# Analysis and Applications of Class-wise Robustness in Adversarial Training

Qi Tian[1], Kun Kuang[1†], Kelu Jiang[1], Fei Wu[1], Yisen Wang[2†]
[1] College of Computer Science and Technology, Zhejiang University
[2] Key Lab of Machine Perception (MoE), School of EECS, Peking University
{tianqics,kunkuang,jiangkelu,wufei}@zju.edu.cn
{yisen.wang}@pku.edu.cn

## ABSTRACT

Adversarial training is one of the most effective approaches to improve model robustness against adversarial examples. However, previous works mainly focus on the overall robustness of the model, and the in-depth analysis on the role of each class involved in adversarial training is still missing. In this paper, we propose to analyze the class-wise robustness in adversarial training. First, we provide a detailed diagnosis of adversarial training on six benchmark datasets, *i.e.*, MNIST, CIFAR-10, CIFAR-100, SVHN, STL-10 and ImageNet. Surprisingly, we find that there are *remarkable robustness discrepancies among classes*, leading to unbalance/unfair class-wise robustness in the robust models. Furthermore, we keep investigating the relations between classes and find that the unbalanced class-wise robustness is pretty consistent among different attack and defense methods. Moreover, we observe that the stronger attack methods in adversarial learning achieve performance improvement mainly from a more successful attack on the vulnerable classes (*i.e.*, classes with less robustness). Inspired by these interesting findings, we design a simple but effective attack method based on the traditional PGD attack, named Temperature-PGD attack, which proposes to enlarge the robustness disparity among classes with a temperature factor on the confidence distribution of each image. Experiments demonstrate our method can achieve a higher attack rate than the PGD attack. Furthermore, from the defense perspective, we also make some modifications in the training and inference phase to improve the robustness of the most vulnerable class, so as to mitigate the large difference in class-wise robustness. We believe our work can contribute to a more comprehensive understanding of adversarial training as well as rethinking the class-wise properties in robust models.

## CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **Security and privacy**;

## KEYWORDS

adversarial training; adversarial robustness; class-wise properties; adversarial examples

**ACM Reference Format:**
Qi Tian[1], Kun Kuang[1†], Kelu Jiang[1], Fei Wu[1], Yisen Wang[2†] . 2021. Analysis and Applications of Class-wise Robustness in Adversarial Training. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '21), August 14–18, 2021, Virtual Event, Singapore.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3447548.3467403

## 1 INTRODUCTION

Deep learning has achieved great success in many applications (such as image classification [11], video processing [35], recommender systems [36]). Unfortunately, the existence of adversarial examples [24] reveals the vulnerability of deep neural networks, which hinders the practical deployment of deep learning models. Adversarial training (training on adversarial examples) [15] has been demonstrated to be one of the most successful defense methods by Athalye et al. [1]. While it can only obtain moderate robustness even for simple image datasets like CIFAR-10, a comprehensive understanding for adversarial training is critical for further robustness improvement.

Previously, some works tried to analyze adversarial training from robust optimization [27], robustness generalization [21, 32], training strategy [4, 15, 19, 29, 34]. However, in these works, they all focus on the averaged robustness over all classes while ignoring the possible difference among different classes. In other fields, there are some works revealing the class-bias learning phenomenon in the standard training (training on natural examples) [25, 28], in which they found that some classes ("easy" classes) are easy to learn and converge faster than other classes ("hard" classes). Inspired by this, a natural question is then raised here:

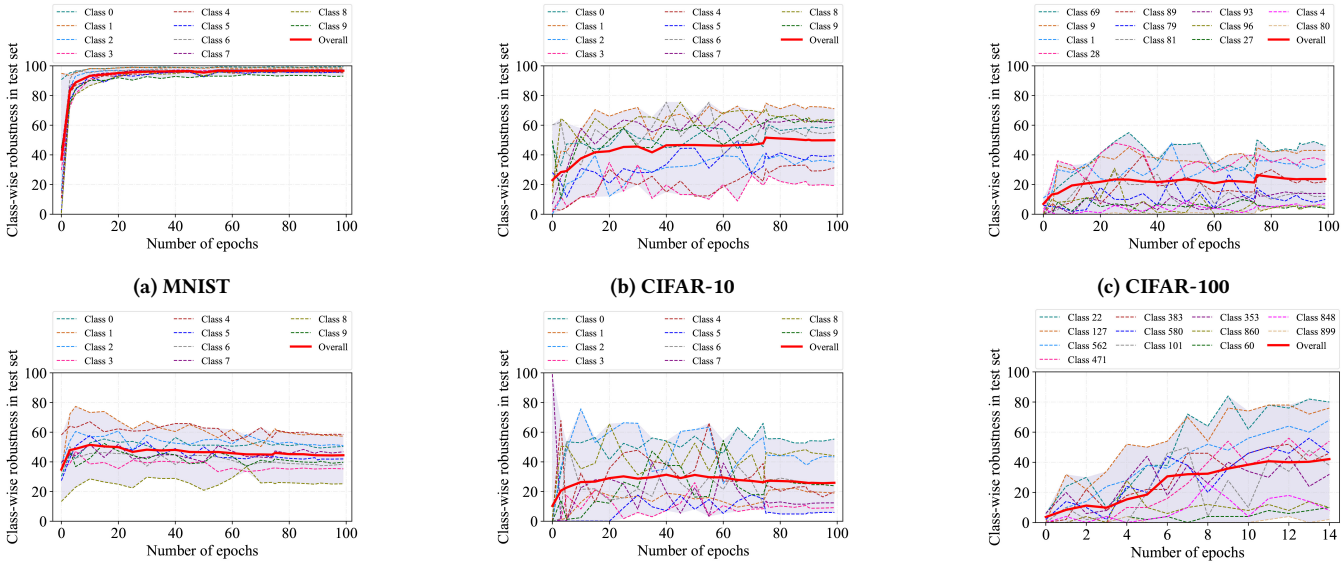
> *Does each class perform similarly in the adversarially trained models? Or is each class equally vulnerable? If not, how would the class-wise robustness affect the performance of classical attack and defense methods in adversarial learning?*

In this paper, we investigate the above questions comprehensively. Specifically, we conduct a series of experiments on several benchmark datasets and find that the class-bias learning phenomenon still exists in adversarial training which is even severe than standard training. For this finding, we have the following questions to explore:


*KDD '21, August 14–18, 2021, Virtual Event, Singapore.*

ACM ISBN 978-1-4503-8332-5/21/08…$15.00
https://doi.org/10.1145/3447548.3467403

[†]Corresponding Authors.




(a) MNIST (b) CIFAR-10 (c) CIFAR-100

(d) SVHN (e) STL-10 (f) ImageNet

**Figure 1: Class-wise robustness at different epochs in the test set**

1) What is the relation between the unbalanced robustness and the properties of the dataset itself?
2) Can we use the class-wise properties to further enlarge the differences among classes?
3) Are there any ways to improve the robustness of vulnerable classes so as to obtain a more balance/fairer robust model?

