Improving Model Robustness with Latent Distribution Locally and Globally

Zhuang QIAN, Shufei ZHANG, Kaizhu HUANG, Qiufeng WANG, Rui ZHANG, Xinping YI

VII. Supplementary Experiments

TABLE IV
The Comparison between Methods

Method	Local	Partial Global	Global	Extra Data	Performance
AT [1]	✓	✗	✗	✗	★☆
TRADES [2]	✓	✓	✗	✗	★☆
FS [3]	✓	✓	✗	✗	★★
OVERFITTING [4]	✓	✗	✗	✗	★★
PRETRAINING [5]	✓	✗	✗	✓	★★
WAR [6]	✓	✗	✗	✓	★★☆
Combination [7]	✓	✗	✗	✓	★★★
FAT [8]	✓	✗	✗	✗	★★
HAT [9]	✓	✓	✗	✓	★★☆
DAJAT [10]	✓	✓	✗	✗	★★
LBGAT [11]	✓	✓	✓	✗	★★
ATLD	✓	✓	✓	✗	★★★

We conduct experiments on the widely-used datasets, CIFAR-10 [12], SVHN [13], CIFAR-100 [12] and Tiny ImageNet [14]. Following the Feature Scattering [3], we leverage the WideResNet [15] as our basic classifier and discriminator model structure. The initial learning rate is empirically set to 0.01 for Tiny ImageNet and 0.1 for others. We train our model 400 epochs on Pytorch and RTX2080TI with transition epoch $60, 90$ and decay rate 0.1. The input perturbation budget is set to $\epsilon = 8$ with the label smoothing rate as 0.5. We use L_∞ perturbation in this paper including all the training and evaluation.

We evaluate the various models on white-box and black-box attacks and report robust accuracy of our proposed models: ATLD (with IMT) and ATLD+ (with limited IMT). Under the white-box attacks, we compare the accuracy of the proposed method with several competitive methods, including 1) the original Wide-Resnet (Standard) trained with natural examples; 2) Traditional Adversarial Training with PGD (AT) [1]; 3) Triplet Loss Adversarial training (TLA) [16]; 4) Layer-wise Adversarial Training (LAT) [17]: injecting adversarial perturbation into the latent space; 5) Bilateral: adversarial perturb on examples and labels both [18]; 6) Feature-Scattering (FS): generating adversarial examples with considering inter-relationship of samples [3]. 7) TRadeoff-inspired Adversarial DEfense via Surrogate-loss minimization (TRADES) [2]: trading adversarial robustness off against accuracy; 8) Robust-overfitting [4]: improving models adversarial robustness by simply using early stop; 9) Pretraining [5]: improving models adversarial robustness with pre-training; 10) Width Adjusted Regularization (WAR) [6]: mitigating the perturbation stability deterioration on wider models; 11) Robust Self-Training (RST) [19]: achieving high robust accuracy with semisupervised learning procedure (self-training); 12) Combination [7]: achieving state-of-the-art results by combining larger models, Swish/SiLU activations and model weight averaging; 13) Friendly Adversarial Training (FAT) [8]: learning the network with the least loss adversarial data among those that are confidently misclassified; 14) Helper-based Adversarial Training (HAT) [9]: training the model with helper examples which are twice far along the perturbation direction but labelled by one standard network. 15) Diverse Augmentation-based Joint Adversarial Training (DAJAT) [10]: training the model with the combination of simple and complex augmentations. 16) Learnable Boundary Guided Adversarial Training (LBGAT) [11]: leveraging the logits from one clean model to guide the learning of the robust model.

These algorithms present state-of-the-art performance in defending against adversarial attacks. To demonstrate more clearly the advantages and disadvantages of our method in comparison with the other methods, we first provide a comparison between the different methods in Table IV where whether local or global information is used, whether additional data is used, and their performance is explicitly marked. Here, more stars mean better performance; a hollow star means half a star.

Under the black-box attacks, we compare four different algorithms used to generate the test time attacks: Vanilla training with natural examples, adversarial training with PGD, FS, and our proposed model.

