FLEXIBLE RELATION PRESERVING FOR ADVERSARIAL TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

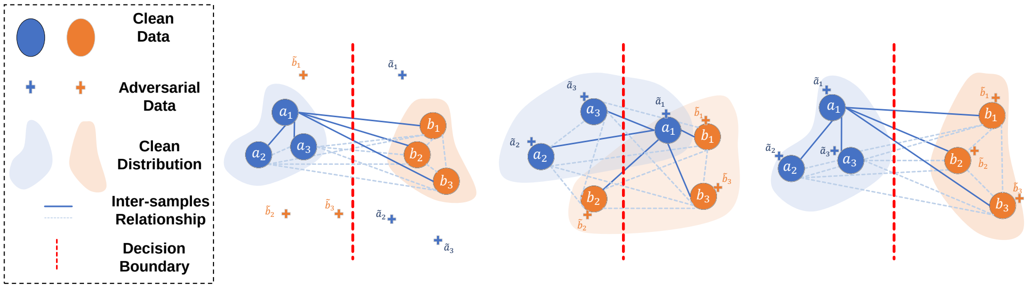
In this study, we revisit the representation learning problem for adversarial training from the perspective of relation preservation. Typical adversarial training methods tend to pull clean and adversarial samples closer to improve robustness. However, our experimental analysis reveals that such operation would lead to cluttered feature representations thus decreasing the accuracy for both clean and adversarial samples. To alleviate the problem, we build a robust discriminative feature space for both clean and adversarial samples by taking into account a relational prior which preserves the relationship between features of clean samples. A flexible relationship preserving adversarial training (FRPAT) strategy is proposed to transfer the well-generalized relational structure of the standard training model into the adversarial training model. Moreover, it acts as an extra regularization term mathematically, making it easy to be combined with various popular adversarial training algorithms in a plug-and-play way to achieve the best of both worlds. Extensive experiments on CIFAR10 and CIFAR100 demonstrate the superiority of our algorithm. Without additional data, it improves clean generalizability up to **8.78%** and robust generalizability up to **3.04%** on these datasets.

1 INTRODUCTION

Deep neural networks have a tremendous impact on various research directions, such as self-driving (Bojarski et al., 2016), speech recognition (Nassif et al., 2019), machine translation (Stahlberg, 2020), and more. However, DNNs are observed to be vulnerable to adversarial examples, which are normal data with human imperceptible perturbations (Szegedy et al., 2014). Recently, various adversarial defense methods (Madry et al., 2017; Xie et al., 2017; Dhillon et al., 2018; Zhang et al., 2019; Bashivan et al., 2021; Sarkar et al., 2021) have been proposed. Adversarial training proves to be the most powerful way to improve adversarial robustness by generating adversarial examples as data augmentation during training (Schott et al., 2018; Pang et al., 2021; Maini et al., 2020).

Compared with standard training, adversarial training methods (Madry et al., 2017; Zhang et al., 2019; Wang et al., 2019; Li et al., 2021) improve the model robustness by aligning the representations of clean data and adversarial samples or classifying the adversarial data correctly. As illustrated in Fig. 1a, generated adversarial samples are distributed differently to natural samples and misclassified by standard training models; by narrowing the distance between natural and corresponding adversarial samples, adversarial training models could handle part of adversarial samples. As shown in Fig. 1b, the feature representation for clean samples are influenced by adversarial training and thus results in cluttered over-smoothing feature space. Some existing works try to mitigate the over-smoothing representation by distilling the logits of the standard training model (Cui et al., 2021; Chen & Lee, 2021; Arani et al., 2020). However, these point-wise distilling models ignore the geometric properties of the feature space which is important for improving the model generalization (Belkin et al., 2006).

In this paper, we qualitatively and quantitatively analyzed the correlation between the adversarial strengths and the inter-sample relationships. Visualizations also proof that standard training will make the features for clean and adversarial samples distributed unevenly while adversarial training tends to make the features less discriminate. We build k-nearest neighbor graph with the features of clean and adversarial samples and then measure the manifold quality by conducting k-NN classification on the graph. Different adversarial strengths under various neighbor numbers are tested. Results show an



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

Figure 1: Illustrations for feature space under different training conditions. (a) In the representation space of the standard training, the clean samples are clearly separated, but the corresponding adversarial samples are far and misclassified. (b) The adversarial training representation is more robust to adversarial attacks, but the clean data representations are negatively affected by the adversarial ones, and the inter-sample relationship is destroyed. (c) Our method maintains the inter-sample relationship while bring clean and adversarial representations together, thus forming a robust and discriminative representation space.

obvious connection between the adversarial strength and the quality of geometric structure of the feature space (See Sec. 2.2 for more details).

Based on these observations, we propose a flexible relation preserving adversarial training (FPRAT) approach to keep the feature relation of standard training models. Two models are involved for FRPAT: one for standard training on clean samples and the other for adversarial training. Graph for each model is built based on the relationship of different samples. Considering the great gap between standard and adversarial models because of adversarial training smoothness effects, we define flexibly the relationship of samples as the probability that different samples are neighbors, and relational distillation is achieved by aligning the probability distributions of the two graphs. As illustrated in Fig. 1c, FPRAT preserves well-generalizable inter-sample relationships of clean samples from the standard model, to avoid the clean representations being pulled away by the adversarial representations during adversarial training. Our contributions are as follows:

- We reveal that adversarial training strength is negatively correlated with inter-sample relationships in representation spaces, which provides a new view for solving the generalization problem of adversarial robustness.
- We propose a *flexible relation preserving regularization* to flexibly preserve the inter-sample relationship structure during adversarial training, which could work in a plug-and-play way combined with various adversarial training approaches.
- Extensive quantitative and qualitative experiments on both CIFAR10 and CIFAR100 datasets show the effectiveness of the proposed FRPAT (maximum 8.78% improvement for the clean sample accuracy and 3.04% for the robust accuracy).