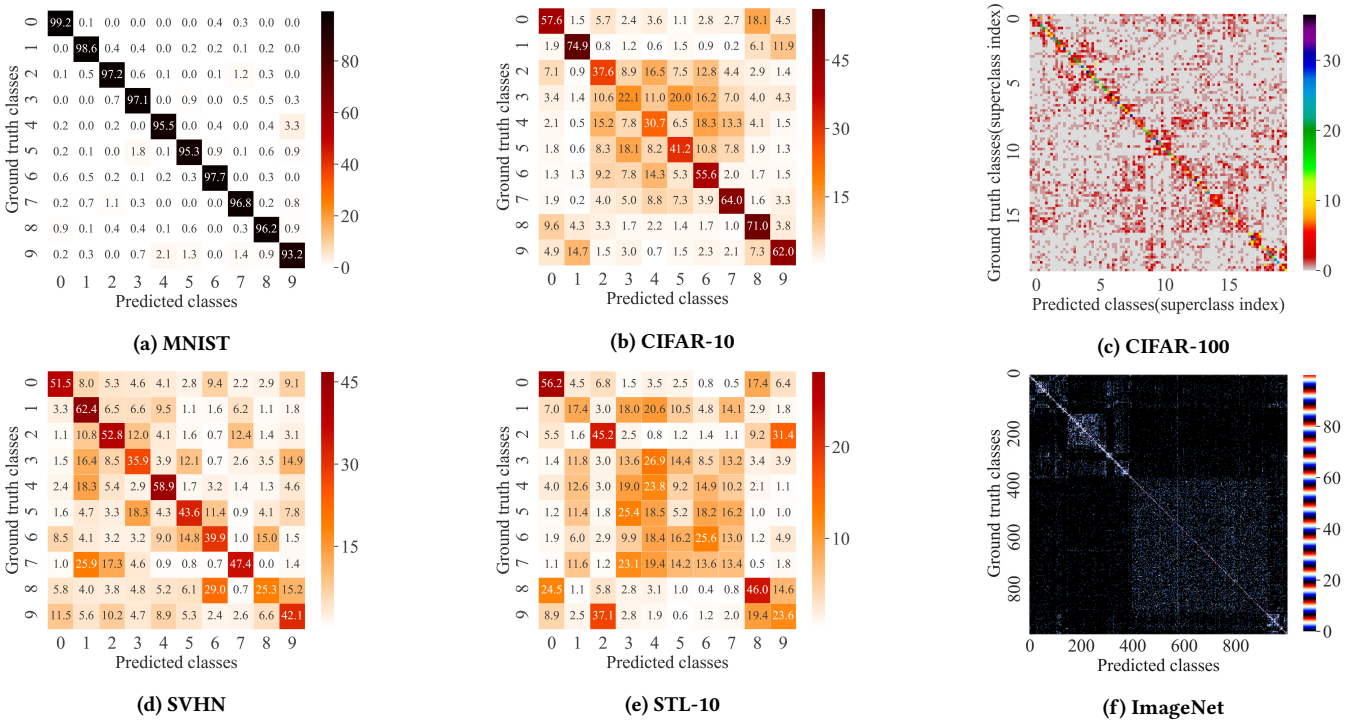
We conduct extensive analysis on the obtained robust models and summarize the following contributions:

- **Analysis on class-wise robustness**
  1) We systematically investigate the relation between different classes and find classes in each dataset can be divided into several groups, and intra-group classes are easily affected by each other.
  2) The relative robustness between each class is pretty consistent among different attack or defense methods, which indicates that the dataset itself plays an important role in the class-wise robustness.
- **Applications for stronger attack**
  1) We make full use of the properties of the vulnerable classes to propose an attack that can effectively reduce the robustness of these classes, thereby increasing the disparity among classes.
- **Applications for stronger defenses**
  1) Training phase: Since the above group-based relation is commonly observed in the dataset, we propose a method that can effectively use this relation to adjust the robustness of the most vulnerable class.
  2) Inference phase: We find that the background of the images may be a potential factor for different classes to be easily flipped by each other. Our experiments show that the robustness of the most vulnerable class can be improved by simply changing the background.

## 2 RELATED WORK

**Class-wise analysis.** Class-wise properties are widely studied in the deep learning community, such as long-tailed data [25] and noisy label [28]. The datasets for these specific tasks are significantly different in each class. *i.e.*, in long-tailed data task, the tail-class (with few training data) usually achieves lower accuracy since it cannot be sufficiently trained. In the asymmetric noisy label task, classes with more label noise usually have lower accuracy. However, in the adversarial community, few people pay attention to class-wise properties because all benchmark datasets seem to be class-balanced. Recently, we notice two parallel and independent works [5, 16] also point out the performance disparity in robust models, but none of them explore the relation between class-wise robustness and the properties of the dataset itself, and our work takes the first step to investigate this problem.

**Attack.** Adversarial attacks are used to craft adversarial examples by adding small and human imperceptible adversarial perturbations to natural examples, which mainly include white-box attacks and black-box attacks. In white-box settings, the attackers know the parameters of the defender model and generate adversarial noise by maximizing the loss function (*e.g.*, Fast Gradient Sign Method (FGSM) [10] and Projected Gradient Descent (PGD) [15] attack maximize cross-entropy loss, while Carlini-Wagner (C&W) [6] attack maximize hinge loss). In black-box settings, there are transfer-based and query-based attacks. The former attacks a substitute model and the generated noise can transfer to the target model [26, 31]. The latter crafts adversarial examples by querying the output of the target model [3, 14]. In this paper, we analyze the class-wise robustness performance of different attacks and propose an attack to illustrate that the unbalanced robustness can be enlarged by carefully using the information of vulnerable classes.




(a) MNIST (b) CIFAR-10 (c) CIFAR-100

(d) SVHN (e) STL-10 (f) ImageNet

**Figure 2: Confusion matrix of robustness in the test set**

**Defense.** Adversarial training [15] is known as the most effective and standard way to against adversarial examples. A range of methods have been proposed to improve adversarial training, including modifying regularization term [19, 29, 34], adding unlabeled data [7] and data augmentation [23]. Since adversarial training is more time-consuming than standard training, Wong et al. [30] propose some solutions to accelerate model training. On the other hand, some researchers [2, 18, 20] try to improve model robustness by preprocessing the image in the inference phase, and these methods are usually complementary to adversarial training. However, none of these methods consider the difference in class-wise robustness, and we have proposed some methods that can improve the robustness of the most vulnerable class so as to obtain a fairer output.

## 3 PRELIMINARY

In this section, we first introduce the formula and notations in adversarial training, then give several definitions about robust/non-robust example and robust/vulnerable/confound class used through this paper.

**Vanilla adversarial training.** Madry et al. [15] formalize the adversarial training as a min-max optimization problem. Given a DNN classifier $h_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, a correctly classified natural example $\boldsymbol{x}$ with class label $y$, cross-entropy loss $\ell(\cdot)$ and an adversarial example $\boldsymbol{x}'$ can be generated by perturbing $\boldsymbol{x}$, then the objective of adversarial training is:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \max_{\left\|\boldsymbol{x}_i'-\boldsymbol{x}_i\right\|_p \leq \epsilon} \ell\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i'\right), y_i\right), \tag{1}$$

where the inner maximization applies the Projected Gradient Descent (PGD) attack to craft adversarial examples, and the outer minimization uses these examples as augmented data to train the model. Since the adversarial perturbation should not be observed by humans, these noises are bounded by $L_p$-norm $\left\|\boldsymbol{x}_i' - \boldsymbol{x}_i\right\|_p \leq \epsilon$.

**TRADRS.** Another popular adversarial training method (TRADRS [34]) is to add a regularization term to the cross-entropy loss:

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right), y_i\right) + \beta \max_{\left\|\boldsymbol{x}_i'-\boldsymbol{x}_i\right\|_p \leq \epsilon} \mathcal{K}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i\right), h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_i'\right)\right), \tag{2}$$

where $\mathcal{K}(\cdot)$ represents Kullback-Leibler divergence and $\beta$ can adjust the relative performance between natural and robust accuracy.

In addition, we define some concepts for the convenience of the following expressions.