A. Ablation Study

To more effectively elucidate the efficacy and characteristics of ATLD and IMT, we conduct ablation experiments for defending against white-box attacks with $\epsilon = 8$ on CIFAR-10, CIFAR-100, and SVHN as illustrated in Figure 8. As observed, ATLD− (ATLD model without applying IMT) outperforms FS on natural data, as well as FGSM-attacked data and CW20-attacked data. Notably, due to the incorporation of

Zhuang QIAN and Shufei ZHANG contributed equally to this work.

Kaizhu HUANG (Corresponding Author) is with Data Science Research Center, Duke Kunshan University. E-mail:{Kaizhu.Huang}@dukekunshan.edu.cn

Zhuang QIAN and Xinping YI are with Department of Electrical Engineering and Electronics, University of Liverpool. E-mail:{Zhuang.Qian, Xinping.Yi}@liverpool.ac.uk

Shufei ZHANG is with Shanghai Artificial Intelligence Laboratory. E-mail:{zhangshufei}@@pjlab.org.cn

Qiufeng WANG is with School of Advanced Technology, Xi'an Jiaotong-Liverpool University. E-mail:qiufeng.wang@xjtlu.edu.cn

Rui ZHANG is with the Department of Foundational Mathematics, Xi'an Jiaotong-Liverpool University. E-mail:rui.zhang02@xjtlu.edu.cn

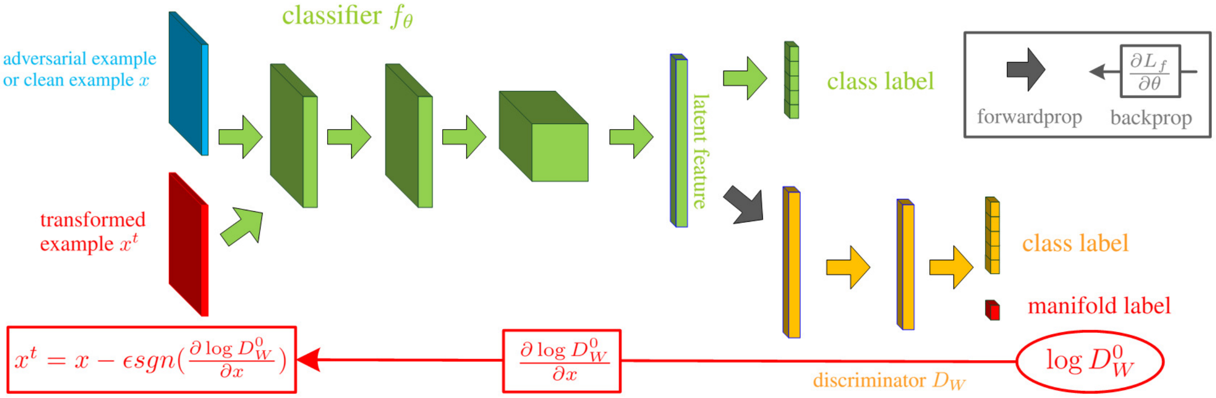
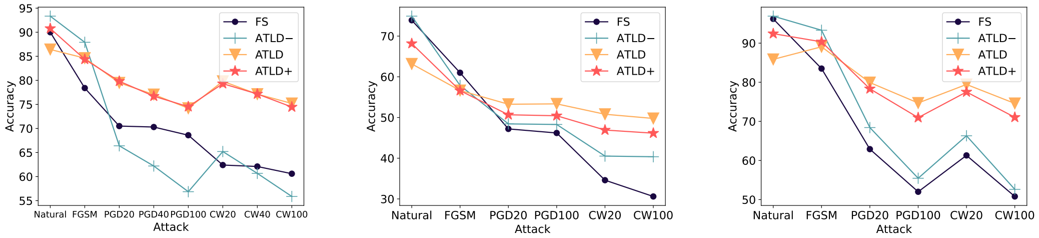

classifier f_θ
adversarial example or clean example x
transformed example x^t
latent feature
class label
forwardprop
backprop
$\frac{\partial L_f}{\partial \theta}$
class label
manifold label
$x^t = x - \epsilon sgn(\frac{\partial \log D_W^0}{\partial x})$
$\frac{\partial \log D_W^0}{\partial x}$
discriminator D_W
$\log D_W^0$

Fig. 7. Detailed Procedure of IMT. 1) The inference sample x (adversarial or not adversarial) is fed into the network, and the discriminator outputs its prediction. The loss $\log D_W$ is computed, and the transformed example x^t (red arrow) is then generated. 2) The transformed sample is fed into the network and classified by the adversarially-trained network.



