Supplementary Material for Topology-preserving Adversarial Training for Alleviating Natural Accuracy Degradation

BMVC 2024 Submission # 168

A Pipeline

To make it easier for the reader to understand our method, Fig. A1 shows the overall process for TRAIN combined with vanilla AT.

Discussion. LBGAT also adopts a two-model framework and transfers the prior knowledge of M to M'. Nonetheless, there exist notable distinctions between our proposed method and LBGAT, which can be summarized as follows:

1. Different perspectives. LBGAT mitigates natural accuracy degradation by focusing on the guidance of the natural classifier boundary. Different from it, our proposed TRAIN emphasizes the importance of the topology of the sample in the representation space. By combining the two perspectives, we can further enhance the model's performance, as confirmed by the experimental results.

2. Different interactions between models. In LBGAT, M and M' affect each other which still has the negative impact of the adversarial samples on the natural samples. However, when M remains independent, optimizing LBGAT becomes challenging due to the inherent differences between the two models. In TRAIN, M unidirectionally influences M' and as an anchor to preserve the original topology of natural samples in the representation space to avoid the negative influence of the adversarial samples on the natural samples. This design choice effectively mitigates the adverse effects of adversarial samples on natural samples.

These differentiating factors highlight the unique contributions of our proposed TRAIN method in addressing the natural accuracy degradation during adversarial training. By considering the topological aspects of samples to avoid the negative impact of adversarial samples, TRAIN offers a novel and effective approach for enhancing model robustness and performance.

B The Flexibility of TRAIN

Different from other methods, TRAIN mitigates natural accuracy degradation by adopting a novel topological perspective. Moreover, TRAIN could be applied to other adversarial

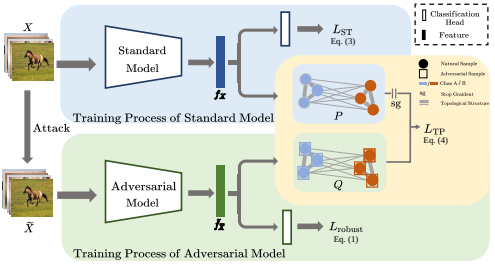

Figure A1: Overall framework of TRAIN. Specifically, we train a standard model M and an adversarial model M'. The standard model takes natural samples X as input and is optimized by cross-entropy loss. On the other hand, the adversarial model takes adversarial samples X' as input and is optimized by robust loss $L_{robust}(\cdot)$ and topology preservation loss $L_{\mathrm{TP}}(\cdot)$. $L_{\mathrm{TP}}(\cdot)$ constructs and aligns the neighborhood relation graph P and Q in the representation spaces of M and M', respectively. It can preserve the topological relationships among samples to reduce the negative effects of the adversarial samples during adversarial training.

training methods, such as vanilla AT [10], TRADES [17], and LBGAT [4], in a plug-and-play way. To validate the effectiveness of our proposed enhancements, we conduct comprehensive validation experiments on these strong baselines. We have introduced the robust loss of vanilla AT in our paper. TRADES [17] improves classification performance by introducing a regularization term, which penalizes the discrepancy between the logits for adversarial examples and their corresponding natural images. Its optimization objective is defined as:

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

where β is a hyper-parameter and its value means the strength of regularization for robustness. TRADES has proven highly effective and remains a strong baseline for adversarial training to this day. Here we will elucidate the specifics of our approach when integrated with another strong baseline LBGAT.

LBGAT leverages the model logits obtained from a standard model to guide the learning process of an adversarial model. It is usually combined with vanilla AT and TRADES. The total loss L_{AT} of adversarial model M' combined with LBGAT and vanilla AT is:

$$\begin{aligned} L_{\mathrm{AT}} &= L(z'_{x'_i}, y_i)+\gamma||logit'_{x'_i}-logit_{x_i}||_2 \\ &+\lambda\sum_i\sum_j\left[p_{i|j}\log\left(\frac{p_{i|j}}{q_{i|j}}\right)+(1-p_{i|j})\log\left(\frac{1-p_{i|j}}{1-q_{i|j}}\right)\right], \end{aligned} \tag{2}$$

and the total loss of adversarial model L_{AT} combined with LBGAT and TRADES is:

