *non-insect invertebrates* achieve below 13%. A similar trend holds when examining superclass robust accuracies on all the individual AT methods, as detailed in the Appendix A.3. These findings empirically confirm the existence of easy and hard classes, with a similar group of superclasses that consistently perform the worst across various AT methods. Moreover, they highlight the limitations of current data augmentation and fair AT strategies in substantially improving the performance of the worst-performing classes.

## 6. Conclusion

In this work, we investigated the effects of perturbation strengths and various adversarial training methods on the fairness in CIFAR-100, with a focus on class-wise clean and robust performance. Our empirical results suggest that certain classes inherently achieve better clean training convergence, which directly correlates with easier adversarial training convergence and improved robustness. Conversely, classes that struggled with clean training convergence exhibited both robust training instability and robust overfitting, resulting in minimal robustness gains. These findings indicate that, instead of adversarial robustness, class-wise clean training stability can serve as a useful predictor of class-wise robust performance. It also indicates that class-wise robust overfitting is the main obstacle in achieving robust fairness. For future work, we aim to explore whether these fairness patterns persist across different model architectures and investigate whether resolving clean training stability and robust overfitting can lead to improvements in robust fairness.

## References

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, Las Vegas, NV, 2016. IEEE. doi: 10.1109/CVPR.2016.90.

Krizhevsky, A. Learning multiple layers of features from tiny images. Technical Report UTML TR 2009, Department of Computer Science, University of Toronto, Toronto, Ontario, 2009.

Lee, S., Lee, H., and Yoon, S. Adversarial vertex mixup: Toward better adversarially robust generalization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 269–278, 2020. doi: 10.1109/CVPR42600.2020.00035.

Li, L. and Spratling, M. W. Data augmentation alone can improve adversarial training. In *Proceedings of the Eleventh International Conference on Learning Representations (ICLR 2023)*, 2023.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *Proceedings of the 6th International Conference on Learning Representations (ICLR 2018)*, Vancouver, BC, Canada, 2018.

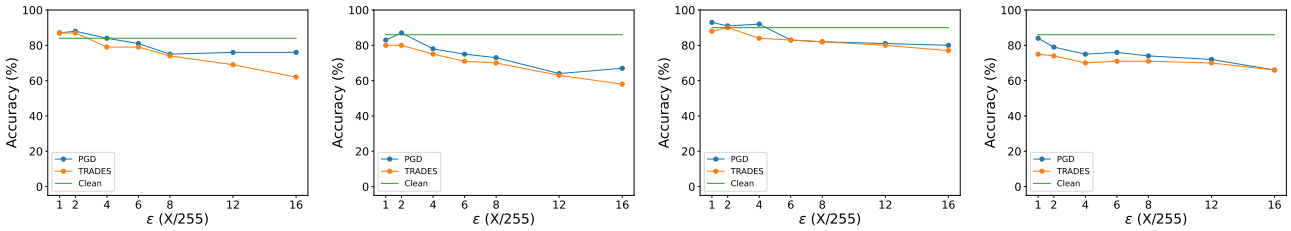
Wei, Z., Wang, Y., Guo, Y., and Wang, Y. Cfa: Class-wise calibrated fair adversarial training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8193–8201. IEEE, 2023. doi: 10.1109/CVPR52729.2023.00792.

Xu, H., Liu, X., Li, Y., Jain, A. K., and Tang, J. To be robust or to be fair: Towards fairness in adversarial training. In *Proceedings of the 38th International Conference on Machine Learning (ICML 2021)*, volume 139, pp. 11492–11501, Virtual Conference, 2021. PMLR.

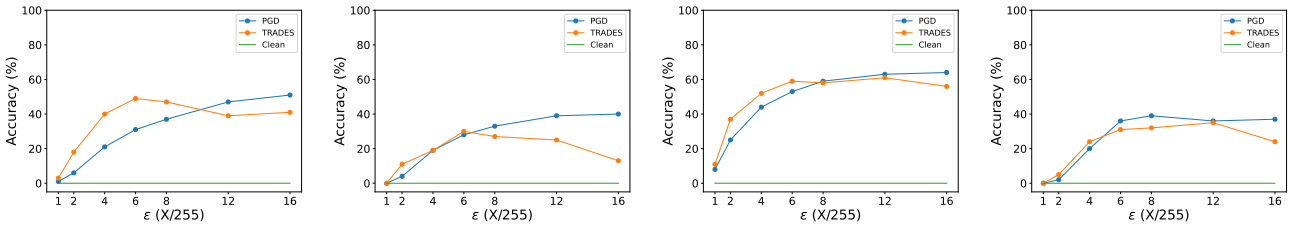
Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *Proceedings of the 36th International Conference on Machine Learning (ICML 2019)*, volume 97, pp. 7472–7482, Long Beach, CA, 2019. PMLR.

