# Topology-preserving Adversarial Training for Alleviating Natural Accuracy Degradation

Xiaoyue Mi[1,2]
mixiaoyue19s@ict.ac.cn
Fan Tang*[1,2]
tfan.108@gmail.com
Yepeng Weng[1,2]
wengyepeng15@mails.ucas.ac.cn
Danding Wang[1,2]
wangdanding@ict.ac.cn
Juan Cao[1,2]
caojuan@ict.ac.cn
Sheng Tang[1,2]
ts@ict.ac.cn
Peng Li*[3]
lipeng@air.tsinghua.edu.cn
Yang Liu[3,4]
liuyang2011@tsinghua.edu.cn

[1] Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, China

[2] University of Chinese Academy of Sciences, Beijing, China

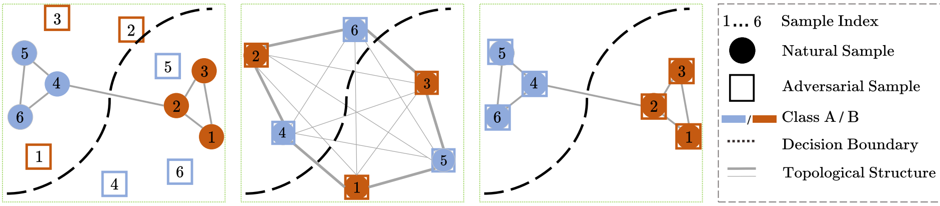
[3] Institute for AI Industry Research (AIR), Tsinghua University, Beijing, China

[4] Department of Computer Science & Technology, Tsinghua University, Beijing, China

**Abstract**

Despite the effectiveness in improving the robustness of neural networks, adversarial training has suffered from the natural accuracy degradation problem, i.e., accuracy on natural samples has reduced significantly. In this study, we reveal that natural accuracy degradation is highly related to the disruption of the natural sample topology in the representation space by quantitative and qualitative experiments. Based on this observation, we propose Topology-pReserving Adversarial traINing (TRAIN) to alleviate the problem by preserving the topology structure of natural samples from a standard model trained only on natural samples during adversarial training. As an additional regularization, our method can be combined with various popular adversarial training algorithms, taking advantage of both sides. Extensive experiments on CIFAR-10, CIFAR-100, and Tiny ImageNet show that our proposed method achieves consistent and significant improvements over various strong baselines in most cases. Specifically, without additional data, TRAIN achieves up to **8.86**% improvement in natural accuracy and **6.33**% improvement in robust accuracy.


*Corresponding author.




((a)) Standard training ((b)) Adversarial training ((c)) TRAIN

Figure 1: Illustrations for representation space under different training strategies.

# 1 Introduction

Adversarial training [10, 20, 25] has been proven to effectively defense adversarial attacks [8, 15, 17] of neural networks [2, 29]. However, models trained by adversarial training strategy have shown a significant reduction of accuracy in natural samples [17], which is usually called *natural accuracy degradation* [6]. This problem hinders the practical application of adversarial training, as natural samples are the vast majority in reality [11].

Existing works attempt to alleviate natural accuracy degradation by data augmentation or extra data collection [24, 32], distilling classifier boundary of the standard model [1, 4, 6], instance reweighting [38], early-stopping [37], adjustments of loss functions [20], and learnable attack strategies during training [12, 14]. Nevertheless, these approaches have not fully closed the natural accuracy gap between adversarial and standard training.

Unlike previous efforts, we attempt to explain the natural accuracy degradation from a new perspective, topology. Topology refers to the neighborhood relation of data in the representation space [23]. Some adversarial training studies [22, 35] have shown the importance of topology in adversarial robustness generalization. However, they do not attenuate the negative impact on the natural samples produced by the adversarial samples, resulting in incomplete topology preservation, the natural accuracy degradation still exists.

We conjecture that adversarial training destroys the topology of natural samples in the representation space, leading to a decrease in natural accuracy. As illustrated in Fig. 1(a), a model after standard training has a well-generalizing topology of natural samples but is vulnerable to adversarial samples, which are usually far from their true class distribution in the representation space. Adversarial training pulls simultaneously the adversarial samples and their corresponding natural samples nearer [18] (Fig. 1(b)) to improve the robustness of the model while leading to the poor topological structure of the natural sample features due to the negative influence of the adversarial samples. Qualitative and quantitative analyses support the intuition that natural accuracy correlates with the topology preservation extent (see Sec. 3.2 for more details).