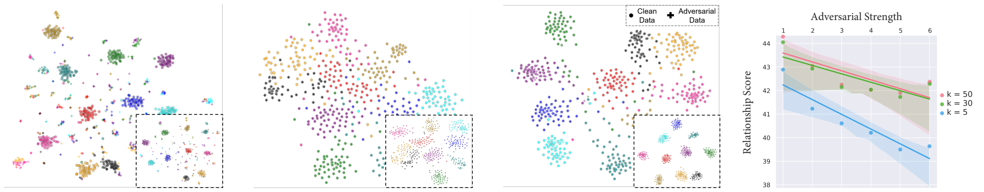
2 Methods

2.1 Preliminary

Adversarial training generates adversarial examples as training data to defend against adversarial attacks. Madry et al. (2017) makes use of projected gradient descent (PGD) to generate adversarial data, and for the first time formally define the goal of adversarial training 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, θ is the model parameters. δ stands for the perturbation applied to x and is usually limited by perturbation size ϵ. $S =$



(a) Standard training (77.39%/0.00%)

(b) TRADES ($\beta = 6.0$) (56.50%/30.93%)

(c) Ours (65.28%/33.97%)

(d) Correlation between adversarial training strength and inter-sample relationships

Figure 2: Analytical experiments on adversarial training strength and sample relationship quality in feature space. (a), (b) and (c) show the difference in feature space between the standard training model, adversarial training model, and ours on CIFAR-100 training set (small plots) and test set (big plots). The red value is the clean accuracy, and the blue value is the PGD-20 accuracy. The penultimate layer of features of the network is visualized using t-SNE. (d) shows negative correlations between adversarial training strength and inter-samples relationship quality.

$\left\{\delta| \left\|\delta\right\|_p \leq \epsilon\right\}$ is the feasible domain for δ. $L(\cdot)$ usually is the cross-entropy loss for classification. By min-max gaming, adversarial training aims to correctly recognize all adversarial examples ($\widetilde{x} = x + \delta$).

2.2 Empirical Analysis of the feature structure for adversarial training

In this section, we analyze how adversarial training influences feature relationships compared with standard training models. Here standard training models refer to the model without adversarial training, and their architectures are ResNet18. Adversarial training models are trained by TRADES with different β which means different adversarial training strengths. We choose the penultimate layer representations (before logits) of the standard training model and adversarial training models for qualitative and quantitative experimental analysis. As shown in Fig. 2a, compared with the standard training model both on the test set and training set, the representation visualization for adversarial training models shows more robustness, but worse relationships among different data resulting in lower discrimination in different classes.

To accurately analyze the correlation between adversarial training strength and the relationship between samples, we conduct quantitative analysis by setting the $\beta = 1, 2, ..., 6$ for TRADES. The larger β represents the greater strength of adversarial training. We use k-NN to evaluate the quality of inter-sample relationships for different models, which is often used in manifold learning (Van der Maaten & Hinton, 2008; McInnes et al., 2018) to evaluate the quality of manifolds. Specifically, we first use PGD-20 ($\epsilon = 0.031$) to generate adversarial data for CIFAR-100 dataset, then we use both clean and adversarial data in the training set as the support set to predict the labels of all the examples in the test set. To verify the reliability of the observation conclusion, we choose $k = 5, 30, 50$, respectively. Finally, the k-NN accuracy is used as the relationship score for the learned feature space. The higher the score, the more reasonable the relationship between the samples. Fig. 2d shows the strength of adversarial training for different models and their corresponding relationship qualities for different k. A negative correlation between the strength of adversarial training and the relationship quality between samples could be observed.

Why does adversarial training destroy inter-sample relationships? Adversarial representations are usually far away from their true class distribution. Therefore, the existing adversarial training algorithms will make the representation of clean samples further away from true class distribution while narrowing the adversarial representation and clean representation. Compared with the standard training model, the relationship between samples of the adversarial training model is worse. Zhang et al. (2021) point out that adversarial training is equivalent to a special kind of regularization and has a strong smoothing effect, which also supports our view. To mitigate the negative impact of close adversarial samples and clean samples, and maintain the inter-sample relationship, we use the inter-sample relationship of the standard training model as prior knowledge to guide the adversarial training and improve the generalization of the model.

2.3 Flexible Relation Preserving Adversarial Training

To relieve the overwhelming smoothing effect Zhang et al. (2021) caused by adversarial training, we consider building a robust discriminative feature space for both clean samples and adversarial samples by transferring the structure prior contained in the model trained on clean data to the adversarial training model.

We train two models simultaneously: a standard training model M on clean input x with cross-entropy loss $L_{st}(\cdot)$ and an adversarial training model $\widetilde{M}$ on $\widetilde{x}$ which is updated by a specific adversarial training algorithm. Furthermore, to prevent the feature space of $\widetilde{M}$ from being over smoothing, we transfer the knowledge of M to $\widetilde{M}$. The overall optimizing objective $L_{at}(\cdot)$ of $\widetilde{M}$ is formulated as:

$$minL_{robust}(\widetilde{x}) + \lambda L_{RP}(M, \widetilde{M}), \tag{2}$$

where $L_{robust}(\cdot)$ stands for the adversarial training loss and $L_{RP}(\cdot)$ works as a regularization item to bound the over smoothing feature distribution of $\widetilde{M}$ by the learned discriminative feature space of M.

Absolute Relationship Preservation. The geometry/structural knowledge means the relationship graph constructed by the similarity between samples or approximated as the manifolds they form. It can be formalized as:

$$\min_{\theta} \mathbb{E}_{(x,y)\in D} \left(F(P, Q) \right), \tag{3}$$

where P and Q stand for the relationship graph constructed by the inter-sample similarity for M and $\widetilde{M}$, respectively. $F(\cdot)$ measures the similarity of two relationships. A straightforward way to construct the relation graph P and Q could be directly applying cosine distance to calculate the distances of any two samples in the high dimensional feature space of M and $\widetilde{M}$:

$$P = \{d_{ij} | 0 < i, j \leq N\}, Q = \{\widetilde{d}_{ij} | 0 < i, j \leq N\}, \tag{4}$$

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

$$d_{ij} = 1 - \frac{f(x_i)^T f(x_j)}{||f(x_i)||_2 ||f(x_j)||_2}, \widetilde{d}_{ij} = 1 - \frac{\widetilde{f}(\widetilde{x}_i)^T \widetilde{f}(\widetilde{x}_j)}{||\widetilde{f}(\widetilde{x}_i)||_2 ||\widetilde{f}(\widetilde{x}_j)||_2}. \tag{5}$$

However, there is a huge difference in the feature space between adversarial training and standard training due to the over-smoothing in adversarial training. Thus direct absolute relationship distillation is difficult to optimize.