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

Fig. 8. Ablation study On CIFAR-10, CIFAR-100 and SVHN.

manifold information, our proposed ATLD− achieves superior performance on clean data across all three datasets. However, IMT appears to have a negative impact on natural and FGSM data, while significantly enhancing performance against PGD and CW. In order to mitigate the impact of IMT on the natural and FGSM data, a threshold is employed to constrain the IMT perturbation based on the output of the discriminator. The perturbation is halved if the output of the discriminator lies within the range of $[0.3, 0.7]$ (ATLD+). In this configuration, our approach could achieve high performance against adversarial attacks without compromising the accuracy of clean data excessively.

Furthermore, we carry out an experiment to evaluate the effectiveness of classification within the discriminator, as presented in Table V. "CLA OF D IN TRA" refers to training the discriminator with its classification loss. When this term is positive, the discriminator is trained to discern whether a sample belongs to an adversarial example or natural sample while providing the correct predict class; "CLA OF D IN PERT" refers to generating adversarial perturbation with the classification loss of the discriminator. The perturbation is generated to make itself more distinguishable by the discriminator while causing the discriminator to give an incorrect prediction. According to Table V, incorporating classification loss in the discriminator can improve the robustness of ATLD (from 78.93% to 80.1% for PGD, and from 78.93% to 80.57% for CW), while having a substantial impact on ATLD+. Utilizing the classification loss of the discriminator for both training and perturbation generation leads to further improvements for both ATLD and ATLD+ for PGD and CW while leading to a slight decrease for natural and FGSM data.

In this paper, we take the ATLD and ATLD+ as our main contributions. We will mainly focus on these two methods in the later discussion.

TABLE V
ABLATION STUDY ON CLASSIFICATION LOSS IN THE DISCRIMINATOR

	CLA OF D IN TRA	CLA OF D IN PERT	NATURAL	FGSM	PGD20	CW20
ATLD	✗	✗	87.13±1.08	83.15±2.15	78.93±3.02	78.93±2.91
ATLD+			90.21±1.66	83.09±2.25	76.59±2.63	76.55±2.83
ATLD	✓	✗	88.33±1.52	84.75±0.55	80.11±2.07	80.57±1.37
ATLD+			91.16±0.78	84.73±0.64	76.80±2.45	77.16±2.36
ATLD	✓	✓	87.58±1.71	84.44±1.20	80.96±1.33	81.16±1.16
ATLD+			88.46±2.06	84.62±1.28	78.87±1.32	78.80±1.29

B. Defending Black-box Attacks

To further verify the robustness of ATLD, we conduct transfer-based black-box attack experiments on CIFAR-10, CIFAR-100 and SVHN. Two different agent models (Resnet-18) are used for generating test time attacks including the Vanilla Training model, and the Adversarial Training with PGD model. As demonstrated by the results in Table VI, our proposed approach can achieve competitive performance almost in all cases. Specifically, ATLD+ outperforms the AT and TRADES in 15 out of 18 cases while it demonstrates comparable or slightly worse accuracy in the other 3 cases. The performance of our two methods shows marginally inferior

TABLE VI
CLASSIFICATION ACCURACY UNDER TRANSFER-BASED BLACK-BOX ATTACKS ON CIFAR-10

DEFENSE MODELS	ATTACKED MODELS (CIFAR-10)				ATTACKED MODELS (CIFAR-100)				ATTACKED MODELS (SVHN)			
	VANILLA TRAINING		ADVERSARIAL TRAINING		VANILLA TRAINING		ADVERSARIAL TRAINING		VANILLA TRAINING		ADVERSARIAL TRAINING	
	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20	FGSM	PGD|CW20
AT [1]	83.59	84.40|84.47	77.63	74.28|73.30	**59.57**	60.30|60.24	56.62	55.42|56.58	88.31	89.54|89.60	77.65	73.43|74.34
TRADES [2]	**84.67**	85.35|85.28	77.71	74.11|73.99	59.29	59.52|59.93	54.90	53.70|55.20	**90.47**	91.90|91.89	83.29	78.10|78.79
ATLD	82.85	85.35|85.67	82.81	76.34|77.36	56.51	57.14|57.32	57.46	55.14|56.97	89.40	91.74|92.15	91.05	87.25|88.39
ATLD+	84.21	**87.32|87.12**	**84.63**	**80.11|80.70**	56.31	**61.14|62.08**	**58.15**	**60.62|61.56**	90.40	**92.71|92.97**	**92.24**	**90.03|90.60**