*Definition 3.1.* (**Robust Example**) Given a natural example $\boldsymbol{x}$ with ground truth class $y$ and a DNN classifier $h_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, if this example does not exist adversarial counterpart in bounded $\epsilon$-ball $\|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon$: $h_{\boldsymbol{\theta}}(\boldsymbol{x}') \equiv y$, the example $\boldsymbol{x}$ is defined as a robust example.

*Definition 3.2.* (**Non-Robust Example and Confound Class**) Given a natural example $\boldsymbol{x}$ with ground truth class $y$ and a DNN classifier $h_{\boldsymbol{\theta}}$ with parameters $\boldsymbol{\theta}$, if the prediction of the model is $y'$ after adding a bounded $\epsilon$-ball $\|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon : h_{\boldsymbol{\theta}}(\boldsymbol{x}') = y' \neq y$, the example $\boldsymbol{x}$ is defined as a non-robust example and $y'$ is defined as confound class of example $\boldsymbol{x}$.

**Table 1: Adversarial robustness (%) (under popular attacks) on CIFAR-10.**

| Defenses(Attacks) | Tot. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | CV | MCD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Madry(FGSM) | 65.5 | 73.7 | 81.2[1] | 51.9 | **41.5**[2] | 54.2 | 49.4 | 73.9 | 72.5 | 78.5 | 78.5 | 191.9 | 39.7 |
| TRADES(FGSM) | 66.9 | 77.5 | 85.9 | 49.7 | **41.9** | 55.8 | 52.8 | 73.0 | 76.8 | 80.7 | 75.3 | 211.2 | 44.0 |
| MART(FGSM) | 67.4 | 73.7 | 84.9 | 54.5 | **45.7** | 50.1 | 51.6 | 76.9 | 75.2 | 83.9 | 77.7 | 206.1 | 39.2 |
| HE(FGSM) | 68.4 | 71.9 | 84.6 | 52.0 | **42.2** | 57.0 | 57.9 | 76.5 | 77.8 | 83.4 | 80.9 | 200.5 | 42.3 |
| Madry($CW_\infty$) | 57.1 | 67.5 | 79.5 | 43.0 | **37.7** | 41.5 | 41.0 | 57.5 | 60.0 | 71.5 | 72.0 | 212.5 | 41.8 |
| TRADES($CW_\infty$) | 59.4 | 69.5 | 85.5 | 39.0 | **38.5** | 43.0 | 46.5 | 57.0 | 67.0 | 77.5 | 70.5 | 258.5 | 47.0 |
| MART($CW_\infty$) | 58.8 | 65.5 | 80.5 | 43.0 | **39.5** | 41.0 | 41.0 | 63.0 | 67.5 | 76.5 | 71.0 | 232.7 | 41.0 |
| HE($CW_\infty$) | 63.6 | 71.2 | 87.5 | 47.1 | **44.4** | 49.8 | 50.1 | 61.4 | 71.2 | 81.5 | 72.7 | 210.8 | 43.1 |
| Madry(PGD) | 52.1 | 63.8 | 71.6 | 39.1 | **25.3** | 36.7 | 38.6 | 57.4 | 59.5 | 63.1 | 66.8 | 224.3 | 46.3 |
| TRADES(PGD) | 56.3 | 67.8 | 80.6 | 37.8 | **29.4** | 40.6 | 43.9 | 59.3 | 66.9 | 71.8 | 65.6 | 263.6 | 51.1 |
| MART(PGD) | 58.2 | 64.5 | 78.0 | 45.1 | **35.4** | 37.7 | 43.5 | 65.3 | 67.5 | 76.3 | 69.5 | 235.1 | 42.6 |
| HE(PGD) | 60.7 | 64.9 | 79.3 | 41.0 | **34.5** | 47.9 | 51.5 | 67.6 | 70.5 | 76.9 | 73.2 | 224.8 | 44.8 |
| Madry(Transfer-based attack) | 80.2 | 84.5 | 87.7 | 71.0 | **68.3** | 78.9 | 69.2 | 86.3 | 82.9 | 87.6 | 86.4 | 56.1 | 19.4 |
| TRADES(Transfer-based attack) | 82.0 | 87.7 | 92.3 | 70.9 | **68.0** | 78.2 | 70.0 | 87.8 | 87.6 | 90.8 | 86.8 | 78.1 | 24.2 |
| MART(Transfer-based attack) | 82.9 | 87.4 | 94.7 | 74.0 | **66.7** | 76.0 | 68.8 | 89.9 | 88.0 | 93.7 | 90.0 | 99.2 | 28.0 |
| HE(Transfer-based attack) | 84.5 | 90.1 | 95.9 | 75.6 | **60.8** | 77.4 | 76.7 | 91.1 | 92.1 | 93.4 | 92.2 | 115.3 | 35.1 |
| Madry($\mathcal{N}$ atacck) | 56.1 | 67.5 | 77.7 | 43.7 | **31.4** | 42.7 | 49.0 | 53.7 | 60.1 | 64.4 | 71.1 | 190.5 | 46.3 |
| TRADES($\mathcal{N}$ atacck) | 64.4 | 73.1 | 87.4 | 46.4 | **44.4** | 49.1 | 61.7 | 56.9 | 71.6 | 79.5 | 74.1 | 200.0 | 43.0 |
| MART($\mathcal{N}$ atacck) | 67.5 | 72.3 | 83.4 | 55.3 | **49.0** | 54.1 | 61.2 | 67.1 | 72.9 | 82.3 | 77.6 | 133.6 | 34.4 |
| HE($\mathcal{N}$ atacck) | 69.7 | 75.9 | 88.3 | 52.7 | **44.7** | 65.4 | 62.6 | 70.1 | 76.0 | 84.5 | 77.5 | 168.6 | 43.5 |

[1] The underscore indicates the most robust class.
[2] The bold indicates the most vulnerable class.

**Table 2: Superclasses in CIFAR-10 and STL-10.**

| Dataset | Transportation | | | | | |
|---|---|---|---|---|---|---|
| CIFAR-10 | Airplane(0) | Automobile(1) | Ship(8) | Truck(9) | | |
| STL-10 | Airplane(0) | Car(2) | Ship(8) | Truck(9) | | |
| Dataset | Animals | | | | | |
| CIFAR-10 | Bird(2) | Cat(3) | Deer(4) | Dog(5) | Frog(6) | Horse(7) |
| STL-10 | Bird(1) | Cat(3) | Deer(4) | Dog(5) | Horse(6) | Monkey(7) |

The number in brackets represents the numeric label of the class in the dataset.

*Definition 3.3.* (**Robust Class and Vulnerable Class**) A class whose robustness is higher than the overall robustness is called a robust class. In contrast, a class whose robustness is lower than the overall robustness is called a vulnerable class.

## 4 CLASS-WISE ROBUSTNESS ANALYSIS

In this section, we focus on analyzing the class-wise robustness, including class-biased learning and class-relation exploring on six benchmark datasets. Moreover, we investigate the class-wise robustness with different attack and defense models.

We use six benchmark datasets in adversarial training to obtain the corresponding robust model, *i.e.*, MNIST [13], CIFAR-10 & CIFAR-100 [12], SVHN [17], STL-10 [8] and ImageNet [9]. Table 2 highlights that the classes of CIFAR-10 and STL-10 can be grouped into two superclasses: *Transportation* and *Animals*. Similarly, CIFAR-100 also contains 20 superclasses with each has 5 subclasses. For the ImageNet dataset, the pipeline of adversarial training follows Wong et al. [30], while the training methods of other datasets follow Madry et al. [15]. See Appendix A.1 for detailed experimental settings.