Flexible Relationship Preservation. Considering the great gap between standard and adversarial models because of adversarial training smoothness effects, we define flexibly the relationship of samples as the probability that different samples are neighbors, and relational distillation is achieved by aligning the probability distributions of the two graphs; we model the conditional probability distribution with a cosine similarity-based affinity metric for relation graph construction:

$$P = \left\{ p_{i|j} \middle| p_{i|j} = \frac{(2 - (d_{ij} - \rho_j))}{\sum_{k=1,k\neq j}^{N} (2 - (d_{jk} - \rho_j))}, 0 < i, j \leq N \right\}, \tag{6}$$

where $p_{i|j}$ is the conditional probability that the i_{th} clean sample is the neighbor of the j_{th} clean sample in the feature space of M. ρ_j represents the distance from the j_{th} data point to its nearest neighbor. Subtracting ρ_j ensures the local connectivity of the manifold, avoiding isolated points and thus better preserves the global structure (McInnes et al., 2018). Similarly, the relationship graph for $\widetilde{M}$ is:

$$Q = \left\{ q_{i|j} \middle| q_{i|j} = \frac{(2 - (\widetilde{d}_{ij} - \rho_j))}{\sum_{k=1,k\neq j}^{N} (2 - (\widetilde{d}_{jk} - \rho_j))}, 0 < i, j \leq N \right\}. \tag{7}$$

We use cross-entropy loss to measure the similarity of P and Q for such flexible relationships. Finally, the L_{RP} for FRPAT is:

$$L_{RP} = CE(P, Q) = \sum_i \sum_j \left[p_{i|j} \log \left(\frac{p_{i|j}}{q_{i|j}} \right) + \left(1 - p_{i|j} \right) \log \left(\frac{1 - p_{i|j}}{1 - q_{i|j}} \right) \right]. \tag{8}$$

Note that the proposed FRPAT could be applied to other adversarial training methods in a plug-and-play way. Alg. 1 shows the overall process for the flexible relation preserving adversarial training method.

Discussion LBGAT (Cui et al., 2021) also involves two models and transfers the prior knowledge of M to $\widetilde{M}$. By directly distilling the logits of clean samples x from M to the logits of corresponding adversarial sample $\widetilde{x}$ on $\widetilde{M}$, LBGAT guides the adversarial training model's feature boundary of different categories to inherit from the clean model. Since the clean model is trained independently from $\widetilde{x}$, the logits of x by M is quite different from that for $\widetilde{x}$ by $\widetilde{M}$. Thus LBGAT needs to update not only $\widetilde{M}$ but also M by $L_{RP}(\cdot)$, which conversely introduces the smoothing effect to M (see Sec. 2.2 for more details). In this study, we propose to build a robust discriminative feature space for $\widetilde{M}$ by transferring the geometry/structural knowledge of M. 4

Algorithm 1 Flexible Relational Preserving for Adversarial Training

```
Require: the step size of perturbations $\epsilon$, batch size $n$, learning rate $\alpha$, attack algorithm optimization iteration times $K$, the number of training epochs $T$, adversarial training model $\widetilde{M}$ with its parameters $\widetilde{\theta}$, standard training model $M$ with its parameters $\theta$, loss weight $\lambda$ and training dataset $(x, y) \in D$
Ensure: robust model $\widetilde{M}$ with $\widetilde{\theta}$
Randomly initialize $\theta$ , $\widetilde{\theta}$
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 $\widetilde{X} = \{\widetilde{x}_1, \widetilde{x}_2, ..., \widetilde{x}_n\}$ through attack algorithms (such as PGD, FGSM)
    $f_X, logit_X = M(X)$
    $\widetilde{f}_{\widetilde{X}}, logit_{\widetilde{X}} = \widetilde{M}(\widetilde{X})$
    Evaluate $L_{st} = CE(softmax(logit_X))$
    Evaluate $L_{at} = \lambda L_{RP} + L_{robust}$
    Update model parameters:
    $\theta = \theta - \alpha \frac{1}{n} \sum_{i=1}^{n} \nabla_{\theta} L_{st}$
    $\widetilde{\theta} = \widetilde{\theta} - \alpha \frac{1}{n} \sum_{i=1}^{n} \nabla_{\widetilde{\theta}} L_{at}$
end for
```

3 Experiments

3.1 Experimental Settings

Datasets We choose CIFAR10 and CIFAR100 datasets (Krizhevsky et al., 2009) to validate our algorithm. CIFAR10 contains $60,000$ images with the category number of ten in total, in which $50,000$ are in the training set and $10,000$ in the test set. CIFAR100 contains $60,000$ images from 100 categories, and the numbers of images in the training and testing set are the same as CIFAR10. Following Cui et al. (2021), the input size of each image is 32×32, and the training data is normalized to $[0, 1]$ after standard data augmentation: random crops of 4 pixels padding size and random horizontal flip. The test set is normalized to $[0, 1]$ without any extra augmentation. We also report the results of Tiny ImageNet (Deng et al., 2009) in supplementary materials.

Training details For fair comparisons with LBGAT (Cui et al., 2021) which also involves two models for adversarial training, we follow the same network configurations: ResNet18 for standard training and WideResNet-34-10 for adversarial training. Following LBGAT, the adopted adversarial attacking method during training is PGD-10, with a perturbation size $\epsilon = 0.031$, a step size of perturbations $\epsilon_1 = 0.007$. The initial learning rate is set to 0.1 with a total of 100 epochs for training and reduced to 0.1x at the 75-th and 90-th epochs. The optimization algorithm is SGD, with the momentum of 0.9 and weight decay of 2×10^{-4}.

Baselines We choose three strong baselines to show our method's effectiveness: Vanilia AT (Madry et al., 2017), TRADES (Zhang et al., 2019), and LBGAT (Cui et al., 2021). For TRADES, we set $\beta = 6.0$. For LBGAT, we conduct experiments based on Vanilla AT and TRADES ($\beta = 6.0$).

Table 1: Quantitative results on CIFAR10. "*" are the results directly quoted from LBGAT.