Inspired by the above intuition, we propose a new approach called Topology-pReserving Adversarial traINing (TRAIN) to alleviate natural accuracy degradation (Fig. 1(c)), which closes the gap between adversarial and corresponding natural samples while preserving the well-generalizing topology of the standard model. A straightforward solution is to distill the natural sample features of the standard model or the relationships based on the absolute distance between samples during adversarial training. However, it suffers from optimization difficulties due to the great gap between standard and adversarial models. So we construct the topological structure of data in the representation space based on the neighbor graph for each model. We define the edge weight of the graph as the probability that different samples are neighbors, and topology preservation is achieved by aligning the standard model's graph and the adversarial model's graph. Meanwhile, the optimization process of the standard model is not affected by the adversarial model, to reduce the negative impact of adversarial samples. Experiments show that benefitting from topology preservation, TRAIN improves both the natural and robust accuracy when combined with other adversarial training algorithms.

Our contributions are as follows:

- We reveal that the topology of natural samples in the representation space plays an important role in the natural accuracy of adversarial models, which provides a new perspective on mitigating natural accuracy degradation.
- We propose a *topology preservation adversarial training* method that preserves the topology structure between natural samples in the standard model representation space, which can be combined with various adversarial training methods.
- Extensive quantitative and qualitative experiments on CIFAR-10, CIFAR-100, and Tiny ImageNet datasets show the effectiveness of the proposed TRAIN (maximum 8.86% improvement for the natural accuracy and 6.33% for the robust accuracy).

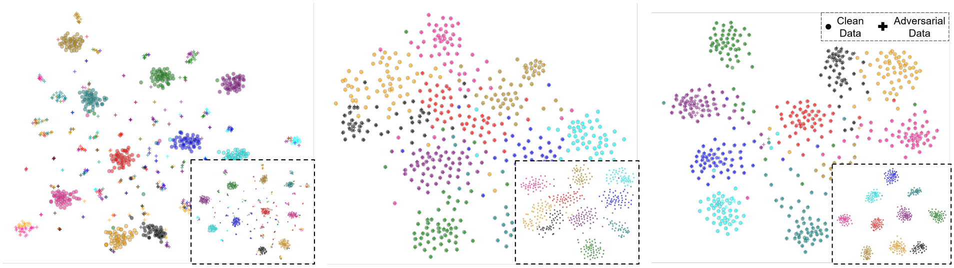
# 2 Related Work

## 2.1 Adversarial Training

Adversarial training [12, 16, 17, 26, 31, 33, 34, 36] is a prevailing method to improve the adversarial robustness of DNNs. However, it decreases the accuracy of natural samples while increasing the adversarial robustness compared with standard training. This phenomenon is called "natural accuracy degradation" or "the trade-off between robustness and accuracy". Several works have been proposed to alleviate this problem. Zhang *et al.* [37] used early-stopping. Rebuffi *et al.* [24] tried to use more training data by data augmentation or adding extra data. Researchers [1, 4, 6] tried to distill the natural sample logits from the standard model to the adversarial model. Zhang *et al.* [38] made use of instance reweighting. Pang *et al.* [20] redefined adversarial training optimization goals. And Jia *et al.* [12] used reinforcement learning to obtain learnable attack strategies. Different from them, we mitigate this problem from the view of the topology of different data in the representation space. Some works [22, 35] show topology is crucial for adversarial robustness generalization but ignore the negative impact of adversarial examples, and still degrade in natural accuracy.

## 2.2 Knowledge Distillation in Adversarial Training

Knowledge distillation can transfer knowledge from a larger, cumbersome model (teacher) to a smaller, more efficient model (student), which is commonly used for model compression. Recently some algorithms have applied knowledge distillation to adversarial training. Some works [9, 39] distilled large robust models for robust model compression. Different from them, researchers [1, 4, 6] distilled the natural data logits of the standard model to enhance adversarial training on natural accuracy. [4] considered additional temperature factors during distillation. However, they did not constrain the topology of samples in the representation space, and their distillation loss updates both standard and adversarial models simultaneously. Therefore, they were still negatively affected by adversarial examples. The experimental section also includes comparative evaluations of different knowledge distillation methods.




((a)) Standard (77.39%/0.00%) ((b)) TRADES (56.50%/30.93%) ((c)) TRAIN (65.28%/33.97%)

Figure 2: Analytical experiments reveal the relationship between topology quality in the representation space and natural accuracy. (a), (b), and (c) show the differences in the representation space for the standard model, adversarial model (trained by TRADES with $\beta = 6.0$), and TRAIN on CIFAR-100 training (small plots) and test sets (large plots). Natural accuracy and PGD-20 accuracy are indicated in red and blue, respectively.