### 4.1 Class-biased Learning

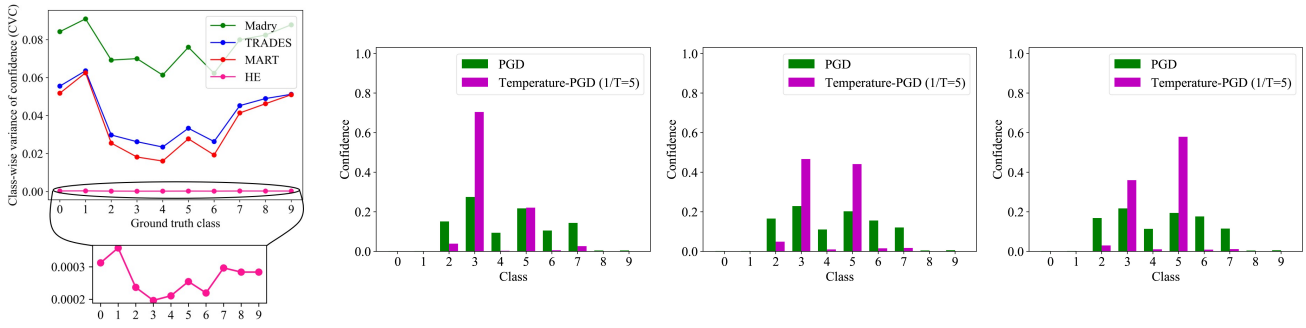
Figure 1 plots the robustness of each class at different epochs in the test set for six benchmark datasets with adversarial training, where the shaded area in each sub-figure represents the robustness gap between different classes across epochs. Considering the large number of classes in CIFAR-100 and ImageNet, we randomly sample 12 classes for a better indication. From Figure 1, we surprisingly find that there are recognizable robustness gaps between different classes for all datasets. Specifically, for SVHN, CIFAR-10, STL-10 and CIFAR-100, the class-wise robustness gaps are obvious and the largest gaps can reach 40%-50% (Figure 1(b)-1(e)). For ImageNet, since the model uses the three-stage training method [30], its class-wise robustness gap increases with the training epoch, and finally up to 80% (Figure 1(f)). Even for the simplest dataset MNIST, on which model has achieved more than 95% overall robustness, the largest class-wise robustness gap still has 6% (Figure 1(a)).

Hence, we can conclude that the class-bias learning phenomenon is also common in adversarial learning, and there are remarkable robustness discrepancies among classes, leading to unbalance/unfair class-wise robustness in adversarial training. Inspired by the phenomenon in Figure 1, we next conduct the analysis to the robust relation between classes and the impact of different attacks and defenses on class-wise robustness in the following subsections.

### 4.2 The relations among different classes

We first systematically investigate the relation of different classes under robust models. Figure 2 shows the confusion matrices of robustness between classes on all the six datasets. The X-axis and Y-axis represent the predicted classes and the ground truth classes, respectively. The grids on the main diagonal line represent the robustness of each class, while the grids on the off-diagonal line represent the non-robustness on one class (Y-axis) to be misclassified to another class (X-axis).

**Observations and Analysis.** From the results reported in Figure 2, we have the following observations and analysis: (i) The confusion matrices on all six benchmark datasets roughly demonstrate one kind of symmetry (*i.e*, the highlight colors of off-diagonal elements are symmetrical about the main diagonal), which indicates that some classes-pair could be easily misclassified between each other. (ii) The symmetry classes-pair in Figure 2 are always similar to some degree, such as similar in shape or belonging to




**(a) Class-wise variance of confidence (CVC) of SOTA defense models** **(b) MART's output for image 127 (class 3) with iteration steps 1** **(c) MART's output for image 127 (class 3) with iteration steps 10** **(d) MART's output for image 127 (class 3) with iteration steps 20**

**Figure 3: Analysis of output confidence**

**Table 3: Adversarial robustness (%) under Temperature-PGD$^{20}$ attack on CIFAR-10.**

| Defense | 1/T | Tot. | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | CV | MCD |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Madry | 2 | 51.8(-0.3)[1] | 63.4(-0.4) | 72.0(+0.4)[2] | 38.8(-0.3) | 25.2(-0.1) | 33.9(-2.8) | 38.5(-0.1) | 56.6(-0.8) | 59.8(+0.3) | 63.0(-0.1) | 66.9(+0.1) | 235.6(+11.3) | 46.8(+0.5) |
| TRADES | 5 | 54.6(-1.7) | 66.8(-1.0) | 80.0(-0.6) | 36.7(-1.1) | 26.2(-3.2) | 35.6(-5.0) | 43.0(-0.9) | 56.0(-3.3) | 66.0(-0.9) | 70.8(-1.0) | 64.9(-0.7) | 291.6(+28.0) | 53.8(+2.7) |
| MART | 5 | 54.3(-3.9) | 62.7(-1.8) | 77.1(-0.9) | 41.5(-3.6) | 26.3(-9.1) | 27.5(-10.2) | 41.5(-2.0) | 60.8(-4.5) | 66.1(-1.4) | 72.8(-3.5) | 67.3(-2.2) | 311.3(+76.2) | 50.8(+8.2) |
| HE | 5 | 57.3(-3.4) | 62.4(-2.5) | 74.8(-4.5) | 38.4(-2.6) | 29.4(-5.1) | 43.1(-4.8) | 47.8(-3.7) | 62.9(-4.7) | 69.1(-1.4) | 74.5(-2.4) | 70.8(-2.4) | 240.8(+16.0) | 45.4(+0.6) |
| HE | 50 | 50.4(-10.3) | 58.2(-6.7) | 71.8(-7.5) | 33.3(-7.7) | 17.6(-16.9) | 23.0(-24.9) | 41.6(-9.9) | 56.2(-11.4) | 66.9(-3.6) | 69.9(-7.0) | 65.7(-7.5) | 363.5(+138.7) | 54.1(+9.3) |

[1] "-" represents the robustness reduction compared with the corresponding element of PGD attack in table 1.
[2] "+" represents the robustness improvement compared with the corresponding element of PGD attack in table 1.

the same superclasses, hence, would be easy misclassified to each other. Specifically, for SVHN, digits with similar shapes are more likely to be flipped to each other, *e.g.*, the number 6 and number 8 are similar in shape and the non-robustness between them (number 6 is misclassified to be number 8 or vice versa) is very high as shown in Figure 2(d). For CIFAR-10 and STL-10, Figures 2(b) and 2(e) clearly show that the classes belonging to the same superclass have high probabilities to be misclassified to each other, for example, both class 3 (cat) and class 5 (dog) in CIFAR-10 belong to the superclass *Animals*, the non-robustness between them is very high in Figure 2(b). (iii) Few misclassifications would happen between two classes with different superclasses. For example, in STL-10, the class 5 (dog) belongs to superclass *Animals*, while class 9 (truck) belongs to *Transportation*, and their non-robustness is almost 0 as shown in figure 2(e).

For CIFAR-100 and ImageNet, we can also observe symmetry properties of confusion matrix in Figure 2(c) and Figure 2(f), which is consistent with the above analysis. Overall, Figure 2 demonstrates that the classes with similar semantic would be easier misclassified (with higher non-robustness) to each other than those with different semantics (*e.g.*, the classes belong to different superclasses).

## 4.3 The class-wise robustness under different attacks and defenses

The above analysis mainly concentrates on the performance under PGD attack. In this subsection, we investigate the class-wise robustness of state-of-the-art robust models against various popular attacks in the CIFAR-10 dataset.

The defense methods we chose include Madry training [15], TRADES [34], MART [29] and HE [19]. We train WideResNet-32-10 [33] following the original papers. White-box attacks include FGSM [10], PGD [15] and $CW_\infty$ [6], and the implementation of $CW_\infty$ follows [7]. Black-box attacks include a transfer-based and a query-based attack. The former uses a standard trained WideResNet-32-10 as the substitute model to craft adversarial examples, and the latter uses $\mathcal{N}$ attack [14]. All hyperparameters see Appendix A.2.