Defense	**Clean Acc.**	**Robust Acc.**			**Relationship Score**	
		PGD-20 Acc.	**C&W-20 Acc.**	**AA Acc.**	**Clean**	**Robust**
Standard Training	94.46	0	0	0	94.94	-
Vanilla AT	86.69	53.45	53.72	48.95	86.51	53.94
Vanilla AT + Ours	**88.85**(↑ 2.16)	**55.64** (↑ 2.19)	**56.18** (↑ 2.46)	**50.89**(↑ 1.94)	**89.11**(↑ 2.60)	**56.55**(↑ 2.61)
Vanilla AT + LBGAT	86.55	54.34	53.35	47.27	86.64	54.26
Vanilla AT + LBGAT + Ours	**89.42**(↑ 2.87)	**56.21**(↑ 1.87)	**57.48**(↑ 4.13)	**51.77**(↑ 4.50)	**89.25**(↑ 2.61)	**56.59**(↑ 2.33)
TRADES	84.42*	56.59*	54.91*	51.91*	85.58	56.73
TRADES + Ours	**87.30**(↑ 2.88)	**58.20**(↑ 1.61)	**56.31** (↑ 1.40)	**53.09** (↑ 1.18)	**90.01**(↑ 4.43)	**58.86**(↑ 2.13)
TRADES + LBGAT	81.98*	**57.78***	55.53*	53.14*	84.57	**57.79**
TRADES + LBGAT + Ours	**87.62**(↑ 5.64)	57.73	**58.08**(↑ 2.55)	**53.64**(↑ 0.50)	**89.50**(↑ 5.00)	57.02

Table 2: Quantitative results on CIFAR-100. "*" are the results directly quoted from LBGAT.

Defense	**Clean Acc.**	**Robust Acc.**			**Relationship Score**	
		PGD-20 Acc.	**C&W-20 Acc.**	**AA Acc.**	**Clean**	**Robust**
Standard Training	77.39	0	0	0	77.07	-
Vanilla AT	60.44	28.06	27.85	24.81	57.17	31.32
Vanilla AT + Ours	**66.39**(↑ 5.95)	**29.88** (↑ 1.82)	**29.84** (↑ 1.99)	**25.81** (↑ 1.00)	**64.70**(↑ 7.53)	**32.84**(↑ 1.52)
Vanilla AT + LBGAT	61.01	**30.10**	28.09	25.63	61.28	30.47
Vanilla AT + LBGAT + Ours	**68.20**(↑ 7.19)	29.83	**30.84**(↑ 2.75)	**25.88**(↑ 0.25)	**66.08**(↑ 4.80)	**32.48**(↑ 2.01)
TRADES	56.50*	30.93*	28.43*	26.87*	52.57	32.17
TRADES + Ours	**65.28(↑ 8.78)**	**33.97**(↑ 3.04)	**30.86**(↑ 2.43)	**28.25** (↑ 1.38)	**65.78** (↑ 13.21)	**34.53** (↑ 2.36)
TRADES + LBGAT	60.43*	35.50*	31.50*	**29.34***	61.06	37.52
TRADES + LBGAT + Ours	**62.62**(↑ 2.19)	**36.27**(↑ 0.77)	**31.72** (↑ 0.22)	29.19	**64.84** (↑ 3.78)	**38.25**(↑ 0.73)

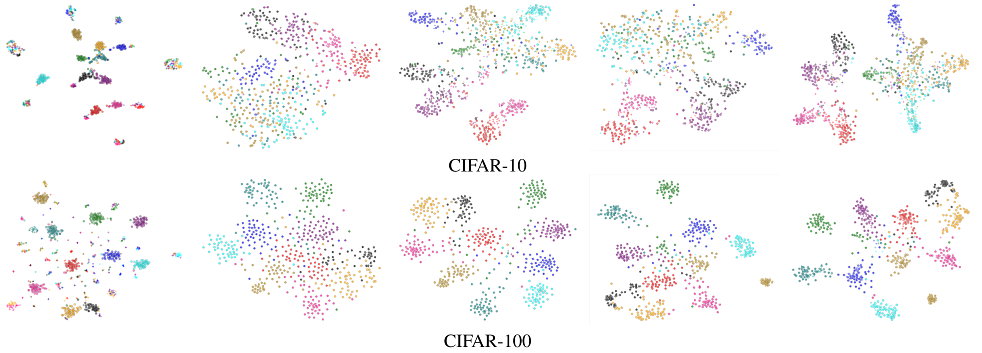
In addition, we combine FRPAT with them to demonstrate the superiority of our approach. All experiments were done on GeForce RTX 3090 with the same training configurations such as the number of epochs and learning rate schedule.

Evaluation metrics In order to evaluate the generalization of the model on clean and adversarial samples, our evaluation metrics are clean data accuracy (Clean Acc.) and robust accuracy (Robust Acc.). Robust accuracy is the model classification accuracy under adversarial attacks. We choose three representative adversarial attack methods for evaluation: PGD-20, C&W-20 (Carlini & Wagner, 2017) and Auto Attack (Croce & Hein, 2020). We denote the model's defense success rate under their attacks separately as *PGD-20 Acc.*, *C&W-20 Acc.*, and *AA Acc.* What's more, we use training sets as support sets and make KNN test accuracy as a relationship score following manifold analysis methods (Van der Maaten & Hinton, 2008; McInnes et al., 2018) under the PGD-20 attack.

3.2 Main Results

We conducted adequate quantitative analysis on CIFAR10 and CIFAR100, and the results show that the proposed flexible relation preserving knowledge transfer can be combined with other adversarial training algorithms to improve the data relational structure of the feature space and increase the accuracy of both clean and adversarial samples. In the following, we will present the experimental results for each of the two datasets.

Results on CIFAR-10 According to Table. 1, our FRPAT gets an improvement by 2.16% compared to Vanilla AT baseline on clean data. It surpasses Vanilla AT on PGD-20, C&W and AA accuracy by 2.19%, 2.46%, 1.94% respectively, indicating its high robustness. Our method also has an edge on LBGAT by 3.6% to 1.3% in all aspects. For another common baseline, TRADES, FRPAT also gets competitive results on both clean and adversarial data. Note that clean accuracy decreases when applying LBGAT to TRADES, so it also brings a large enhancement when combined with our method. For the relationship score which is measured by KNN accuracy, FRPAT could boost the performance by a large margin. Since KNN classification is based only on inter-sample relationships, such results prove that FRPAT could help build a discriminative feature space for both clean and adversarial samples.



(a) Standard (b) TRADES (c) TRADES + Ours (d) TRADES + LBGAT (e) TRADES + LBGAT + Ours

Figure 3: CIFAR-10/100 t-SNE visualizations. A total of 512 samples are selected. Crosses and circles are adversarial samples and clean samples, respectively. Different colors represent different classes.