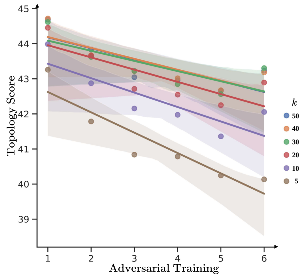



Figure 3: Quantitative analysis reveals a negative correlation between the adversarial strength and the topology score.

# 3 Topology's Role in Adversarial Training

## 3.1 Formulation

Following vanilla AT [17], the goal of adversarial training is defined as:

$$\arg\min_{\theta} \mathbb{E}_{(x,y)\in D} \left( \max_{\delta \in S} L(x+\delta, y; \theta) \right), \tag{1}$$

where $D$ is the data distribution for input $x$ and its corresponding label $y$, $\theta$ is the model parameters. $\delta$ stands for the perturbation applied to $x$ and is usually limited by perturbation size $\varepsilon$. $S = \left\{ \delta | \left\| \delta \right\|_p \leq \varepsilon \right\}$ is the feasible domain for $\delta$. $L(\cdot)$ usually is the cross-entropy loss for classification. By min-max gaming, adversarial training aims to correctly recognize all adversarial examples ($x' = x + \delta$). For descriptive purposes, we refer to models trained only on natural samples as *standard models* and those trained using adversarial training as *adversarial models* in the latter part.

## 3.2 Empirical Analysis

In this section, we analyze how adversarial training influences topological relationships compared with standard models. We find that the quality of the topology is positively correlated with natural accuracy, while negatively correlated with adversarial strength. Adversarial models are trained by TRADES [36], and here we consider the weights of the adversarial loss function $\beta$ as adversarial strength. Larger $\beta$ represents the greater strength of adversarial training. We choose the penultimate layer representations (before logits) of the standard model and adversarial models for qualitative and quantitative experimental analysis. **See supplementary material for more details on experimental settings.**

**Qualitative analysis.** As shown in Fig. 2, compared with the standard model on both training the test sets, the representation visualization for adversarial models shows more robustness, but a worse topology of data resulting in lower discrimination in different classes.

**Quantitative analysis.** We conduct quantitative analysis by setting the $\beta = 1, 2, ..., 6$ for TRADES as different adversarial strengths, and use $k$NN accuracy as the *topology score* to evaluate the quality of topology for different models, which is often used in manifold learning [19, 30] to evaluate the quality of topology in dimension reduction. The higher the score, the more reasonable topology between the samples. Specifically, we use both natural and adversarial data (generated by PGD-20) in the CIFAR-100 training set as the support set to predict the labels of natural and adversarial data in the test set. To verify the reliability of the observation conclusion, we choose $k = 5, 10, 20, 30, 40, 50$, respectively.

Fig. 3 shows the strength of adversarial training and their corresponding topology qualities for different $k$. A negative correlation between the strength of adversarial training and the topology quality could be observed.

**Why does adversarial training destroy topological relationships?** Adversarial representations are usually far away from their true class distribution, while natural samples are not. Adversarial training narrowing the adversarial representations and natural representations concurrently usually makes the representation of natural samples further away from the original distribution, and hurts the topology and discrimination of natural data representations. Zhang *et al*. [38] points out that adversarial training is equivalent to a special kind of regularization and has a strong smoothing effect, which also supports our intuition.

# 4 Topology-Preserving Adversarial Training

## 4.1 Overall Framework

To reduce the negative impact of adversarial samples, we propose a method TRAIN that focuses on preserving the topology of natural features from the standard model during adversarial training. As shown in Alg. 1, we train two models simultaneously: a standard model $M$ with a cross-entropy loss $L_{\mathrm{ST}}(\cdot)$ and an adversarial model $M'$ which is updated by a specific adversarial training algorithm. For natural sample $x_i$, the outputs of $M(x_i)$ are the feature of the last layer $f_{x_i}$ and logit $logit_{x_i}$. Similarly, the outputs of $M'(x_i')$ are $f'_{x_i'}$ and $logit'_{x_i'}$ for aversarial sample $x_i'$ and $f'_{x_i}$ and $logit'_{x_i}$ for natural sample $x_i$.

The loss $L_{\mathrm{ST}}(\cdot)$ of $M$ is formulated as:

$$L_{\mathrm{ST}} = L(z_{x_i}, y_i), z_{x_i} = \frac{exp(logit_{x_i})}{\sum_{j=1}^{N} exp(logit_{x_j})}, \tag{2}$$

where $L$ is cross-entropy loss. And the overall loss $L_{\mathrm{AT}}(\cdot)$ of $M'$ is formulated as follows:

$$L_{\mathrm{AT}} = L_{\mathrm{robust}}(x') + \lambda L_{\mathrm{TP}}(M, M'), \tag{3}$$

where $L_{\mathrm{robust}}(\cdot)$ denotes the adversarial robustness loss, which is determined by the specific adversarial training algorithm employed. Additionally $L_{\mathrm{TP}}(\cdot)$ serves as a regularization item to preserve the topology of natural samples from $M$ and updates only $M'$. A comprehensive discussion regarding the specifics of $L_{\text{TP}}(\cdot)$ will be discussed in the next subsection.