In order to quantitatively measure the robustness unbalance (or discrepancy) among classes, we give the definition of two statistical metrics: class-wise variance (CV) and maximum class-wise discrepancy (MCD) as follows

*Definition 4.1.* (**Class-wise Variance ,CV**) Given one dataset containing $C$ classes, the accuracy of each class $c$ is $a_c$, the average accuracy over all classes is $\overline{a} = \sum_{c=1}^{C} a_c/C$, and then CV is defined as:

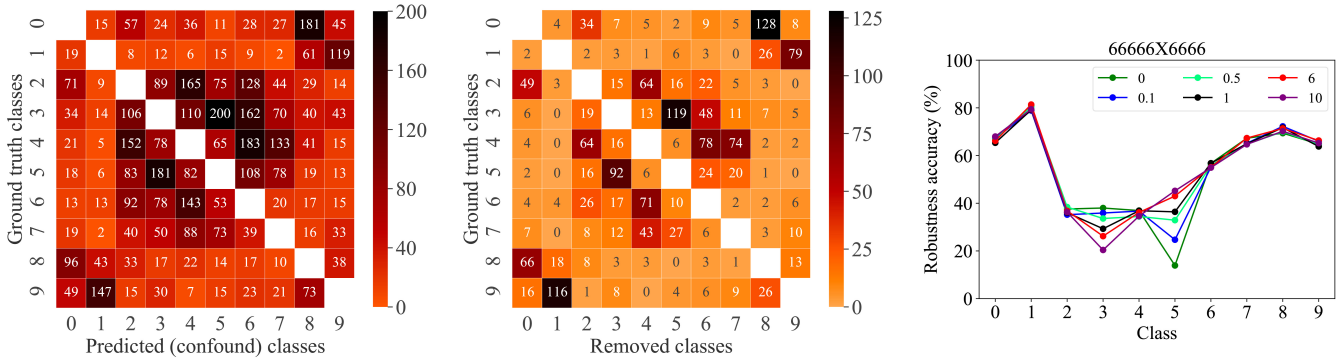
$$CV = \frac{1}{C}\sum_{c=1}^{C}(a_c - \overline{a})^2.$$

*Definition 4.2.* (**Maximum Class-wise Discrepancy, MCD**) Given one dataset, let $a_{max}$ and $a_{min}$ represent the maximum and minimum accuracy of class, then MCD is defined as:

$$MCD = a_{max} - a_{min}.$$

The insight of these two metrics is to measure the average discrepancy and the most extreme discrepancy among classes. Intuitively, these metrics will be large if there are huge class-wise differences.

**Observations and Analysis.** Based on the CIFAR-10 dataset, we check the class-wise robustness of different attack and defense models and report the results in Table 1. From the results, we can have the following observations and analysis: (i) In all models and




**(a) Misclassified confusion matrix** **(b) Homing confusion matrix** **(c) Adjust class 3 robustness**

**Figure 4: Case study about adjusting class 3 robustness in the training phase**

attacks, there are remarkable robustness gaps between different classes, and class 1 and class 3 are always the most robust and vulnerable class in all settings, which suggests the relative robustness of each class has a strong correlation with the dataset itself. (ii) Stronger attacks in white-box settings are usually more effective for vulnerable classes. For example, comparing FGSM and PGD of the same defense method, the robustness reduction of the vulnerable classes (*e.g.*, class 3) is obviously larger than that of robust classes (*e.g.*, class 1), resulting in larger class-wise variance (CV) and maximum class-wise discrepancy (MCD). (iii) In black-box settings, the main advantage of the query-based attack over the transfer-based attack is also concentrated in vulnerable classes. One explanation is that many examples of these classes are closer to the decision boundary, making it easier to be attacked.

In addition, we have also checked that the CV and MCD of the adversarial training are significantly larger than the standard training in all datasets. For example, in terms of the most popular dataset CIFAR-10, the CV of adversarial training is 28 times that of standard training, and the MCD of adversarial training is 5 times that of standard training. This shows that class-wise properties in the robustness model are worthy of attention.

## 5 IMPROVING ADVERSARIAL ATTACK VIA CLASS-WISE DISCREPANCIES

Although Section 4.3 have shown the class-wise robustness discrepancies are commonly observed in adversarial settings, we believe that this gap can be further enlarged if the attacker makes full use of the properties of vulnerable classes. Specifically, since the images near decision boundary usually have smooth confidence distributions, popular attacks cannot find the effective direction in the iterative process, and Figure 3(b)-3(d) clearly show an example of the failed attack with PGD (*i.e.*, the bar for ground truth class 3 is always the highest). To solve this problem, we propose to use a temperature factor to change this distribution, so as to create *virtual power* in the possible adversarial direction.

For a better formulation, we assume that the DNN is $f$, the input example is $x$, the number of classes in the dataset is $C$, then the softmax probability of this sample $x$ corresponding to class $k$ ($k \in C$) is

$$\mathbb{S}(f(x))_k = \frac{e^{f(x)_k/T}}{\sum_{c=1}^{C} e^{f(x)_c/T}}. \tag{3}$$

Using this improved softmax function, the adversarial perturbation crafted at $t^{\text{th}}$ step is

$$\delta^{t+1} = \prod_{\epsilon}(\delta^t + \alpha \cdot \text{sign}(\nabla \ell(\mathbb{S}(f(x+\delta^t)), y))). \tag{4}$$

Where $\prod_{\epsilon}$ is the projection operation, which ensures that $\delta$ is in $\epsilon$-ball. $\ell(\cdot)$ is the cross-entropy loss. $\alpha$ is the step size. $y$ is the ground truth class.

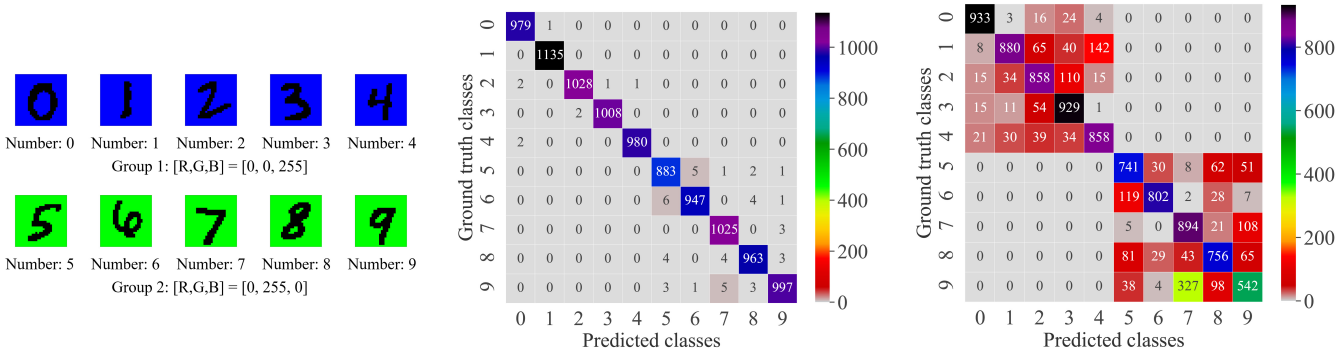
The bar corresponding to Temperature-PGD (1/T=5) in Figure 3(b)-3(d) is a good example of how our proposed method works. To better understand the impact of our method on different defense models, the class-wise variance of confidence (CVC) is proposed to measure the smoothness of the confidence output of these models.