# A. Appendix

## A.1. More Results of Class-wise Perturbation Strength Impact

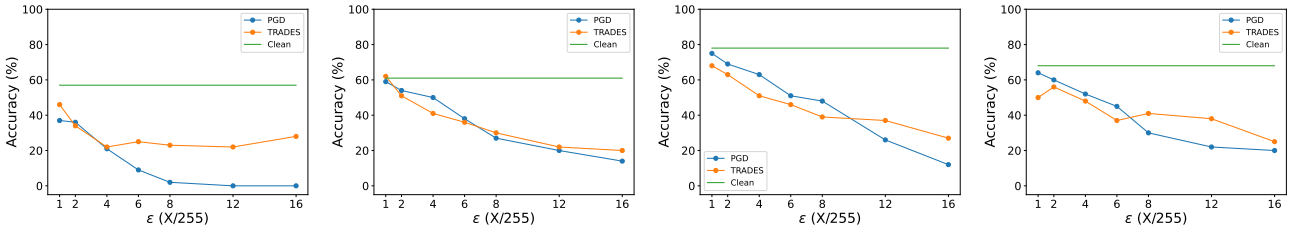



(a) Apple Clean Accuracy (b) Chimpanzee Clean Accuracy (c) Orange Clean Accuracy (d) Sea Clean Accuracy

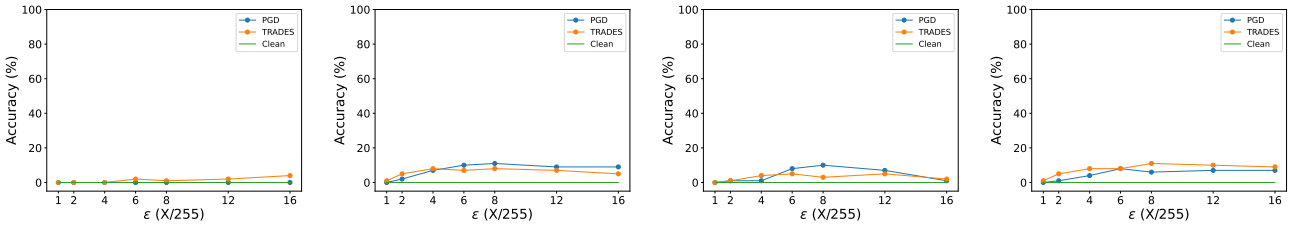



(e) Apple Robust Accuracy (f) Chimpanzee Robust Accuracy (g) Orange Robust Accuracy (h) Sea Robust Accuracy

*Figure 5.* Testing clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9 and different perturbation strengths.