## 4.2 Topology Preservation in Adversarial Training

The topological structure is typically based on a neighborhood relation graph constructed by the similarity among samples in the representation space [19, 23, 30]. In this graph, each point is a sample in the representation space, while the edges are relationships among the samples, and the weights assigned to the edges are determined by the similarity between the samples. Consequently, the topology preservation can be precisely formulated as follows:

$$L_{\text{TP}} = \mathbb{E}_{(x,y)\in D}\left(F(P,Q)\right), \tag{4}$$

where $P$ and $Q$ represent the neighborhood relation graph constructed by the inter-sample similarity for $M$ and $M'$, respectively. $F(\cdot)$ measures the similarity between two graphs.

**Absolute relationship preservation.** Directly applying cosine similarity to calculate the pairwise distances $d_{ij}$ and $d'_{ij}$ between samples in representation spaces of $M$ and $M'$ to construct the neighborhood relation graph $P$ and $Q$ is a straightforward way:

$$P = \{d_{ij}|0 < i, j \leq N\}, Q = \{d'_{ij}|0 < i, j \leq N\}, \tag{5}$$

where $d_{ij}$ and $d'_{ij}$ are defined as:

$$d_{ij} = 1 - \frac{f_{x_i}^T f_{x_j}}{||f_{x_i}||_2||f_{x_j}||_2}, \widetilde{d'_{ij}} = 1 - \frac{f'^T_{x'_i} f'_{x'_j}}{||f'_{x'_i}||_2||f'_{x'_j}||_2}. \tag{6}$$

However, there exists a substantial difference in the representation space between the adversarial model and the standard model, making it challenging to optimize the preservation of direct absolute relationships.

**Relative relationship preservation.** Considering the significant gap between standard and adversarial models, our objective is to use conditional probability distribution for modeling the relationships between samples. Specifically, we define the edge weights of the neighborhood relation graph as the probability that distinct samples are neighbors, thus ensuring topology preservation through the alignment of the probability distributions of the two graphs.

Different from manifold learning [19, 30] which uses the regular Kernel Density Estimation (KDE) for approximations of the conditional probabilities, we use the cosine similarity-based affinity metric. This choice is motivated by the excessive hyper-parameter tuning requirements and unacceptable training costs associated with KDE in adversarial training.

$$\begin{aligned} K_{cos}(f_{x_i}, f_{x_j}) =& \frac{1}{2}\left(\frac{f_{x_i}^T f_{x_j}}{||f_{x_i}||_2||f_{x_j}||_2} + 1\right), \\ =& \frac{1}{2}(2 - d_{ij}), \end{aligned} \tag{7}$$

where $K_{cos}$ is cosine similarity-based affinity metric value for $x_i$ and $x_j$.

Moreover, we add a special term $\rho_j$ to better preserve the global structure of representation space. $\rho_j$ represents the distance from the $j_{th}$ data point to its nearest neighbor. Subtracting $\rho_j$ ensures the local connectivity of the graph, avoiding isolated points and thus better preserves the global structure.

$$\widetilde{d_{ij}} = d_{ij} - \rho_j, \ \widetilde{d'_{ij}} = d'_{ij} - \rho'_j. \tag{8}$$

**Algorithm 1** Topology-Preserving Adversarial Training

**Require:** the step size of perturbations $\varepsilon$, batch size $n$, learning rate $\alpha$, attack algorithm optimization iteration times $K$, the number of training epochs $T$, adversarial model $M'$ with its parameters $\theta'$, standard model $M$ with its parameters $\theta$, loss weight $\lambda$ and training dataset $(x,y) \in D$

**Ensure:** robust model $M'$ with $\theta'$

Randomly initialize $\theta$ , $\theta'$
**for** $i = 1,...,T$ **do**
    Sampling a random mini-batch $X = \{x_1, x_2, ..., x_n\}$ and corresponding labels $Y = \{y_1, y_2, ..., y_n\}$ from $D$
    Generating adversarial data $X' = \{x'_1, x'_2, ..., x'_n\}$ through attack algorithms (such as PGD-K, FGSM)
    $f_X, \text{logit}_X = M(X)$
    $f'_{X'}, \text{logit}'_{X'} = M'(X')$
    Evaluate $L_{\text{ST}}$ Eq. (2)
    Evaluate $L_{\text{AT}} = \lambda L_{\text{TP}} + L_{\text{robust}}$ Eq. (3)
    Update model parameters:
    $\theta = \theta - \alpha \frac{1}{n} \sum_{i=1}^{n} \nabla_\theta L_{\text{ST}}$
    $\theta' = \theta' - \alpha \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta'} L_{\text{AT}}$
**end for**