*Definition 5.1.* (**Class-wise Variance of Confidence, CVC**) Assume that there are $C$ classes in the test set, class $k$ has $N$ images, the confidence output of one image $i$ is $\boldsymbol{p} = (p_1, \cdots, p_c, \cdots, p_C)$ and the average confidence of this image is $\overline{p}^i = \sum_{c=1}^{C} p_c^i/C$, then CVC of class $k$ is defined as

$$CVC_k = \frac{1}{NC} \sum_{i=1}^{N} \sum_{c=1}^{C} (p_c^i - \overline{p}^i)^2.$$

Intuitively, this value will be small if the output confidence of one class is smooth. From the results of Figure 3(a) and Table 2, we can find that in all defense models, the CVC of superclass *Animals* is smaller than that of superclass *Transportation*, and the CVC of class 4 and class 3 is the smallest and second-smallest. Combined with the information of Table 1, the class with low robustness is closer to the classification boundary, so the confidence distribution is smoother, which is consistent with our previous analysis. On the other hand, the overall CVC of HE is much smaller than other defense methods, which means that popular attacks (*i.e.*, PGD) may be very inefficient for this defense model.

**Overall results.** In practice, we perform a grid search on the hyperparameter 1/T $\in [2, 5, 10, 50]$, and report the best performance. The results of Table 3 verify the effectiveness of our proposed method. Specifically, (i) The CV and MCD of all defense models




**(a) Numbers after adding background** **(b) Confusion matrix for natural images** **(c) Confusion matrix for adversarial images**

**Figure 5: Experiments about adding background on MNIST**

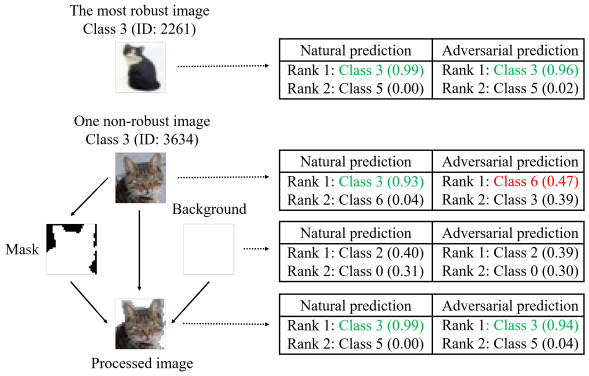



**Figure 6: Changing the background of class 3 images in the CIFAR-10 test set**

have become larger, which means that the disparity between the classes is enlarged. (ii) Since the confidence output of the vulnerable classes is smoother (Figure 3(a)), our method can significantly reduce the robustness of these classes. (*e.g.*, class 3 and class 4). (iii) The Madry's model has a steeper confidence distribution, so it is not sensitive to Temperature-PGD attack. On the contrary, because the confidence outcomes of the HE's model is extremely smooth, increasing the temperature factor to make the outcome steeper can significantly improve the attack rate, *i.e.*, total robustness reduces 10.3% (when 1/T=50), of which class 3 and class 4 reduce 16.9% and 24.9% respectively. Overall, the success of Temperature-PGD is effective evidence that stronger attackers can further increase the class-wise robustness difference.

# 6 IMPROVING THE ROBUSTNESS OF THE VULNERABLE CLASS

In this section, we propose two methods to mitigate the difference in class-wise robustness. Specifically, our goal is to improve the robustness of the most vulnerable subgroup in CIFAR-10 (*i.e.*, class 3), because it has the lowest robustness as described in the previous analysis.

## 6.1 Adjust robustness at the training phase

Figure 2 analyzes the relation of class-wise robustness in detail. Here we further explore the more fine-grained relation between these classes by removing the confound class (Definition 3.2). Specifically, for the example $\boldsymbol{x}$ from class $y$ is attacked to the confound class $y'$, we are curious if we remove confound class $y'$ (*i.e.,* remove all examples of ground truth class $y'$ in the training set) and re-train the model, will example $\boldsymbol{x}$ become a robust example WITHOUT being maliciously flipped to a new confound class[1]?

*Definition 6.1.* (**Homing Property**) Given an adversarial example $x'$ from class $y$ which is misclassified as the confound class $y'$ by a model, this example satisfies homing property if it becomes a robust example after we re-train the model via removing confound class $y'$.

To explore the above question, we conduct extensive experiments and the results are reported in Figure 4. Figure 4(a) and Figure 2(b) are similar, and the difference is that the values in Figure 4(a) represent the number of examples instead of percentage, and the main diagonal elements (the number of examples correctly classified) are hidden for better visualization and comparison. Thus this figure is called the Misclassified confusion matrix. To check the *homing property*, we alternatively remove each confound class to re-train the model and plot the results in Figure 4(b), where the element in the $i^{\text{th}}$ row and $j^{\text{th}}$ column (indexed by the classes starting from 0) indicates how many adversarial examples with ground truth class $i$ and confound class $j$ that satisfy *homing property* (*i.e.,* these examples will become robust examples after removing the confound class $j$), so this figure is defined as the Homing confusion matrix.

Figure 4 clearly shows *homing property* is widely observed in many misclassified examples. For example, we can focus on the $3^{\text{rd}}$ row and the $5^{\text{th}}$ column of Figure 4(a) and 4(b). 200 in Figure 4(a) means that 200 examples of class 3 are misclassified as class 5, and 119 in Figure 4(b) means that if we remove class 5 and re-train the model, 119 of 200 examples will *home* to the correct class 3 (*i.e.,* become robust examples). This suggests that changing the robustness of class 3 only needs to carefully handle the relation with

[1]We usually think that there are many decision boundaries in bounded $\epsilon$-ball, so a new confound class is likely to appear even if one decision boundary is removed.

**Table 4: Robust model prediction results for images of class 3 in the test set (1000 images) under Temperature-PGD$^{20}$ attack.**

| Line number | Test set | Class 0 | Class 1 | Class 2 | **Class 3(correct)** | Class 4 | Class 5 | Class 6 | Class 7 | Class 8 | Class 9 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | Original image (Natural) | 18[1] | 4 | 43 | **710** | 38 | 88 | 64 | 15 | 6 | 14 |
| 2 | + white background (Natural) | 39(+21)[2] | 10(+6) | 48(+5) | **742(+32)** | 13(-25) | 59(-29) | 63(-1) | 5(-10) | 3(-3) | 18(+4) |
| 3 | + training adjustment method (Natural) | 35(+17) | 15(+11) | 56(+13) | **758(+48)** | 23(-15) | 24(-64) | 61(-3) | 6(-9) | 8(+2) | 14(+0) |
| 4 | Original image (Adversarial) | 34 | 16 | 82 | **262** | 113 | 254 | 139 | 53 | 14 | 33 |
| 5 | + white background (Adversarial) | 76(+42) | 22(+6) | 72(-10) | **403(+141)** | 80(-33) | 136(-118) | 116(-23) | 38(-15) | 7(-7) | 50(+17) |
| 6 | + training adjustment method (Adversarial) | 93(+59) | 31(+15) | 80(-2) | **435(+173)** | 72(-41) | 84(-170) | 111(-28) | 35(-18) | 3(-11) | 56(+23) |

[1] The number represents how many images in the corresponding test set are predicted to be the corresponding class.
[2] "+" represents the increase in the number of images compared with the corresponding element of the original image test set, and "-" is vice versa.

class 5. Interestingly, these group-based relations are commonly observed in CIFAR-10, *e.g.*, class 1 (automobile)-class 9 (truck) and class 0 (airplane)-class 8 (ship).