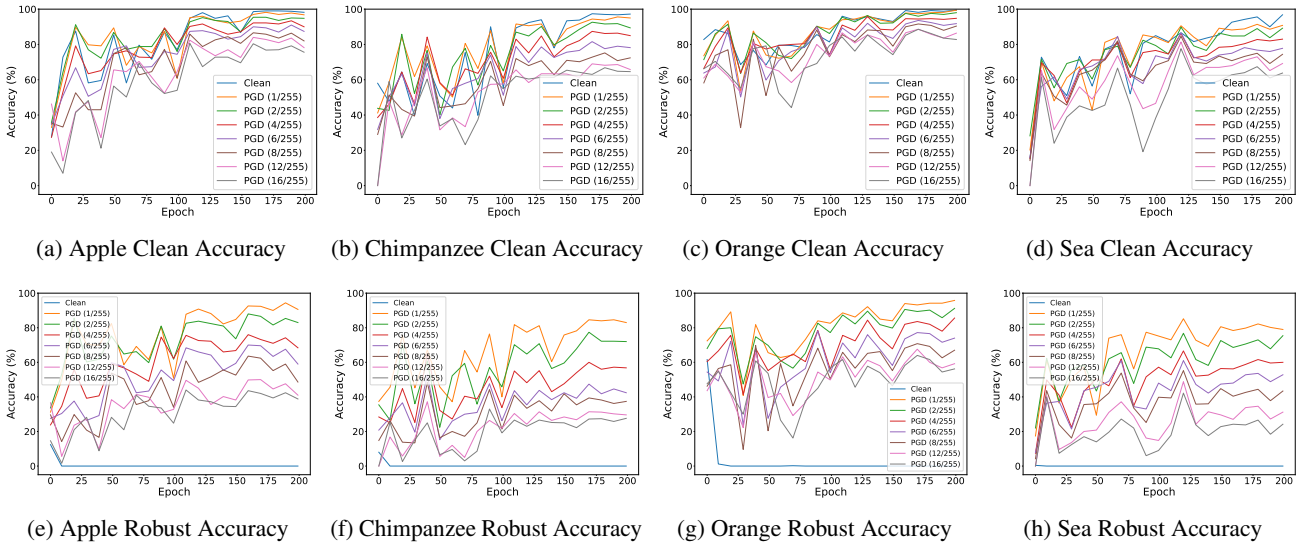



(a) Squirrel Clean Accuracy (b) Table Clean Accuracy (c) Tiger Clean Accuracy (d) Tulip Clean Accuracy




(e) Squirrel Robust Accuracy (f) Table Robust Accuracy (g) Tiger Robust Accuracy (h) Tulip Robust Accuracy

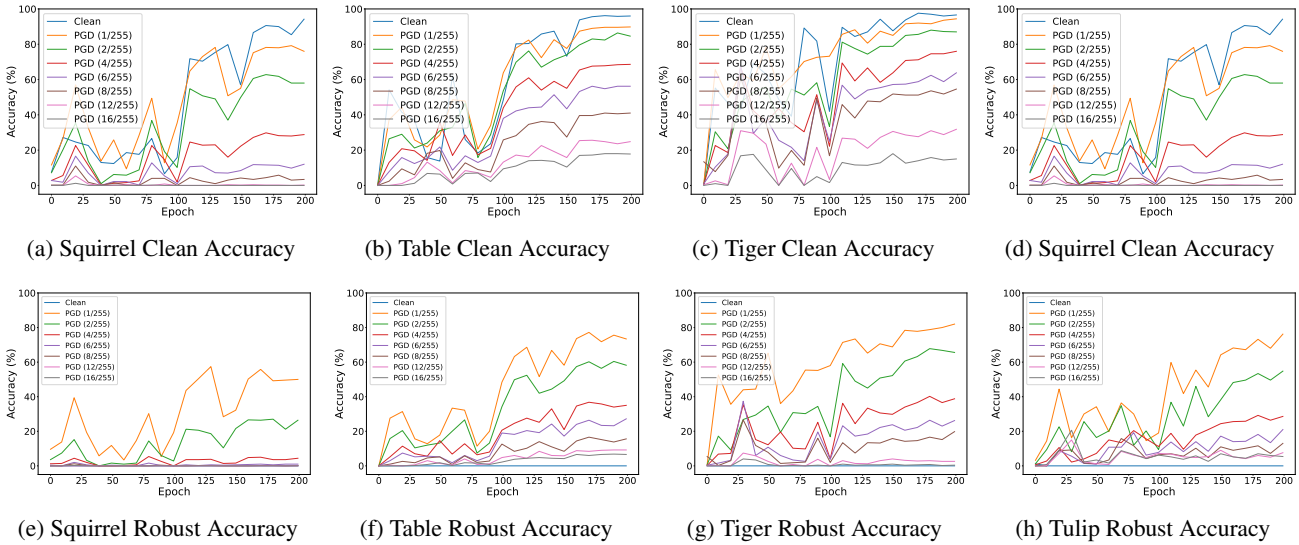
*Figure 6.* Testing clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9 and different perturbation strengths.




(a) Apple Clean Accuracy (b) Chimpanzee Clean Accuracy (c) Orange Clean Accuracy (d) Sea Clean Accuracy

(e) Apple Robust Accuracy (f) Chimpanzee Robust Accuracy (g) Orange Robust Accuracy (h) Sea Robust Accuracy

*Figure 7.* Training clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9 and different perturbation strengths.