After normalization, we obtain the $p_{i|j}$, which represents the conditional probability that the $i_{th}$ natural sample is a neighbor of the $j_{th}$ natural sample in the representation space of $M$.

$$p_{i|j} = \frac{2 - \widetilde{d}_{ij}}{\sum_{k=1,k\neq j}^{N} (2 - \widetilde{d}_{jk})}. \tag{9}$$

Similarly, for the adversarial model $M'$:

$$q_{i|j} = \frac{2 - \widetilde{d}'_{ij}}{\sum_{k=1,k\neq j}^{N} (2 - \widetilde{d}'_{jk})}. \tag{10}$$

So the neighborhood relation graph construction of $M$ can be formalized as:

$$P = \left\{ p_{i|j} \middle| p_{i|j} = \frac{2 - \widetilde{d}_{ij}}{\sum_{k=1,k\neq j}^{N} (2 - \widetilde{d}_{jk}))}, 0 < i, j \leq N \right\}. \tag{11}$$

Similarly, the relationship graph for $M'$ is:

$$Q = \left\{ q_{i|j} \middle| q_{i|j} = \frac{2 - \widetilde{d}'_{ij}}{\sum_{k=1,k\neq j}^{N} (2 - \widetilde{d}'_{jk})}, 0 < i, j \leq N \right\}. \tag{12}$$

We use cross-entropy loss to measure the similarity of $P$ and $Q$ for such flexible relationships. Finally, the $L_{\text{TP}}$ for TRAIN is formalized as:

$$\begin{aligned} L_{\text{TP}} &= CE(P, Q) \\ &= \sum_i \sum_j \left[ p_{i|j} \log\left(\frac{p_{i|j}}{q_{i|j}}\right) + (1 - p_{i|j}) \log\left(\frac{1 - p_{i|j}}{1 - q_{i|j}}\right) \right]. \end{aligned} \tag{13}$$

((a)) Standard training ((b)) TRADES ($\beta = 6.0$) ((c)) TRAIN

Figure 4: t-SNE visualizations of penultimate layer features on CIFAR-10. Crosses and circles are adversarial samples and natural samples, respectively. Different colors represent different classes.

# 5 Experiments

**Experimental settings** Following [6, 12, 20], we conduct extensive evaluations on popular datasets in adversarial training, including CIFAR-10, CIFAR-100 [13]. ResNet-18 is the backbone of standard models, and WideResNet-34-10 is the backbone of adversarial models. The adopted adversarial attacking method during training is PGD-10, with a perturbation size $\varepsilon = 0.031$, a step size of perturbations $\varepsilon_1 = 0.007$. For different experiment settings, we choose different $\lambda$. We set $\lambda = 5$ on CIFAR-10 dataset, and $\lambda = 20a$ on CIFAR-100 dataset, where $a = \frac{2}{1+e^{-\frac{10t}{100}}-1}$ and $t$ is the current $t$-th epoch during training. Finally, all experiments were done on GeForce RTX 3090.

Our evaluation metrics are natural data accuracy (Natural Acc.) and robust accuracy (Robust Acc.). Robust accuracy is the model classification accuracy under adversarial attacks. Following previous works, we choose three representative adversarial attack methods for evaluation: PGD-20, C&W-20 [3], and Auto Attack [5]. We denote the model's defense success rate under those attacks separately as *PGD-20 Acc.*, *C&W-20 Acc.*, and *AA Acc.*. To provide a comprehensive evaluation and comparison with other state-of-the-art adversarial training methods, we use their original hyperparameters in our settings, and include baselines: Vanilia AT [17], TRADES [36], LBGAT, MART [31], FAT [37], GAIRAT [38], AWP [34], SAT [27], LAS [12], and ECAS [14]. For TRADES, we set $\beta = 6.0$. For LBGAT, we conduct experiments based on vanilla AT and TRADES ($\beta = 6.0$). We also provide **details of experimental settings** and **experiments on Tiny ImageNet [7] in supplementary materials**.

## 5.1 Main Results

**Quantitative results.** As shown in Tables 1 and 2, TRAIN achieves a better trade-off between natural accuracy and adversarial robustness compared with the most popular adversarial training algorithms. TRADES, LBGAT, and ECAS achieve significant improvement in natural accuracy by combining with TRAIN, and the robust accuracy is also relatively improved or preserved.