**The proposed method.** Based on the above discovery, we try to use this group-based relation to adjust the class-wise robustness. Our method is based on TRADES [34] as shown in the Equation (5). Specifically, Zhang et al. [34] set $\beta$ as a constant to adjust natural accuracy and robust accuracy, while we modify $\beta$ to a vector $\boldsymbol{\beta} = (\beta_1, ..., \beta_c, ..., \beta_C)$ to adjust class-wise robustness, where $c \in C$ is the class id, thus the loss function of class $c$ is

$$\min_{\boldsymbol{\theta}} \frac{1}{n} \sum_{i=1}^{n} \ell\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_c^i\right), y_c^i\right) + \beta_c \max_{\left\|\boldsymbol{x}_c'^i - \boldsymbol{x}_c^i\right\|_p \leq \epsilon} \mathcal{K}\left(h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_c^i\right), h_{\boldsymbol{\theta}}\left(\boldsymbol{x}_c'^i\right)\right). \tag{5}$$

Since class 3 and class 5 have an obvious one-to-one relation in Figure 4(b). We only change the $\beta_c$ of class 5 to adjust the robustness of class 3, while fixes the $\beta_c$ of other classes. The result is shown in Figure 4(c). Each line represents the class-wise robustness under the Temperature-PGD attack. The title in the figures represents the $\beta_c$ value of each class $c$, that is, '66666X6666' stands for $\beta_c = 6$ ($\forall c \in C$ and $c \neq 5$), and the number 6 is chosen to be comparable to the experiment of Zhang et al. [34] ($\forall c \in C, \beta_c = 6$).

**Overall results.** Figure 4(c) demonstrates that the robustness of class 3 can be improved or reduced by adjusting the value of $\beta_5$. Specifically, when $\beta_5 = 6 \rightarrow \beta_5 = 0.5$, the robustness of class 3 changes from 26.2% to 33.6% and the robustness of class 5 changes from 43.0% to 34.1%, which shows that our method can effectively adjust the robustness of the most vulnerable class 3, thereby reducing the class-wise disparity. Intuitively, Other group-based relations can also be used to further balance the overall robustness.

## 6.2 Adjust robustness at the inference phase

MNIST and CIFAR-10 are the most commonly used datasets for adversarial training. However, the overall performance of robust models in MNIST usually exceeds 95%, while this is only 50%-60% in CIFAR-10. We speculate that the unified background of MNIST is one of the potential reasons why its performance is better.

To verify our assumption, we add different backgrounds to each class of images in MNIST to explore the role of the background. Specifically, we first modify the original images into three-channel images and then add two sets of background colors to the training set and test set of each class, as shown in Figure 5(a). In the training and inference phase, $\epsilon$ is set to 0.5 to highlight the robust relation between classes, and other settings are consistent with Section 4. In addition, we have also verified that this background-changing dataset has almost no effect on the accuracy of standard training. Therefore, we only report the confusion matrices of adversarial training as shown in Figure 5(b) and Figure 5(c).

The confusion pattern in Figure 5(c) is completely consistent with the background relation of each class in Figure 5(a), which is the evidence that the class-wise robust relation can be changed through the background. One possible explanation is that the model mistakenly learned the spurious correlation [22] between the foreground and the background during the training process, *e.g.*, the model may think that the number 2 and the number 3 are more similar since they have the same background, while the number 2 and the number 5 are vice versa. However, from the perspective of causality [22], the intrinsic feature to judge whether numbers are similar should be the shape rather than the background. In fact, Shen et al. [22] has proved that this phenomenon has a negative impact on model prediction, but comparing the results of Figure 5(b) and Figure 5(c), it is clearly demonstrated that the influence of the background on the adversarial examples is much greater than that on the natural examples, which makes this factor very important in adversarial settings. To the best of our knowledge, this is the first step to explore the connection between background and model robustness.

**The proposed method.** Inspired by the above phenomenon, we believe that the complex background in the CIFAR-10 dataset may affect the robustness of each class and we can use this property to adjust class-wise robustness. To check this, we first select the images with the ground truth class 3 in the test set and then record the confidence of adversarial prediction corresponding to the class 3 of each image (*i.e.*, $\mathbb{S}(h_{\boldsymbol{\theta}}(\boldsymbol{x}_i'))_{y=3}$, where $\mathbb{S}$ is the softmax function) and visualize images according to the confidence from high to low. Surprisingly, we find that the backgrounds of the highly robust images in class 3 are pure white color. Figure 6 shows the most robust image (ID: 2261) in class 3 has this white background.

Therefore, we manually extract the mask that can locate the background from one non-robust image (ID: 3634) of class 3 in the test set, and then replace the original background with a white background to investigate the change of prediction. As shown in Figure 6, the boxes represent the natural and robust prediction of the corresponding image. 'Rank 1' and 'Rank 2' represent the classes with the highest and the second-highest confidence, and the value in brackets represents the specific confidence. The result indicates this non-robust image can become a robust one by replacing the background, while it slightly affects the natural prediction.

We apply the above image processing method to all images of class 3 (1000 images) in the test set to verify whether the above phenomenon can be generalized. As shown in Table 4. The number

in each row represents how many images in the corresponding test set are predicted to be the corresponding class. Since the ground truth of all test images is class 3, the column corresponding to class 3 is the number of images that are correctly predicted.

**Overall results.** As illustrated in Line 4 and Line 5 of Table 4, many non-robust examples become robust after adding a white background (*i.e.*, the robustness changed from 26.2% to 40.3%), while Line 1 and Line 2 indicate natural predictions are not sensitive to the background, which proves that the background mainly has a great influence on the model's adversarial prediction. Furthermore, we combine the modified training method mentioned in Section 6.1, and the robustness of class 3 becomes 43.5% (Line 6), which means that the robustness of the most vulnerable class in CIFAR-10 has been greatly improved.

## 7 CONCLUSION

In this paper, we have a closer look at the class-wise properties of the robust model based on the observation that robustness between each class has a recognizable gap. We conduct systematic analysis and find: 1) In each dataset, classes can be divided into several subgroups, and intra-group classes are easily flipped by each other. 2) The emergence of the unbalanced robustness is closely related to the intrinsic properties of the datasets. Furthermore, we make full use of the properties of the vulnerable classes to propose an attack that can effectively reduce the robustness of these classes, thereby increasing the disparity among classes. Finally, in order to alleviate the robustness difference between classes, we propose two methods to improve the robustness of the most vulnerable class in CIFAR-10 (*i.e.*, class 3): 1) At the training phase: Modify loss function according to group-based relation between classes. 2) At the inference phase: Change the background of the original images. We believe our work can contribute to a more comprehensive understanding of adversarial training and let researchers realize that the class-wise properties are crucial to robust models.

## 8 ACKNOWLEDGEMENTS

This work is supported in part by Key R&D Projects of the Ministry of Science and Technology (No. 2020YFC0832500), National Natural Science Foundation of China (No. 61625107, No. 62006207), National Key Research and Development Program of China (No. 2018AAA0101900), the Fundamental Research Funds for the Central Universities and Zhejiang Province Natural Science Foundation (No. LQ21F020020). Yisen Wang is partially supported by the National Natural Science Foundation of China under Grant 62006153, and CCF-Baidu Open Fund (OF2020002).