Results on CIFAR-100 The overall results on CIFAR-100 are similar to CIFAR-10. As shown in Table. 2, FRPAT performs better than Vanilla AT and LBGAT and gets a further improvement when deployed with LBGAT simultaneously. For TRADES, our method surpasses LBGAT by a large margin (4.85%) on clean data, with the cost of slight drops (around 1%) on robust data. Adding LBGAT to FRPAT causes a decrease in clean accuracy but achieves the best accuracy in PGD-20 and C&W-20. The above results show that the proposed FRPAT could be applied to popular adversarial training pipelines for achieving SOTA performance on both clean accuracy and robust accuracy. Furthermore, we report the results of using the experimental setting as Jia et al. (2022) in supplementary materials.

Qualitative analysis To demonstrate that our algorithm can indeed help the adversarial training model construct a uniform discriminative feature space, we use t-SNE to visualize samples from ten randomly selected categories in CIFAR-100 test set and all categories of the CIFAR-10 test set for qualitative analysis. Fig. 3 shows the results. For standard training (Fig. 3a), the clean data are well clustered, however, the adversarial samples are out of place, resulting in poor performance in robust accuracy. TRADES guides the clean and adversarial data to be close to each other, which could improve the robust accuracy but the feature space is less discriminative (Fig. 3b). As shown in Fig. 3c, applying the proposed FRPAT to TRADES could drive the cluster for each category be more compact. Compared with TRADES+LBGAT (Fig. 3d), we have fewer misclassified samples in the middle section. Finally, Fig. 3e shows the result of the combination of FRPAT and LBGAT on TRADES, and the visualization for the clustering effect is also a composite of them.

3.3 Ablation Studies

In this section, we delve into FPRAT to study its effectiveness in many aspects. All the ablation experiments are based on CIFAR-100 dataset. The experimental settings are the same as the main results.

Different relation preserving approaches. In this study, we propose a flexible relation preserving knowledge transfer approach to learn discriminative features for both clean samples and robust samples. In this section, we compare our method with possible alternatives: one metric learning approach MCA (Yang et al., 2021) and two absolute relationship distillation methods RKD (Park et al., 2019) and CRD (Tian et al., 2019). MCA applies supervised contrastive loss into adversarial training. RKD takes the absolute value of the cosine distance between samples as the relationship as discussed in Sec. 2.3. To distill the teacher model's structure in feature space, CRD requires that the same sample representation of the student model is closer to the teacher model, and farther from the representation of other samples in the teacher model.

Table 4 shows the statistics. Compared with vanilla AT, MCA improves robustness while decreasing clean generalization, which is consistent with our conjecture: only supervised contrastive loss will exacerbate the negative influence of the clean samples from the adversarial samples in the feature space. Compared with RKD, CRD prefers to be exactly the same as the teacher model, but this is not necessary for the adversarial training task: we only need to maintain the relative relationship between the clean examples of the standard trained model. What's more, adversarial training models are very different from standard training models, and forcing samples to have exactly the same features is detrimental to learning. It can also be seen from the results that our method of maintaining the relative relationship between samples achieves the optimum results.

Time costing for training. Table 3 shows the time statistics for training one epoch (with batch size equals 128) by different baselines. It takes additional 28 seconds when combined with Vanilia AT and 30 seconds on TRADES for FRPAT, which is as fast as LBGAT.

Table 3: Time cost comparisons. We show the number of seconds required for different algorithms to train an epoch on one RTX 3090 GPU.

Vanilla AT	**Vanilla AT + LBGAT**	**Vanilla AT + Ours**	**TRADES**	**TRADES + LBGAT**	**TRADES + Ours**
821	848	849	1079	1106	1109

The impact of different batch sizes. As shown in Table 5, we tried 128, 256, 384 samples per batch for relation calculating. Among them, a batch size of 256 achieves the best results, but the difference among different configurations is not large. Overall our method is not sensitive to different batch sizes.

Table 4: The ablation experiment about different relation preserving approaches.

Methods	**Clean Acc**	**Robust Acc**		
		PGD-20 Acc	**C&W-20 Acc**	**AA Acc**
Vanilla AT	60.44	28.06	27.85	24.81
MCA	57.18	29.31	27.23	25.76
Vanilla AT + RKD	64.00	28.32	27.92	24.92
Vanilla AT + CRD	62.22	27.47	27.42	24.53
Vanilla AT + Ours	**66.39**	**29.88**	**29.84**	**25.81**

Table 5: The ablation experiment about different batch sizes.

Batch Size	128	256	384
Clean Acc.	66.39	**66.55**	66.26
PGD-20 Acc.	29.88	**31.08**	30.60
C&W-20 Acc.	29.84	**30.72**	30.16
AA Acc.	25.81	**26.07**	25.41

Sensitivity analysis of hyper-parameter λ. As Table 6 shows, with the increase of λ in Eq. 2, clean accuracy always gets higher, while the PGD-20 accuracy rises at first and then decreases. It is reasonable because a large λ forces the manifold to be highly close to that of standard training, thus the robustness of the model is weakened.

Table 6: Sensitivity analysis of hyper-parameter λ.

	0	5	10	20	50
Clean Acc.	57.99%	61.52%	63.21%	65.28%	66.40%
PGD-20 Acc.	31.53%	32.31%	33.47%	33.90%	33.62%

4 Related Work

Adversarial Training Adversarial training is known as one of the most effective methods to improve the adversarial robustness of DNNs. Most adversarial training algorithms (Madry et al., 2017; Zhang et al., 2019; Wang et al., 2019; Wong et al., 2020; Wu et al., 2020; Li et al., 2021; Jia et al., 2022) focused on getting clean samples close to the representation of the adversarial samples they generate. Madry et al. (2017) generated adversarial examples by PGD attack method as model input during training, and based on that, Zhang et al. (2019) proposed TRADES by punishing the model for outputting different logits of adversarial examples and their corresponding natural images, which as a regularization term adding to the cross-entropy loss. Considering the influence of misclassified data, Wang et al. (2019) introduced Misclassification Aware adveRsarial Training (MART), which emphasizes misclassified examples by higher weights. Due to the large computational cost of adversarial training, Wong et al. (2020); Li et al. (2021) use single-step attacks to obtain adversarial examples that greatly reduce training time. Wu et al. (2020) improved the adversarial robustness by flattening the weight loss landscape. Jia et al. (2022) proposed LAS from the view of automatically generating adversarial examples of proper epsilon. However, They all ignored the impact of adversarial training on the relationship between samples, resulting in cluttered feature space.

There are also works (Mao et al., 2019; Yang et al., 2021; Fan et al., 2021; Bui et al., 2021) that relate to the relationship between different samples. Mao et al. (2019) introduced metric learning in the field of adversarial training, using triplet loss to enhance model generalization. Yang et al. (2021); Fan et al. (2021); Bui et al. (2021) further used contrastive learning. However, they recognized the huge difference in the distribution of adversarial and clean samples, but did not notice the excessive impact on clean samples when narrowing them, which resulted in that they did not effectively improve the sample relationship.