**Qualitative analysis.** To showcase the efficacy of our algorithm in assisting the adversarial model in constructing a well-generalizing topology in the representation space, we use t-SNE to visualize samples from ten randomly selected categories in the CIFAR-100 test set and all categories of the CIFAR-10 test set for qualitative analysis. Figs. 2 and 4 show the results of CIFAR-100/10 datasets, respectively. For standard training (Figs. 2(a) and 4(a)), the natural data exhibit clear clustering, while the adversarial samples appear disjointed, resulting in poor performance on robust accuracy. The TRADES approach facilitates the alignment of

| Defense | Natural Acc. | Robust Acc. | | |
|---|---|---|---|---|
| | | PGD-20 Acc. | C&W-20 Acc. | AA Acc. |
| Vanilla AT* [17] | 85.17 | 55.08 | 53.91 | 51.69 |
| MART* [31] | 84.17 | <u>58.56</u> | 54.58 | 51.10 |
| FAT* [37] | **87.97** | 49.86 | 48.65 | 47.48 |
| GAIRAT* [38] | 86.30 | **59.54** | 45.57 | 40.30 |
| AWP* [34] | 85.57 | 58.13 | 56.03 | 53.90 |
| ECAS [14] | 84.57 | 55.86 | 54.65 | 52.10 |
| ECAS+TRAIN | 85.26(↑ 0.69) | 56.23(↑ 0.37) | 54.77(↑ 0.12) | 52.22(↑ 0.12) |
| TRADES* [36] | 85.72 | 56.10 | 53.87 | 53.40 |
| LAS-TRADES* [12] | 85.24 | 57.07 | 55.45 | 54.15 |
| TRADES + TRAIN | <u>87.07</u>(↑ 1.35) | 58.51(↑ 2.41) | **56.81**(↑ 2.94) | **54.70**(↑ 1.30) |
| TRADES + LBGAT [6] | 80.20 | 57.41 | 54.84 | 53.32 |
| TRADES + LBGAT +TRAIN | 86.69(↑ 6.49) | 58.04(↑0.63) | <u>56.75</u>(↑ 1.91) | <u>54.47</u>(↑ 1.15) |

Table 1: Results on CIFAR-10. "*" are the results directly quoted from LAS. The best and second best results are **bolded** and <u>underlined</u>, respectively.

| Defense | Natural Acc. | Robust Acc. | | |
|---|---|---|---|---|
| | | PGD-20 Acc. | C&W-20 Acc. | AA Acc. |
| Vanilla AT* [17] | 60.89 | 31.69 | 30.10 | 27.86 |
| SAT* [27] | 62.82 | 27.17 | 27.32 | 24.57 |
| AWP* [34] | 60.38 | 33.86 | 31.12 | 28.86 |
| ECAS [14] | 64.60 | 35.41 | <u>33.39</u> | 29.55 |
| ECAS+TRAIN | 65.24(↑ 0.64) | **35.83**(↑ 0.42) | **33.50**(↑ 0.11) | **30.69**(↑ 1.14) |
| TRADES* [36] | 58.61 | 28.66 | 27.05 | 25.94 |
| LAS-TRADES* [12] | 60.62 | 32.53 | 29.51 | 28.12 |
| TRADES + TRAIN | **67.47**(↑ 8.86) | 34.99(↑ 6.33) | 31.61(↑ 4.56) | 28.95(↑ 3.01) |
| TRADES + LBGAT* [6] | 60.64 | 34.75 | 30.65 | 29.33 |
| TRADES + LBGAT+ TRAIN | <u>65.40</u>(↑ 4.76) | <u>35.46</u>(↑ 0.71) | 32.36(↑ 1.71) | <u>30.17</u>(↑ 0.84) |

Table 2: Results on CIFAR-100. "*" are the results directly quoted from LAS. The best and second best results are marked in **bold** and <u>underline</u>.

natural and adversarial data to enhance robust accuracy. Nonetheless, it is noteworthy that this alignment process can unintentionally disrupt the integrity of natural feature topologies, as it lacks any defensive measures to counteract this effect (refer to Figs. 2(b) and 4(b) in the paper for visual representations of this phenomenon). As shown in Figs. 2(c) and 4(c), applying the proposed TRAIN to TRADES could drive the cluster for each category to be more compact, thereby preserving the topology more effectively.