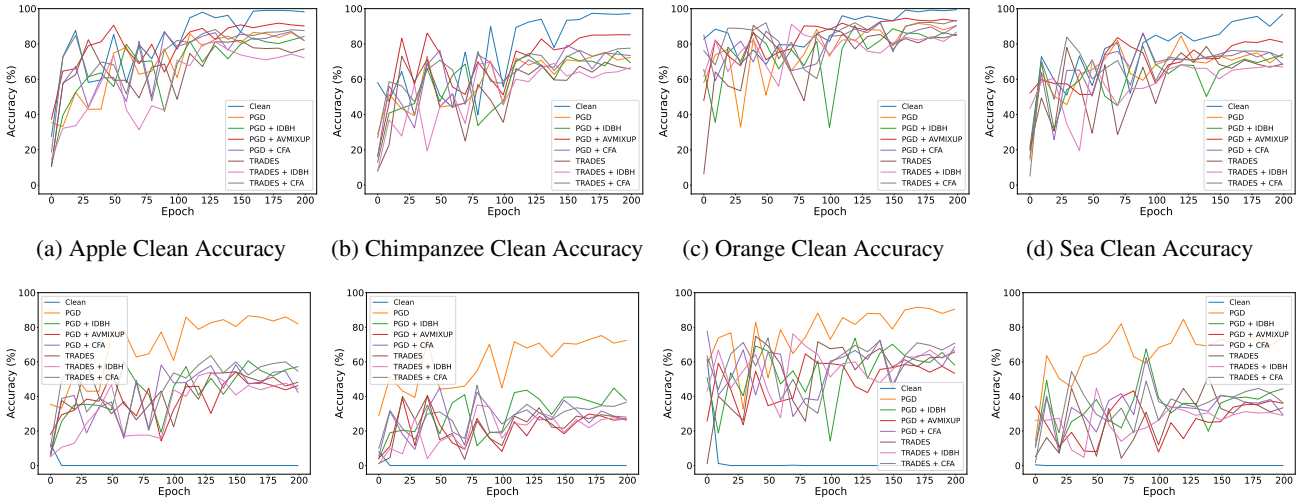



(a) Squirrel Clean Accuracy (b) Table Clean Accuracy (c) Tiger Clean Accuracy (d) Squirrel Clean Accuracy

(e) Squirrel Robust Accuracy (f) Table Robust Accuracy (g) Tiger Robust Accuracy (h) Tulip Robust Accuracy

*Figure 8.* Training clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9 and different perturbation strengths.

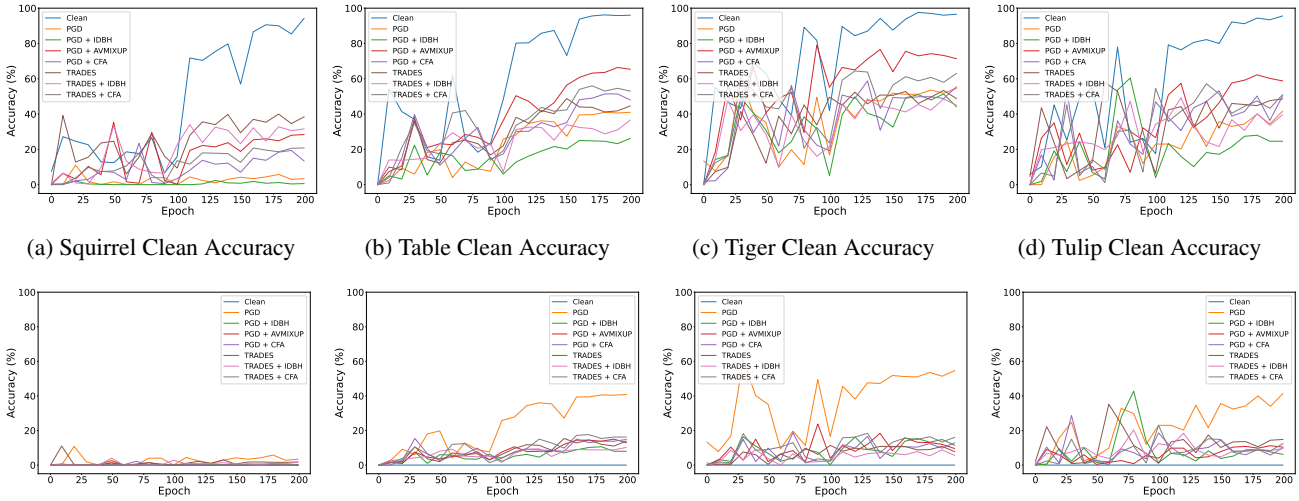
## A.2. More Results of Class-wise Fair Adversarial Training Impact




(a) Apple Clean Accuracy (b) Chimpanzee Clean Accuracy (c) Orange Clean Accuracy (d) Sea Clean Accuracy

(e) Apple Robust Accuracy (f) Chimpanzee Robust Accuracy (g) Orange Robust Accuracy (h) Sea Robust Accuracy

*Figure 9.* Training clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9.