Knowledge Distillation in Adversarial Training Knowledge distillation (Passalis & Tefas, 2018; Tung & Mori, 2019; Park et al., 2019; Zhu et al., 2021) can transfer the ability of the teacher network to the student network and is often used to achieve model compression. Goldblum et al. (2020); Zi et al. (2021) distilled large robust models for robust model compression. Arani et al. (2020); Cui et al. (2021); Chen & Lee (2021) distilled the clean data logits of the standard training model to enhance adversarial training on clean accuracy. Chen & Lee (2021) considered additional temperature factor during distillation. However, they did not constrain the relationship between samples, and their distillation loss updates both standard and adversarially-trained models at the same time. Therefore, they were also negatively affected by adversarial examples.

5 Conclusions and Future Work

Adversarial training shows significant over smoothing in the model feature space and results in poor generalization. Different from previous algorithms, we propose Flexible Relation Preserving for Adversarial Training (FRPAT) from the perspective of inter-sample relationships. It improves the clustering in the adversarial training feature space by migrating the relationships between clean samples of the standard training model. Because the adversarial training will make the clean samples close to the feature distribution of their generated adversarial samples, the clean sample generalization and robust generalization will be improved. FRPAT is simple yet effective. On the CIFAR10 and CIFAR100 datasets, we get a maximum improvement of 8.78% in clean sample accuracy and 3.04% in robust accuracy, demonstrating that our method does help the adversarial training model to constitute a better inter-sample relationship through visualization.

There are a few future directions we plan to pursue. First, FRPAT is a label-free algorithm, so it can be naturally applied to various unlabeled data. Second, we can distill the inter-sample relationships of pre-trained models, even across modalities, since FRPAT is only related to the feature space regardless of the label space and the model architecture. Third, FRPAT introduces a new perspective to the combination of transfer learning and adversarial robustness. We transfer the relationship between samples as a whole instead of one-to-one, which could be an inspiration for other works.

Reproducibility Statement

We provide core source codes in the supplementary material, and detailed experimental settings in the paper.

Ethics Statement

FRPAT can be used in combination with other adversarial training algorithms to further improve the classification accuracy of the model for unknown samples, enhance the robust security of the model, and reduce the possibility of criminals using adversarial attacks to cause nefarious effects. But there is still a lot of room for improvement to achieve a fully robust model.

REFERENCES

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.

Pouya Bashivan, Reza Bayat, Adam Ibrahim, Kartik Ahuja, Mojtaba Faramarzi, Touraj Laleh, Blake Richards, and Irina Rish. Adversarial feature desensitization. *Advances in Neural Information Processing Systems*, 34, 2021.

Mikhail Belkin, Partha Niyogi, and Vikas Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of machine learning research*, 7(11), 2006.

Mariusz Bojarski, Davide Del Testa, Daniel Dworakowski, Bernhard Firner, Beat Flepp, Prasoon Goyal, Lawrence D Jackel, Mathew Monfort, Urs Muller, Jiakai Zhang, et al. End to end learning for self-driving cars. *arXiv preprint arXiv:1604.07316*, 2016.

Anh Bui, Trung Le, He Zhao, Paul Montague, Seyit Camtepe, and Dinh Phung. Understanding and achieving efficient robustness with adversarial supervised contrastive learning. *arXiv preprint arXiv:2101.10027*, 2021.

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

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

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

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*, pp. 15721–15730, 2021.

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*, pp. 248–255. Ieee, 2009.

Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. *arXiv preprint arXiv:1803.01442*, 2018.

Lijie Fan, Sijia Liu, Pin-Yu Chen, Gaoyuan Zhang, and Chuang Gan. When does contrastive learning preserve adversarial robustness from pretraining to finetuning? *Advances in Neural Information Processing Systems*, 34:21480–21492, 2021.

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

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

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

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

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

Pratyush Maini, Eric Wong, and Zico Kolter. Adversarial robustness against the union of multiple perturbation models. In *International Conference on Machine Learning*, pp. 6640–6650. PMLR, 2020.

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

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

Ali Bou Nassif, Ismail Shahin, Imtinan Attili, Mohammad Azzeh, and Khaled Shaalan. Speech recognition using deep neural networks: A systematic review. *IEEE access*, 7:19143–19165, 2019.

Tianyu Pang, Xiao Yang, Yinpeng Dong, Hang Su, and Jun Zhu. Bag of tricks for adversarial training. In *International Conference on Learning Representations*, 2021.

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*, pp. 3967–3976, 2019.

Nikolaos Passalis and Anastasios Tefas. Learning deep representations with probabilistic knowledge transfer. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pp. 268–284, 2018.

Anindya Sarkar, Anirban Sarkar, Sowrya Gali, and Vineeth N Balasubramanian. Get fooled for the right reason: Improving adversarial robustness through a teacher-guided curriculum learning approach. *arXiv preprint arXiv:2111.00295*, 2021.

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.

Felix Stahlberg. Neural machine translation: A review. *Journal of Artificial Intelligence Research*, 69:343–418, 2020.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations (ICLR)*, 2014.

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

Frederick Tung and Greg Mori. Similarity-preserving knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1365–1374, 2019.

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

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.

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

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

Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. *arXiv preprint arXiv:1711.01991*, 2017.

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)*, pp. 807–816. IEEE, 2021.

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

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

Jinguo Zhu, Shixiang Tang, Dapeng Chen, Shijie Yu, Yakun Liu, Mingzhe Rong, Aijun Yang, and Xiaohua Wang. Complementary relation contrastive distillation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9260–9269, 2021.

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

Supplementary Materials for “FLEXIBLE RELATION PRESERVING FOR ADVERSARIAL TRAINING”

Here we will introduce more details about our method and experiments.

A More Details about Our Methods

Here we show more details about how we get the relationships between samples. Inspired by t-SNE and UMAP (Van der Maaten & Hinton, 2008; McInnes et al., 2018), we want to use the conditional probability distribution to represent relationships between samples. t-SNE and UMAP make use of the regular Kernel Density Estimation (KDE) for approximations of the conditional probabilities. It has too many hyper-parameters to tune, and the training cost is unacceptable for us. So followed by PKT (Passalis & Tefas, 2018), we use the cosine similarity-based affinity metric.

$$d_{ij} = 1 - \frac{f(x_i)^T f(x_j)}{||f(x_i)||_2||f(x_j)||_2}, \tag{1}$$

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

where $f(x_i)$ is the feature of the i-th sample, d_{ij} is the cosine distance between x_i and x_j, and K_{cos} is cosine similarity-based affinity metric value for x_i and x_j.