## 5.2 Ablation Studies

In this section, we delve into TRAIN to study its effectiveness in different relation-preserving methods We present a comparative analysis of our proposed method with alternative approaches: a metric learning approach called MCA [35] and two absolute relationship distillation methods, namely RKD [21] and CRD [28]. MCA applies a supervised contrastive loss into adversarial training. RKD takes the absolute value of the cosine distance between samples as the relationship as discussed in Sec. 4. CRD requires that a sample's representation in the student model be closer to its corresponding representation in the teacher model while

| Methods | Natural Acc | Robust Acc | | |
|---|---|---|---|---|
| | | PGD-20 Acc | C&W-20 Acc | AA Acc |
| Vanilla AT [17] | 60.44 | 28.06 | 27.85 | 24.81 |
| MCA [35] | 57.18 | 29.31 | 27.23 | 25.76 |
| Vanilla AT + RKD [21] | 64.00 | 28.32 | 27.92 | 24.92 |
| Vanilla AT + CRD [28] | 62.22 | 27.47 | 27.42 | 24.53 |
| Vanilla AT + TRAIN′ | 62.10 | 29.43 | 29.66 | 25.78 |
| Vanilla AT + TRAIN | **66.39** | **29.88** | **29.84** | **25.81** |

Table 3: Ablation results on different relation-preserving methods.

| **Vanilla AT** | **Vanilla AT + LBGAT** | **Vanilla AT + TRAIN** |
|---|---|---|
| 821 | 848 | 849 |
| **TRADES** | **TRADES + LBGAT** | **TRADES + TRAIN** |
| 1,079 | 1,106 | 1,109 |

Table 4: Training time in second of an epoch on one RTX 3090 GPU.

being farther from the representations of other samples in the teacher model. TRAIN′ means adversarial training will influence the standard models during training. Table 3 shows the effectiveness of the relative relationship preservation TRAIN′ means adversarial training will influence the standard models during training, and it is important to reduce the negative influence of adversarial samples (comparison between TRAIN and TRAIN′). All the ablation experiments are based on the CIFAR-100 dataset and combined with TRADES, and we provide its other training details in the supplementary material.

**Time complexity.** Our method is based on batch computation, and its time complexity is $O(N(mz'K)) + O(N(bz(fz' + fz) + mz)$, where $mz'$ and $mz$ is the number of neurons of the adversarial model (48.32 M) and standard model (11.22 M), $bz$ is the batch size (128), $fz'$ and $fz$ is the feature size of the standard model (512) and adversarial model (640), and $K$ is the number of iterations for generating adversarial examples (10). For classic adversarial training, its time complexity is $O(N(mz'K)$. Since $bz(fz' + fz) + mz << mz'K$, the additional time overhead of our method is negligible.

Note that, the primary time-consuming factor in adversarial training algorithms lies in the need for additional backpropagation during the generation of adversarial samples, whereas the TRAIN algorithm does not incur any extra computational cost in this regard. Table 4 shows the time statistics for training one epoch (with batch size equals 128) by different baselines. It takes an additional 28 seconds when combined with Vanilia AT and 3% (30 seconds) on TRADES for TRAIN, which is as fast as LBGAT. We also analyze **the influence of batch size, hyper-parameter $\lambda$, and model architectures in supplementary material.**

# 6 Conclusion

Compared with standard training, adversarial training shows significant natural accuracy degradation. Different from previous algorithms, we assume this is due to topology disruption of natural features, and confirm it by empirical experiments. Based on that, we propose Topology-pReserving Adversarial traINing (TRAIN). While improving the adversarial robustness of the model, it preserves the topology of natural samples in the representation space of the standard model. Our method has been rigorously validated through both quantitative and qualitative experiments, demonstrating its effectiveness and reliability.

# 7 Acknowledgments

This work is supported by the National Natural Science Foundation of China (62203425).

# References

[1] Elahe Arani, Fahad Sarfraz, and Bahram Zonooz. Adversarial concurrent training: Optimizing robustness and accuracy trade-off of deep neural networks. *arXiv preprint*

*arXiv:2008.07015*, 2020.

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International conference on machine learning*, pages 274–283. PMLR, 2018.

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

[4] Erh-Chung Chen and Che-Rung Lee. Ltd: Low temperature distillation for robust adversarial training. *arXiv preprint arXiv:2111.02331*, 2021.

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

[6] Jiequan Cui, Shu Liu, Liwei Wang, and Jiaya Jia. Learnable boundary guided adversarial training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15721–15730, 2021.

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[8] Haoran Gao, Hua Zhang, Xin Zhang, Wenmin Li, Jiahui Wang, and Fei Gao. Enhanced covertness class discriminative universal adversarial perturbations. *Neural Networks*, 165:516–526, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.06.006. URL https://www.sciencedirect.com/science/article/pii/S0893608023003076.