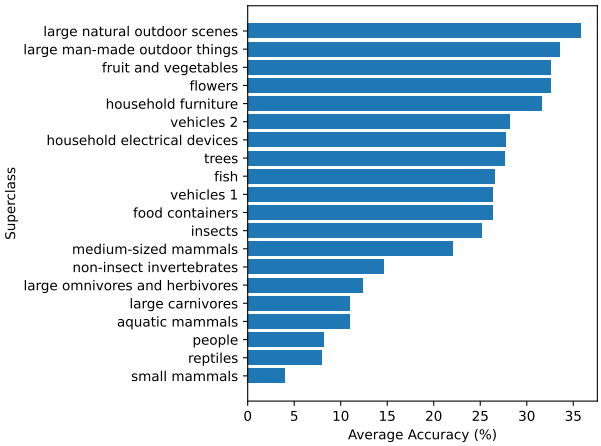



(a) Squirrel Clean Accuracy (b) Table Clean Accuracy (c) Tiger Clean Accuracy (d) Tulip Clean Accuracy

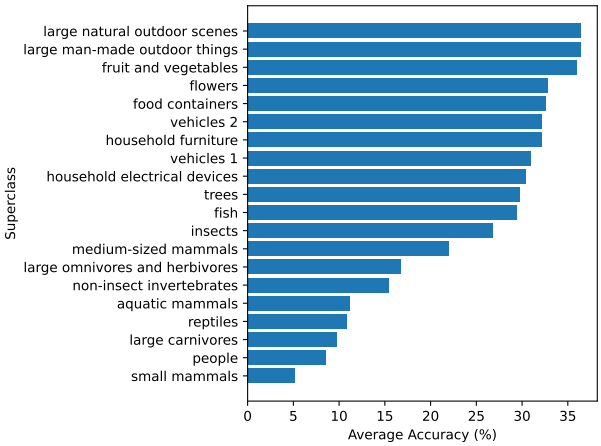
(e) Squirrel Robust Accuracy (f) Table Robust Accuracy (g) Tiger Robust Accuracy (h) Tulip Robust Accuracy

*Figure 10.* Training clean and robust accuracies on CIFAR-100 classes with clean and adversarial training methods on ResNet9.

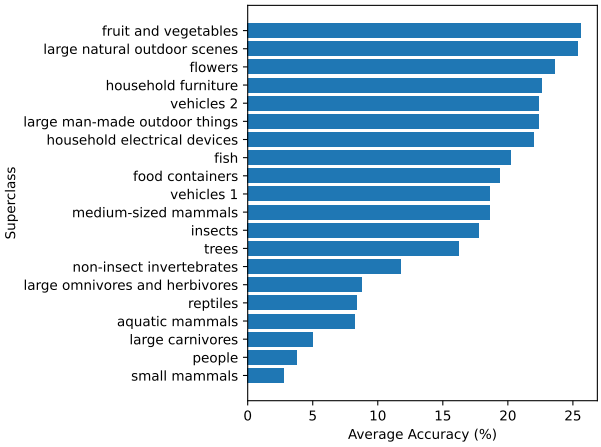
## A.3. More Results of Superclass Analysis