to AT or TRADES against FGSM data, while our method outperforms AT and TRADES significantly against PGD20 and CW20 adversarial attacks both on the Vanilla Training model, and the Adversarial Training with PGD model.

In subsequent sections, we will delve into two sophisticated attacks: AutoAttack (AA) [20] and Rays [21]. Within the ensemble attacks of AA, there is an integration of a black-box attack, specifically the Square Attack [22]. Additionally, Rays operates as a black-box attack, relying only on the target model's hard-label output (namely the prediction label). By examining these two additional black-box attacks, we aim to provide a comprehensive assessment of ATLD's robustness against black-box attacks.

C. Time Consuming Analysis

To evaluate the computational and time burden introduced by the additional discriminator, we conduct the experiment using NVIDIA RTX-2080TI GPU card on CIFAR10 about the training and inference time consuming of the proposed ATLD compared with AT and FS as shown in Table VII.

In particular, our proposed ATLD requires less expensive computations in each epoch, unlike other methods such as FS, which requires costly computations on optimal transport, and AT, which demands multiple backpropagations. Furthermore, our proposed ATLD training from scratch both for the classifier and the discriminator does not incur extra computations. Each epoch in ATLD takes around 21m37s, which is less time-consuming than FS (25m14s) and standard AT with PGD-7 (28m53s). Although it looks like ATLD requires more epochs to achieve optimal robustness, it outperformed all compared methods at the 200th epoch, attaining a 77.74% robust accuracy against PGD20 and 79.03% against CW20, with performance only slightly inferior to that at the 400th epoch. Note the training epoch reported in Table VII is the best-performance epoch, which is based on empirical experience and previous study [4].

Furthermore, compared to AT and FS, our proposed IMT does not require additional inference time consumption. As observed in Table VII, our proposed method consumes almost the same inference time as AT and FS (i.e., all take around 25 seconds), as the computational cost incurred in solving the optimization problem appears negligible compared to data reading and transmission between the CPU and GPU. In our CIFAR-10 experiments, the inference time difference for the entire test dataset between ATLD (25s) and ATLD− (without IMT) (25s) was negligible for natural data.

TABLE VII
TRAINING AND INFERENCE TIME CONSUMING ON CIFAR-10

	TRAINING TIME PER EPOCH (SEC)	EPOCHS	TOTAL TIME (HOUR)	INFERENCE TIME (SEC)
AT [1]	1733	62	29.85	25
FS [3]	1514	200	84.11	25
ATLD	1297	400	144.11	25

D. Analysis of the Validity of LMAEs

We conducted the ablation experiment to examine the potency of the LMAEs and IMT. Particularly, taking the dataset of CIFAR10 as one illustrative example, we executed experiments where the model was trained using the ATLD mechanism but replaced LMAEs with alternative attack techniques, including PGD, TRADES, and FS. As presented in Table VIII, when substituting LMAEs with the other prevalent training attack methodologies, both natural and robust accuracies are significantly lower than those of the ATLD with LMAEs. This substantiates the effectiveness of LMAEs in our proposed ATLD framework.

Moreover, the efficacy of the proposed IMT has been established in different frameworks, consistently improving robustness. However, ATLD+ seems less effective at alleviating the performance degradation caused by IMT on natural samples. This discrepancy might stem from the different distribution of training attacks, which consequently affects the Discriminator's output distribution. As a result, the efficacy range of [0.3-0.7] of ATLD+ might not hold for other types of attacks. This aspect will be a subject of our future investigations.

Nevertheless, it can be noted that the proposed ATLD− may still be vulnerable to sophisticated attacks. As anticipated, the proposed LMAEs, which consider global manifold information, are not designed to deceive the classifier during training entirely. It is plausible that approaches incorporating global and manifold defenses may not withstand special local perturbations identifiable by AA and Rays because the model has not seen the worst-case scenario. Similar occurrences have been observed in FS, which exhibits high robustness against PGD and CW but is unsuccessful against AA.