[9] Micah Goldblum, Liam Fowl, Soheil Feizi, and Tom Goldstein. Adversarially robust distillation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 3996–4003, 2020.

[10] Lirong He, Qingzhong Ai, Xincheng Yang, Yazhou Ren, Qifan Wang, and Zenglin Xu. Boosting adversarial robustness via self-paced adversarial training. *Neural Networks*, 167:706–714, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.08.063. URL https://www.sciencedirect.com/science/article/pii/S0893608023004938.

[11] Pei Huang, Yuting Yang, Minghao Liu, Fuqi Jia, Feifei Ma, and Jian Zhang. $\varepsilon$-weakened robustness of deep neural networks. In *Proceedings of the 31st ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA 2022, page 126–138, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450393799. doi: 10.1145/3533767.3534373. URL https://doi.org/10.1145/3533767.3534373.

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

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

[14] Hui Kuurila-Zhang, Haoyu Chen, and Guoying Zhao. Adaptive adversarial norm space for efficient adversarial training. In *34th British Machine Vision Conference Proceedings*. BMVA Press, 2023.

[15] Huanhuan Li, Wenbo Yu, and He Huang. Strengthening transferability of adversarial examples by adaptive inertia and amplitude spectrum dropout. *Neural Networks*, 165:925–937, 2023. ISSN 0893-6080. doi: https://doi.org/10.1016/j.neunet.2023.06.031. URL https://www.sciencedirect.com/science/article/pii/S089360802300343X.

[16] Tao Li, Yingwen Wu, Sizhe Chen, Kun Fang, and Xiaolin Huang. Subspace adversarial training. *arXiv preprint arXiv:2111.12229*, 2021.

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

[18] Chengzhi Mao, Ziyuan Zhong, Junfeng Yang, Carl Vondrick, and Baishakhi Ray. Metric learning for adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[19] Leland McInnes, John Healy, and James Melville. UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[20] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. 2022.

[21] Wonpyo Park, Dongju Kim, Yan Lu, and Minsu Cho. Relational knowledge distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3967–3976, 2019.

[22] Zhuang Qian, Shufei Zhang, Kaizhu Huang, Qiufeng Wang, Rui Zhang, and Xinping Yi. Improving model robustness with latent distribution locally and globally. *arXiv preprint arXiv:2107.04401*, 2021.

[23] Juan Ramon Rabunal, Julian Dorado, and Alejandro Pazos Sierra. *Encyclopedia of artificial intelligence*. IGI Global, 2009.

[24] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan Andrei Calian, Florian Stimberg, Olivia Wiles, and Timothy A Mann. Data augmentation can improve robustness. *Advances in Neural Information Processing Systems*, 34:29935–29948, 2021.

[25] Lukas Schott, Jonas Rauber, Matthias Bethge, and Wieland Brendel. Towards the first adversarially robust neural network model on mnist. *arXiv preprint arXiv:1805.09190*, 2018.

[26] Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36, 2023.

[27] Chawin Sitawarin, Supriyo Chakraborty, and David Wagner. Sat: Improving adversarial training via curriculum-based loss smoothing. In *Proceedings of the 14th ACM Workshop on Artificial Intelligence and Security*, pages 25–36, 2021.

[28] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive representation distillation. In *International Conference on Learning Representations*, 2019.

[29] Florian Tramer, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. *Advances in Neural Information Processing Systems*, 33:1633–1645, 2020.

[30] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of machine learning research*, 9(11), 2008.

[31] Yisen Wang, Difan Zou, Jinfeng Yi, James Bailey, Xingjun Ma, and Quanquan Gu. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

[32] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 36246–36263. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/wang23ad.html.

[33] Eric Wong, Leslie Rice, and J Zico Kolter. Fast is better than free: Revisiting adversarial training. *arXiv preprint arXiv:2001.03994*, 2020.

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

[35] Shuo Yang, Zeyu Feng, Pei Du, Bo Du, and Chang Xu. Structure-aware stabilization of adversarial robustness with massive contrastive adversaries. In *2021 IEEE International Conference on Data Mining (ICDM)*, pages 807–816. IEEE, 2021.

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

[37] Jingfeng Zhang, Xilie Xu, Bo Han, Gang Niu, Lizhen Cui, Masashi Sugiyama, and Mohan Kankanhalli. Attacks which do not kill training make adversarial learning stronger. In *International conference on machine learning*, pages 11278–11287. PMLR, 2020.

[38] Jingfeng Zhang, Jianing Zhu, Gang Niu, Bo Han, Masashi Sugiyama, and Mohan S Kankanhalli. Geometry-aware instance-reweighted adversarial training. In *ICLR*, 2021.

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