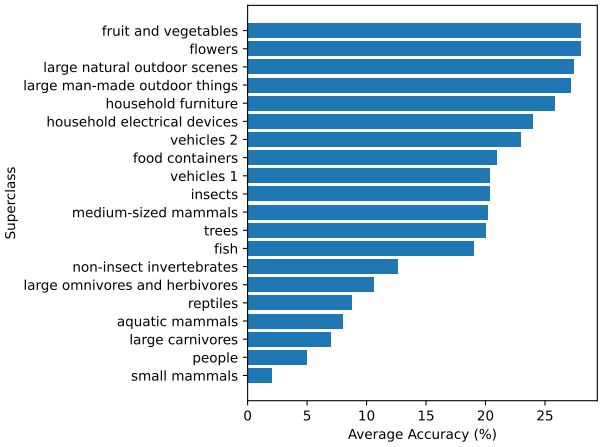

(a) PGD




(b) PGD + IDBH

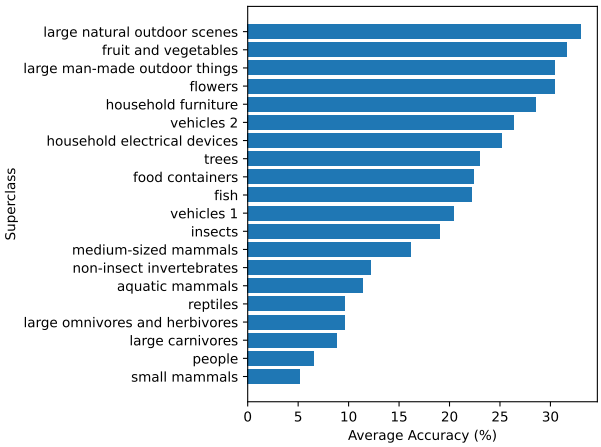



(c) PGD + AVmixup

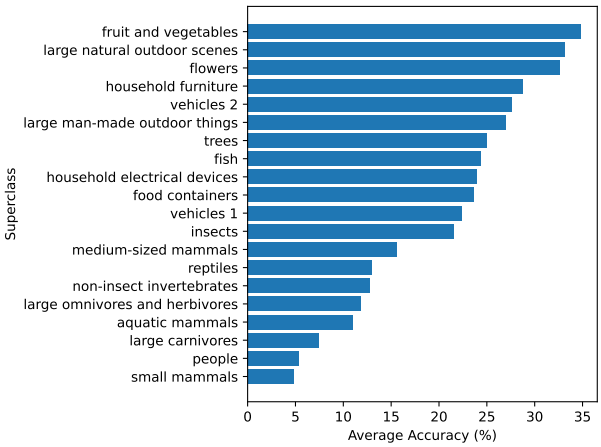



(d) PGD + CFA

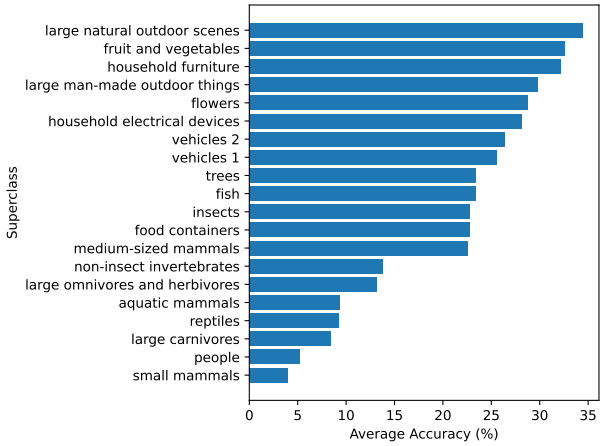
*Figure 11.* Superclasses ordered by their robust accuracies from PGD adversarial training methods.



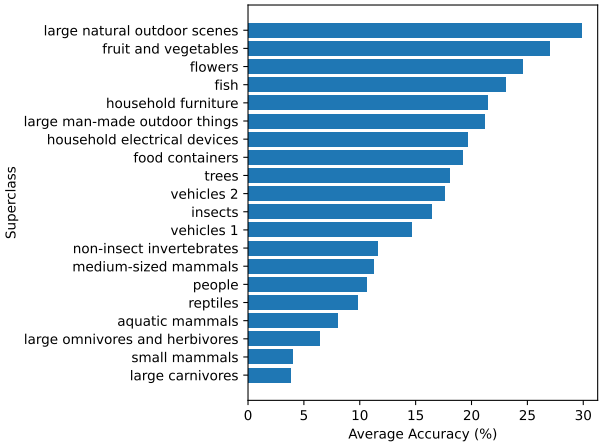

(a) TRADES




(b) TRADES + IDBH




(c) TRADES + CFA




(d) TRADES + FRL

*Figure 12.* Superclasses ordered by their robust accuracies from TRADES adversarial training methods.