Conversely, several recent studies have questioned the appropriateness of AA as a benchmark. As evidenced by [23], [24] that even simple detection algorithms exhibit near-perfect detection rates for AA samples. In contrast, other attack methods present a more challenge to detect and do not achieve comparable detection rates as effortlessly. These investigations contend that AA is an excessively severe attack. Consequently, when evaluating model robustness, it is essential to consider performance against several attacks rather than solely focusing on strong attacks.

Our experiments are consistent with this phenomenon since our proposed IMT improved the robust accuracy against AA from 34.49% to 70.49%, attributable to the discriminator's effective prediction of the manifold for AA-attacked samples. As depicted in Figure 4, IMT can predict whether a sample belongs to the natural or adversarial manifold. Accordingly, IMT utilizes the discriminator gradient to push adversarial samples into the natural manifold, rendering them more distinct and easier for classification. As an integral component of ATLD, IMT excels in predicting strong attacks. Nonetheless, for weak attacks or natural samples, IMT's predictions may not be entirely accurate.

TABLE VIII
Robust Accuracy against White-box Attacks with Different Training Attacks on CIFAR 10

Training Attack	Inference Method	Natural	FGSM	PGD20	CW20	AA
PGD	ATLD-	84.56	62.61	49.65	47.09	40.69
	ATLD	76.08	62.71	59.17	58.71	60.22
	ATLD+	80.72	60.80	52.65	52.07	56.02
FS	ATLD-	89.90	77.18	62.55	53.68	35.95
	ATLD	80.82	74.91	65.72	62.78	54.91
	ATLD+	80.85	74.98	65.66	62.37	54.90
TRADES	ATLD-	87.54	71.89	57.08	50.64	41.19
	ATLD	76.28	70.02	64.43	62.23	64.91
	ATLD+	77.13	69.49	64.46	62.40	64.10
LMAEs	ATLD-	93.34	87.91	66.40	65.21	34.49
	ATLD	86.42	84.62	79.48	79.81	70.49
	ATLD+	90.78	84.34	79.82	79.31	70.60

E. Robustness Evaluation under Adaptive Attacks

The proposed IMT utilizes information from the discriminator during the inference time, so what if the discriminator is white-box attacked during inference? To verify the effectiveness of the proposed IMT. we further perform additional experiments to evaluate adaptive attacks focusing on both the classifier and the discriminator as shown in Table IX. We find that when the adversary attacks both the classifier and discriminator adaptively, the proposed ATLD still has high robustness. Here 'D' means the binary classification of the discriminator. 'Cla' means the classification loss of the classifier such as the cross entropy loss and CW loss. '(Cla+D)-step20' means that both the loss of and classifier and the discriminator are used to compute the attacking gradients with 20 steps, 'Cla-step20 + D-step1' means that the classification loss is first used to compute attacking gradients with 20 steps then followed by 1 step compute attacking gradients with the discriminating loss. Note that FGSM only takes 1 step.

F. Model Robustness on Higher Resolution Images

We further evaluate the effectiveness of the proposed ATLD on the higher resolution dataset Tiny-ImageNet [14]. The Tiny-ImageNet database is a subset collected from the ImageNet database, which covers 200 classes with 600 images in the training set and 50 images in the validation set in the size of 64 × 64 for each class. As there are no labels for test images of Tiny ImageNet, following the common practice, we evaluate the different methods on the validation set.

TABLE IX
Robustness Accuracy under Different Adaptive Attacks

	Attacked Models (CIFAR-10)			
Methods	(Cla+D)-step20 FGSM/PGD/CW	Cla-step20 + D-step1 FGSM/PGD/CW	D-step1	D-step20
ATLD	84.66/80.4/79.38	86.07/81.16/81.13	87.51	85.72
ATLD+	85.4/80.31/79.49	85.72/83.54/83.32	87.68	87.75
	Attacked Models (CIFAR-100)			
Methods	(Cla+D)-step20 FGSM/PGD/CW	Cla-step20 + D-step1 FGSM/PGD/CW	D-step1	D-step20
ATLD	56.91/53.95/50.37	57.81/58.59/57.57	59.62	58.84
ATLD+	57.01/51.60/47.44	58.08/60.32/58.62	59.28	61.50
	Attacked Models (SVHN)			
Methods	(Cla+D)-step20 FGSM/PGD/CW	Cla-step20 + D-step1 FGSM/PGD/CW	D-step1	D-step20
ATLD	89.16/79.64/79.18	87.13/83.63/83.19	89.75	82.85
ATLD+	89.03/78.86/78.14	87.31/85.18/85.14	90.24	85.14