Moreover, we are also inspired by UMAP to better preserve the global structure of feature space by adding ρ_j, which is the distance between x_j and its nearest neighbor.

$$d'_{ij} = d_{ij} - \rho_j. \tag{3}$$

Finally, after normalization, we can get the probability that x_i is a neighbor of x_j by K_{cos} for the standard training model.

$$\begin{aligned} p_{i|j} &= \frac{2 - d_{ij}^{st'}}{\sum_{k=1,k\neq j}^{N}(2 - d_{jk}^{st'})}, \\ &= \frac{(2 - (d_{ij}^{st} - \rho_j))}{\sum_{k=1,k\neq j}^{N}(2 - (d_{jk}^{st} - \rho_j))}. \end{aligned} \tag{4}$$

Similarly, for the adversarial training model:

$$\begin{aligned} q_{i|j} &= \frac{2 - d_{ij}^{at'}}{\sum_{k=1,k\neq j}^{N}(2 - d_{jk}^{at'})}, \\ &= \frac{(2 - (d_{ij}^{at} - \rho_j))}{\sum_{k=1,k\neq j}^{N}(2 - (d_{jk}^{at} - \rho_j))}, \end{aligned} \tag{5}$$

To make it easier for the reader to understand our method, we show the pipeline of the method in Fig. 1

What’s more, when combined with Vanilla AT, $L_{robust} = CE(softmax(\widetilde{M}(\widetilde{X})))$;

when combined with LBGAT and Vanilla AT, $L_{robust} = CE(softmax(\widetilde{M}(\widetilde{X}))) + MSE(M(\widetilde{X}), \widetilde{M}(\widetilde{X}))$;

when combined with TRADES, $L_{robust} = CE(softmax(\widetilde{M}(X))) + \beta * KL(\widetilde{M}(X), \widetilde{M}(\widetilde{X}))$;

and when combined with LBGAT and TRADES, $L_{robust} = MSE(M(\widetilde{X}), \widetilde{M}(\widetilde{X})) + \beta * KL(\widetilde{M}(X), \widetilde{M}(\widetilde{X}))$

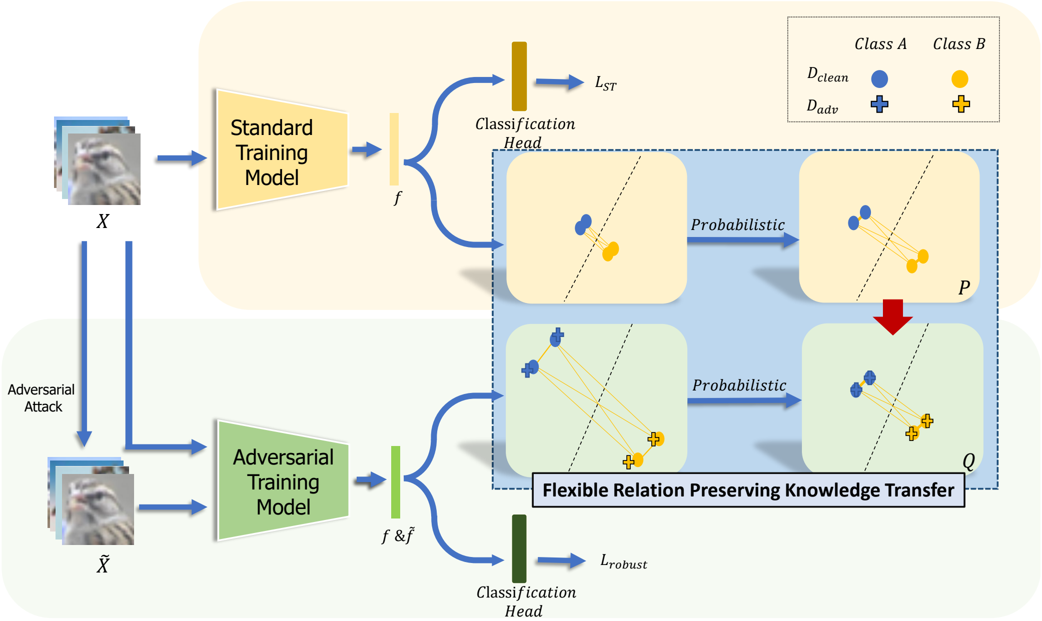


Figure 1: The whole framework of FPRAT.

Table 1: Quantitative experiment on Tiny ImageNet. "*" are the results directly quoted from LBGAT.

Defense	**Clean Acc.**	**PGD-20 Acc.**
Vanilla AT*	30.65	6.81
Vanilla AT + LBGAT*	36.50	14.00
ALP*	30.51	8.01
LBGAT + ALP*	33.67	14.55
TRADES (β = 6.0)*	38.51	13.48
TRADES (β = 6.0) + LBGAT*	39.26	16.42
TRADES (β = 6.0) + Ours	41.12	16.18
TRADES (β = 6.0) + LBGAT+ Ours	**41.53**	**17.09**

B More Results

Training details Here we will add more training details. Firstly, for different experiment settings, we choose different λ. We set $\lambda = 5$ on CIFAR-10 dataset, and $\lambda = 20\gamma$ on CIFAR-100 dataset, where $\gamma = \frac{2}{1+e^{\frac{-10t}{100}}-1}$ and t is the current t-th epoch during training. Moreover, all our experimental results are reproducible with random seed = 1. Finally, we also provided our core code in the supplementary material, and all the existing assets we used chose MIT license.

Experiments on Tiny ImageNet. To verify the effectiveness of our method on a larger dataset, we conduct new experiments on Tiny ImageNet (Deng et al., 2009), which contains $120,000$ 64×64 color images in classes. Experimental settings are the same as LBGAT. * are the results directly quoted from LBGAT. For the training set, we resize the image from 64×64 to 32×32, and the data augmentation is random crops with 4 pixels of padding; finally, we normalize pixel values to [0,1]. For test set, we resize the image to 32×32, and normalize pixel values to [0,1]. Others are the same as CIFAR-100 and CIFAR-10.

As shown in Table. 1, combined with our algorithm can further improve the clean sample generalization and robustness of TRADSES and LBGAT.