## REFERENCES

[1] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*.
[2] Yang Bai, Yan Feng, Yisen Wang, Tao Dai, Shu-Tao Xia, and Yong Jiang. 2019. Hilbert-Based Generative Defense for Adversarial Examples. In *ICCV*.
[3] Yang Bai, Yuyuan Zeng, Yong Jiang, Yisen Wang, Shu-Tao Xia, and Weiwei Guo. 2020. Improving query efficiency of black-box adversarial attack. In *ECCV*.
[4] Yang Bai, Yuyuan Zeng, Yong Jiang, Shu-Tao Xia, Xingjun Ma, and Yisen Wang. 2021. Improving Adversarial Robustness via Channel-wise Activation Suppressing. In *ICLR*.
[5] Philipp Benz, Chaoning Zhang, Adil Karjauv, and In So Kweon. 2020. Robustness may be at odds with fairness: An empirical study on class-wise accuracy. *arXiv preprint arXiv:2010.13365* (2020).
[6] Nicholas Carlini and David Wagner. 2017. Towards evaluating the robustness of neural networks. In *S&P*.
[7] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, and John Duchi. 2019. Unlabeled data improves adversarial robustness. In *NeurIPS*.
[8] Adam Coates, Andrew Ng, and Honglak Lee. 2011. An analysis of single-layer networks in unsupervised feature learning. In *AISTATS*.
[9] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *CVPR*.
[10] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.
[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.
[12] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images. (2009).
[13] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.
[14] Yandong Li, Lijun Li, Liqiang Wang, Tong Zhang, and Boqing Gong. 2019. Nattack: Learning the distributions of adversarial examples for an improved black-box attack on deep neural networks. In *ICML*.
[15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR*.
[16] Vedant Nanda, Samuel Dooley, Sahil Singla, Soheil Feizi, and John P Dickerson. 2021. Fairness Through Robustness: Investigating Robustness Disparity in Deep Learning. In *FAccT*.
[17] Yuval Netzer, Tao Wang, Adam Coates, Alessandro Bissacco, Bo Wu, and Andrew Y Ng. 2011. Reading digits in natural images with unsupervised feature learning. (2011).
[18] Tianyu Pang, Kun Xu, and Jun Zhu. 2019. Mixup Inference: Better Exploiting Mixup to Defend Adversarial Attacks. In *ICLR*.
[19] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Jun Zhu, and Hang Su. 2020. Boosting adversarial training with hypersphere embedding. In *NeurIPS*.
[20] Edward Raff, Jared Sylvester, Steven Forsyth, and Mark McLean. 2019. Barrage of random transforms for adversarially robust defense. In *CVPR*.
[21] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. 2019. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032* (2019).
[22] Zheyan Shen, Peng Cui, Kun Kuang, Bo Li, and Peixuan Chen. 2018. Causally regularized learning with agnostic data selection bias. In *MM*.
[23] Chuanbiao Song, Kun He, Jiadong Lin, Liwei Wang, and John E Hopcroft. 2019. Robust Local Features for Improving the Generalization of Adversarial Training. In *ICLR*.
[24] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199* (2013).
[25] Kaihua Tang, Jianqiang Huang, and Hanwang Zhang. 2020. Long-Tailed Classification by Keeping the Good and Removing the Bad Momentum Causal Effect. In *NeurIPS*.
[26] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. 2021. A unified approach to interpreting and boosting adversarial transferability. In *ICLR*.
[27] Yisen Wang, Xingjun Ma, James Bailey, Jinfeng Yi, Bowen Zhou, and Quanquan Gu. 2019. On the Convergence and Robustness of Adversarial Training.. In *ICML*.
[28] Yisen Wang, Xingjun Ma, Zaiyi Chen, Yuan Luo, Jinfeng Yi, and James Bailey. 2019. Symmetric cross entropy for robust learning with noisy labels. In *ICCV*.
[29] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. 2019. Improving adversarial robustness requires revisiting misclassified examples. In *ICLR*.
[30] Eric Wong, Leslie Rice, and J Zico Kolter. 2019. Fast is better than free: Revisiting adversarial training. In *ICLR*.
[31] Dongxian Wu, Yisen Wang, Shu-Tao Xia, James Bailey, and Xingjun Ma. 2019. Skip Connections Matter: On the Transferability of Adversarial Examples Generated with ResNets. In *ICLR*.
[32] Dongxian Wu, Shu-Tao Xia, and Yisen Wang. 2020. Adversarial Weight Perturbation Helps Robust Generalization. In *NeurIPS*.
[33] Sergey Zagoruyko and Nikos Komodakis. 2016. Wide residual networks. *arXiv preprint arXiv:1605.07146* (2016).
[34] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael I Jordan. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*.
[35] Shengyu Zhang, Ziqi Tan, Zhou Zhao, Jin Yu, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, and Fei Wu. 2020. Comprehensive information integration modeling framework for video titling. In *KDD*.
[36] Shengyu Zhang, Dong Yao, Zhou Zhao, Tat-Seng Chua, and Fei Wu. 2021. CauseRec: Counterfactual User Sequence Synthesis for Sequential Recommendation. In *SIGIR*.

# A APPENDIX: HYPERPARAMETERS FOR REPRODUCIBILITY

## A.1 Hyperparameters for defenses in Section 4

**MNIST setup.** Following Zhang et al. [34], we use a four-layers CNN as the backbone. In the training phase, we adopt the SGD optimizer with momentum 0.9, weight decay $2 \times 10^{-4}$ and an initial learning rate of 0.01, which is divided by 10 at the $55^{\text{th}}$, $75^{\text{th}}$ and $90^{\text{th}}$ epoch (100 epochs in total). Both the training and testing attacker are 40-step PGD ($\text{PGD}^{40}$) with random start, maximum perturbation $\epsilon = 0.3$ and step size $\alpha = 0.01$.

**CIFAR-10 & CIFAR-100 setup.** Like Wang et al. [29] and Zhang et al. [34], we use ResNet-18 [11] as the backbone. In the training phase, we use the SGD optimizer with momentum 0.9, weight decay $2 \times 10^{-4}$ and an initial learning rate of 0.1, which is divided by 10 at the $75^{\text{th}}$ and $90^{\text{th}}$ epoch (100 epochs in total). The training and testing attackers are $\text{PGD}^{10}$/$\text{PGD}^{20}$ with random start, maximum perturbation $\epsilon = 0.031$ and step size $\alpha = 0.007$.

**SVHN & STL-10 setup.** All settings are the same to CIFAR-10 & CIFAR-100, except that the initial learning rate is 0.01.

**ImageNet setup.** Following Wong et al. [30], we use ResNet-50 [11] as the backbone. Specifically, in the training phase, we use the SGD optimizer with momentum 0.9 and weight decay $2 \times 10^{-4}$. A three-stage learning rate schedule is used as the same with Wong et al. [30]. The training attacker is FGSM [10] with random start, maximum perturbation $\epsilon = 0.007$, and the testing attacker is $\text{PGD}^{50}$ with random start, maximum perturbation $\epsilon = 0.007$ and step size $\alpha = 0.003$.

## A.2 Hyperparameters for attacks in Section 4.3

**FGSM setup.** Random start, maximum perturbation $\epsilon = 0.031$.

**PGD setup.** Random start, maximum perturbation $\epsilon = 0.031$. For RST model, step size $\epsilon = 0.01$ and steps $\alpha = 40$, following Carmon et al. [7]. For other models, step size $\epsilon = 0.003$ and steps $\alpha = 20$.

**$\text{CW}_\infty$ setup.** Binary search steps $b = 5$, maximum perturbation times $n = 1000$, learning rate $lr = 0.005$, initial constant $c_0 = 0.01$, $\tau$ decrease factor $\gamma = 0.9$. Similar to Carmon et al. [7], we randomly sample 2000 images to evaluate model robustness, and 200 images per class.

**Transfer-based attack setup.** All settings are the same to PGD for the substitute standard model.

**$\mathcal{N}$ attack setup.** Random start, maximum perturbation $\epsilon = 0.031$, population size $n_{pop} = 300$, noise standard deviation $\sigma = 0.1$ and learning rate $lr = 0.02$. Similar to Li et al. [14], we randomly sample 2000 images to evaluate model robustness, and 200 images per class.