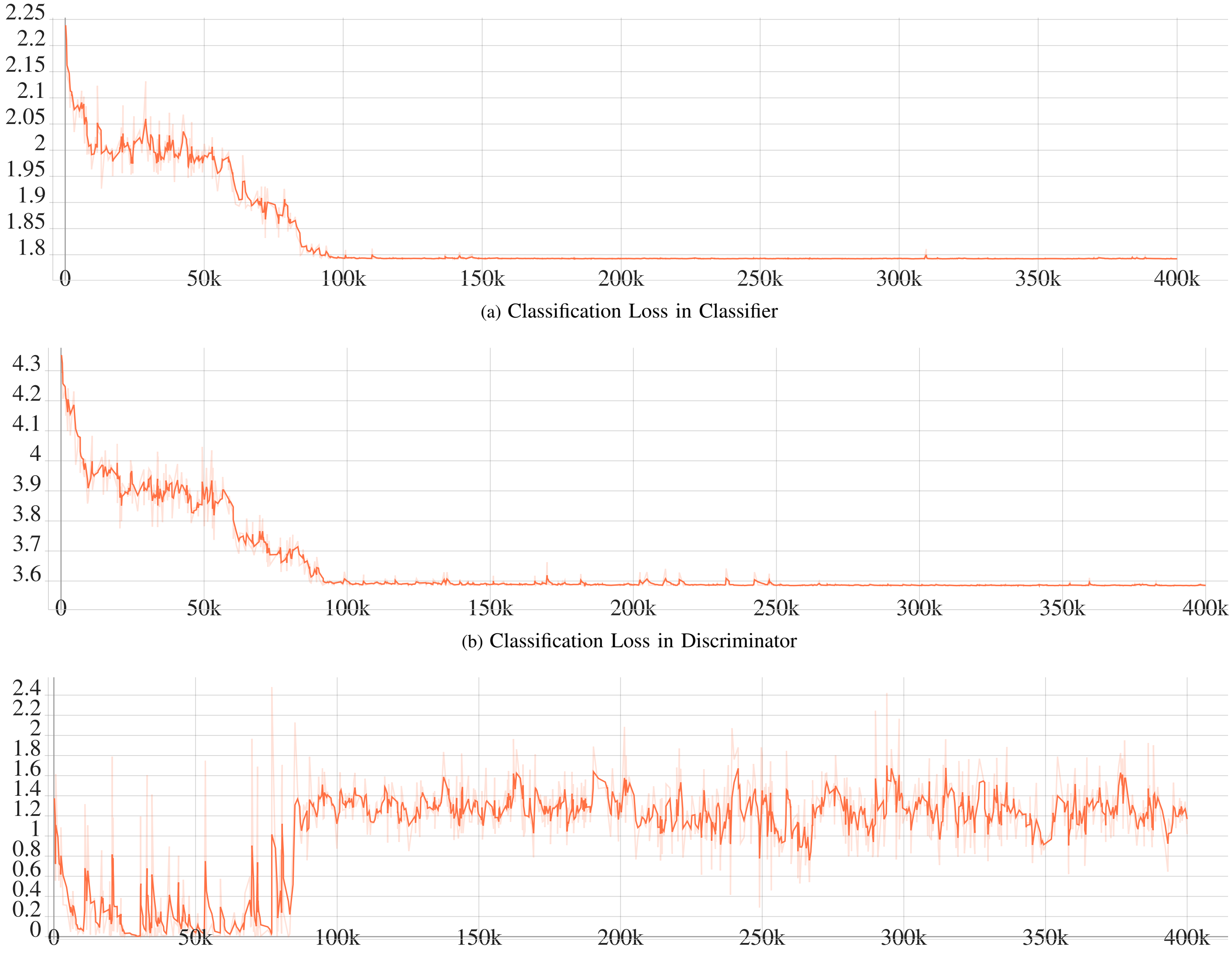
Table X demonstrates clear evidence that the ATLD techniques present a balanced performance profile. In natural accuracy, both ATLD and ATLD+ have a distinct advantage. Additionally, ATLD displays strong robustness against the AA attack, with an accuracy of 32.96%. While ATLD+ has a slight drop in defending to some adversarial attacks compared to ATLD, it still holds its ground, especially against AA. Overall, the ATLD methods provide a compelling alternative to traditional methods, suggesting their potential utility in enhancing adversarial defence strategies.