Table 2: The ablation experiment about different standard training model architecture.

Model	Method	Teacher Model	Clean Acc	Robust Acc		
				PGD-20 Acc	C&W-20 Acc	AA Acc
ResNet-18	Standard Training	None	77.39	0	0	0
WideResNet34-10	TRADES+Ours	ResNet-18	62.62	36.27	31.72	29.19
WideResNet34-10	Standard Training	None	78.11	0	0	0
WideResNet34-10	TRADES+Ours	WideResNet34-10	63.09	35.54	30.41	28.76

Table 3: Quantitative experiment on TRADES ($\beta = 0$) on CIFAT-10 and CIFAR-100.

Dataset	Transfer Methods	Clean Acc	Robust Acc		
			PGD-20 Acc	C&W-20 Acc	AA Acc
CIFAR-10	LBGAT	87.64	57.16	55.52	52.38
	LBGAT+Ours	**88.09**	**57.55**	**56.91**	**52.87**
CIFAR-100	LBGAT	**69.39**	33.05	29.74	26.59
	LBGAT+Ours	69.20	**33.22**	**30.68**	**27.59**

Ablations on Different teacher architectures. As shown in Table. 2, our approach is not sensitive to standard training model architectures, as ResNet18 has achieved comparable results to WideresNet34-10 on the CIFAR datasets.

Here we also show that our algorithm can further improve the performance of the LBGAT algorithm under the TRADES ($\beta = 0$) experimental setting (Table. 3).

Moreover, we also show our methods can alleviate the overfitting of adversarial training and get better performance both on the best epoch model and the last epoch model. (Table. 5 and Table. 5)

More discussion about LBGAT. As shown in Table. 6, we find that if only the adversarial training model is updated for the distillation loss of LBGAT ($LBGAT_{detach}$), the clean accuracy of the model will be further improved, but the robustness will be decreased, and it is difficult to converge in the training stage, which verifies our speculation.

C NEW RESULTS

New experimental settings To compared with LAS (Jia et al., 2022), a state-of-the-art adversarial training method, we also follow Jia et al. (2022) to rerun our experiments. ϵ is 8/255. The initial learning rate is set to 0.1 with a total of 110 epochs for training and reduced to 0.1x at the 100-th and 105-th epochs. Weight decay is 5e-4; other experimental Settings are the same as in the main text.

Quantitative experiment results As shown in Table. 7 and Table. 8, our algorithm has obvious advantages in clean accuracy and achieves better or comparable results in defense accuracy of different attack methods compared with most popular adversarial training algorithms. TRADES and LBGAT achieve significant improvement in clean sample generalization by combining with our algorithm, and the robust generalization is also relatively improved or maintained. Compared

Table 4: Quantitative experiment on CIFAR-10. GAP is the best epoch model's accuracy minus the last epoch model's accuracy. ↑ means the higher, the better, and ↓ means the lower, the better.

Model	Baseline	Transfer Methods	Last Epoch Clean Acc (↑)	Last Epoch PGD-20 Acc (↑)	PGD-20 GAP (↓)
WideResNet34-10	Vanilia AT	None	86.35	49.68	3.77
		Ours	**89.10**	**52.16**	**3.48**
		LBGAT	85.91	51.77	2.57
		LBGAT+Ours	**89.42**	**53.88**	**2.33**
	TRADES	None*	85.35	53.24	3.35
		Ours	**88.28**	**56.59**	**1.61**
		LBGAT*	82.31	**57.74**	0.30
		LBGAT+Ours	**87.62**	57.73	**0.00**

Table 5: Quantitative experiment on CIFAR-100. GAP is the best epoch model's accuracy minus the last epoch model's accuracy. ↑ means the higher, the better, and ↓ means the lower, the better.

Model	Baseline	Transfer Methods	Last Epoch Clean Acc (↑)	Last Epoch PGD-20 Acc (↑)	PGD-20 GAP (↓)
WideResNet34-10	Vanilia AT	None	59.83	26.03	**2.03**
		Ours	**66.45**	**27.15**	2.73
		LBGAT	59.37	27.43	2.67
		LBGAT+Ours	**68.39**	**28.03**	**1.9**
	TRADES	None	57.94	28.48	2.45
		Ours	**64.97**	**32.23**	**1.74**
		LBGAT	60.43	35.11	**0.97**
		LBGAT+Ours	**62.89**	**35.29**	0.98

Table 6: Analysis experiments on LBGAT on CIFAR-100.

Methods	Clean Acc	Robust Acc		
		PGD-20 Acc	C&W-20 Acc	AA Acc
Vanilla AT + $LBGAT_{detach}$	**71.30**	28.16	**28.65**	24.09
Vanilla AT + LBGAT	61.01	**30.10**	28.09	**25.63**

with SOTA method LAS-TRADES, We also have clear advantages in clean sample accuracy and robustness.

Table 7: New quantitative experiment on CIFAR10. "*" are the results directly quoted from Jia et al. (2022).

Defense	Clean Acc.	Robust Acc.		
		PGD-20 Acc.	C&W-20 Acc.	AA Acc.
Vanilla AT*	85.17	55.08	53.91	51.69
MART*	84.17	58.56	54.58	51.10
FAT*	**87.97**	49.86	48.65	47.48
GAIRAT*	86.30	**59.54**	45.57	40.30
AWP*	85.57	58.13	56.03	53.90
TRADES*	85.72	56.10	53.87	53.40
LAS-TRADES*	85.24	57.07	55.45	54.15
TRADES + Ours	87.07	58.51	**56.81**	**54.70**
TRADES + LBGAT	80.20	57.41	54.84	53.32
TRADES + LBGAT +Ours	86.69	58.04	56.75	54.47

Table 8: New quantitative experiment on CIFAR100. "*" are the results directly quoted from Jia et al. (2022).

Defense	**Clean Acc.**	**Robust Acc.**		
		PGD-20 Acc.	**C&W-20 Acc.**	**AA Acc.**
Vanilla AT*	60.89	31.69	30.10	27.86
SAT*	62.82	27.17	27.32	24.57
AWP*	60.38	33.86	31.12	28.86
TRADES*	58.61	28.66	27.05	25.94
LAS-TRADES*	60.62	32.53	29.51	28.12
TRADES + Ours	**67.47**	34.99	31.61	28.95
TRADES + LBGAT*	60.64	34.75	30.65	29.33
TRADES + LBGAT+ Ours	65.40	**35.46**	**32.36**	**30.17**