TABLE X
Robust Accuracy against White-box Attacks on Tiny-ImageNet

Method	Natural	PGD50	CW50	AA
AT [1]	43.98	19.98	17.6	13.78
TRADES [2]	39.16	15.74	12.92	12.32
AWP [25]	41.48	22.51	19.02	17.34
LAS-AWP [26]	45.26	23.42	19.88	18.42
ATLD	51.60	33.10	33.14	32.96
ATLD+	55.22	29.38	29.42	28.90

G. Analysis of Loss Dynamics in ATLD Training

We also present the training loss plots for the proposed ATLD to provide insight into the learning efficiency and stability of the model, as shown in Fig. 9. The x-axis of the plot delineates the training iterations, while the y-axis quantifies the value of the loss function. For manifold loss (D_W^0), initially, the loss exhibits high variability with sharp peaks, indicating the discriminator's struggles to adapt to the manifolds. This high variability in the early stages is typical as the discriminator has not yet learned enough to effectively differentiate features from natural or adversarial examples. As training progresses, the loss undergoes notable fluctuations, but shows a decreasing trend. This suggests an adjustment phase where the discriminator begins to better identify the distinguishing patterns, improving its accuracy. During this period, the classifier also enhances its ability to correctly output labels, which requires continual adaptation by the



(a) Classification Loss in Classifier

(b) Classification Loss in Discriminator

(c) Manifold Loss in Discriminator

Fig. 9. Loss Dynamics in Training Procedure of ATLD, smoothing rate 0.6. (Notes the oscillatory loss in manifold loss is not necessarily problematic but is indicative of the complex and dynamic interactions characteristic of adversarial training processes, which is similar to GANs [27], [28].)

discriminator. Eventually, the plot shows a trend towards stabilization and lower-amplitude fluctuations in the loss, signaling that the discriminator is achieving a state of equilibrium with the classifier. This phase indicates that the discriminator has learned the manifolds of the natural and adversarial samples, resulting in a more consistent performance in assessing the authenticity of the samples.

References

[1] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," in *International Conference on Learning Representations (ICLR)*. OpenReview.net, 2018.

[2] H. Zhang, Y. Yu, J. Jiao, E. P. Xing, L. E. Ghaoui, and M. I. Jordan, "Theoretically principled trade-off between robustness and accuracy," in *International Conference on Machine Learning (ICML)*, vol. 97. PMLR, 2019, pp. 7472–7482.

[3] H. Zhang and J. Wang, "Defense against adversarial attacks using feature scattering-based adversarial training," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 1829–1839.

[4] L. Rice, E. Wong, and J. Z. Kolter, "Overfitting in adversarially robust deep learning," in *International Conference on Machine Learning (ICML)*, vol. 119. PMLR, 2020, pp. 8093–8104.

[5] D. Hendrycks, K. Lee, and M. Mazeika, "Using pre-training can improve model robustness and uncertainty," in *International Conference on Machine Learning (ICML)*, vol. 97. PMLR, 2019, pp. 2712–2721.

[6] B. Wu, J. Chen, D. Cai, X. He, and Q. Gu, "Do wider neural networks really help adversarial robustness?" in *Neural Information Processing Systems (NeurIPS)*, 2021, pp. 7054–7067.

[7] S. Gowal, C. Qin, J. Uesato, T. A. Mann, and P. Kohli, "Uncovering the limits of adversarial training against norm-bounded adversarial examples," *CoRR*, vol. abs/2010.03593, 2020. [Online]. Available: https://arxiv.org/abs/2010.03593

[8] J. Zhang, X. Xu, B. Han, G. Niu, L. Cui, M. Sugiyama, and M. S. Kankanhalli, "Attacks which do not kill training make adversarial learning stronger," in *International Conference on Machine Learning (ICML)*, vol. 119. PMLR, 2020, pp. 11 278–11 287.

[9] R. Rade and S.-M. Moosavi-Dezfooli, "Helper-based adversarial training: Reducing excessive margin to achieve a better accuracy vs. robustness trade-off," in *ICML 2021 Workshop on Adversarial Machine Learning*, 2021.

[10] S. Addepalli, S. Jain *et al.*, "Efficient and effective augmentation strategy for adversarial training," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 35, pp. 1488–1501, 2022.

[11] J. Cui, S. Liu, L. Wang, and J. Jia, "Learnable boundary guided adversarial training," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2021, pp. 15 701–15 710.

[12] A. Krizhevsky, V. Nair, and G. Hinton, "The cifar-10 dataset," in *online:*

https://www.cs.toronto.edu/˜kriz/cifar.html, 2014.

[13] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, and A. Y. Ng, "Reading digits in natural images with unsupervised feature learning," 2011.

[14] Y. Le and X. Yang, "Tiny imagenet visual recognition challenge," *CS 231N*, vol. 7, no. 7, p. 3, 2015.

[15] S. Zagoruyko and N. Komodakis, "Wide residual networks," in *arXiv preprint arXiv:1605.07146*, 2016.

[16] C. Mao, Z. Zhong, J. Yang, C. Vondrick, and B. Ray, "Metric learning for adversarial robustness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 478–489.

[17] N. Kumari, M. Singh, A. Sinha, H. Machiraju, B. Krishnamurthy, and V. N. Balasubramanian, "Harnessing the vulnerability of latent layers in adversarially trained models," in *International Joint Conference on Artificial Intelligence (IJCAI)*. ijcai.org, 2019, pp. 2779–2785.

[18] J. Wang and H. Zhang, "Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks," in *IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE, 2019, pp. 6628–6637.

[19] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. Liang, "Unlabeled data improves adversarial robustness," in *Advances in Neural Information Processing Systems (NeurIPS)*, 2019, pp. 11 190–11 201.

[20] F. Croce and M. Hein, "Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks," in *International Conference on Machine Learning (ICML)*, ser. Proceedings of Machine Learning Research, vol. 119. PMLR, 2020, pp. 2206–2216.

[21] J. Chen and Q. Gu, "Rays: A ray searching method for hard-label adversarial attack," in *ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2020, pp. 1739–1747.

[22] M. Andriushchenko, F. Croce, N. Flammarion, and M. Hein, "Square attack: A query-efficient black-box adversarial attack via random search," in *Computer Vision - ECCV European Conference*, A. Vedaldi, H. Bischof, T. Brox, and J. Frahm, Eds.

[23] P. Lorenz, P. Harder, D. Strassel, M. Keuper, and J. Keuper, "Detecting autoattack perturbations in the frequency domain," in *[ICML 2021 workshop on A Blessing in Disguise]*, 2021, pp. 1–7.

[24] P. Lorenz, D. Strassel, M. Keuper, and J. Keuper, "Is robustbench/autoattack a suitable benchmark for adversarial robustness?" in *The AAAI-22 Workshop on Adversarial Machine Learning and Beyond*, 2022.

[25] D. Wu, S. Xia, and Y. Wang, "Adversarial weight perturbation helps robust generalization," in *Advances in Neural Information Processing Systems: Annual Conference on Neural Information Processing Systems (NeurIPS)*, H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, Eds., 2020.

[26] X. Jia, Y. Zhang, B. Wu, K. Ma, J. Wang, and X. Cao, "LAS-AT: adversarial training with learnable attack strategy," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2022, pp. 13 388–13 398.

[27] P. Grnarova, K. Y. Levy, A. Lucchi, N. Perraudin, I. Goodfellow, T. Hofmann, and A. Krause, "A domain agnostic measure for monitoring and evaluating gans," *Advances in neural information processing systems (NeurIPS)*, vol. 32, 2019.

[28] J. Brownlee, *How to Identify and Diagnose GAN Failure Modes*. Machine Learning Mastery, 2021.