$$\begin{aligned} L_{\mathrm{AT}} &= L(z'_{x_i}, y_i)+\beta KL(z'_{x_i}||z'_{x'_i})+\gamma||logit'_{x'_i}-logit_{x_i}||_2 \\ &+\lambda\sum_i\sum_j\left[p_{i|j}\log\left(\frac{p_{i|j}}{q_{i|j}}\right)+(1-p_{i|j})\log\left(\frac{1-p_{i|j}}{1-q_{i|j}}\right)\right], \\ z'_{x_i} &= \frac{exp(logit'_{x_i})}{\sum_{j=1}^{N}exp(logit'_{x_j})}, z'_{x'_i}=\frac{exp(logit'_{x'_i})}{\sum_{j=1}^{N}exp(logit'_{x'_j})}, \end{aligned} \tag{3}$$

where KL is Kullback–Leibler divergence, which is commonly used to concretely implement the robust regularization of TRADES. γ is the hyper-parameter of LBGAT, and β is the hyper-parameter of TRADES.

C Experimental Settings

Datasets. Following [4, 6, 17], we conduct extensive evaluations on popular datasets, including CIFAR-10, CIFAR-100 [8] and Tiny ImageNet [5] dataset to validate the effectiveness of our algorithm. The CIFAR-10 and CIFAR-100 datasets consist of a total of $60,000$ color images with dimensions of 32×32 pixels. Among these $50,000$ images are designated for training and the remaining $10,000$ images are reserved for testing. Furthermore, CIFAR-10 has 10 categories while CIFAR-100 has 100 categories. Furthermore, the Tiny ImageNet dataset includes $120,000$ color images with dimensions of 64×64 pixels in 200 categories, with each category containing 500 training images, 50 validation images, and 50 testing images. This dataset offers a larger image size and a broader range of categories, enabling a more challenging evaluation of our algorithm's performance.

Baselines. We choose three strong baselines to demonstrate the effectiveness of our method: Vanilia AT [10], TRADES [17], and LBGAT. For TRADES, we set $\beta = 6.0$. For LBGAT, we conduct experiments based on vanilla AT and TRADES ($\beta = 6.0$). We also add ALP [7] as a baseline in the Tiny ImageNet dataset. In addition, we combine TRAIN with them to demonstrate the superiority of our approach. To provide a comprehensive evaluation and comparison with other state-of-the-art adversarial training methods, we include additional baseline: MART [15], FAT [18], GAIRAT [19], AWP [16], SAT [13], LAS [6], and ECAS [9].

Evaluation metrics. To evaluate the generalization of the model on natural and adversarial samples, our evaluation metrics are natural data accuracy (Natural Acc.) and robust accuracy (Robust Acc.). Robust accuracy is the model classification accuracy under adversarial attacks. As specified in the respective publications, we choose three representative adversarial attack methods for evaluation: PGD-20, C&W-20 [2], and Auto Attack [3]. We denote the model's defense success rate under those attacks separately as *PGD-20 Acc.*, *C&W-20 Acc.*, and *AA Acc.*. Similar to manifold learning, we make kNN test accuracy as a topology score due to kNN relying solely on the relationships among samples to classify. We utilize training sets as support sets (natural samples and adversarial samples generated by PGD-20) and methods [11, 14]. In Fig. 3 we set k as $5, 10, 20, 30, 40, 50$, and we observe that the choice of k does not affect the relative ranking of the topological relationships among samples in different representation spaces. So in Tables A1 and A2, we set k as 30.

Data pre-process. Similar to LBGAT [4], for CIFAR-10/100 datasets, 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, and the test set is normalized to $[0, 1]$ without any extra augmentation; For the training set of Tiny ImageNet, 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 $[0, 1]$, and for the test set, we resize the image to 32×32 and normalize pixel values to $[0, 1]$. Others are the same as CIFAR datasets.

Training details. For Tables 1 and 2, we follow state-of-the-art adversarial training method LAS [6]. ε is $8/255$, and 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 5×10^{-4}, and random seed is 1. ResNet-18 is the backbone of standard models, and WideResNet-

Defense	Natural Acc.	Robust Acc. PGD-20 Acc.	Robust Acc. C&W-20 Acc.	Robust Acc. AA Acc.	Topology Score Natural	Topology Score Robust
Standard Training	94.46	0.00	0.00	0.00	94.94	-
Vanilla AT [22]	86.69	53.45	53.72	48.95	86.51	53.94
Vanilla AT + TRAIN	**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 [4]	86.55	54.34	53.35	47.27	86.64	54.26
Vanilla AT + LBGAT + TRAIN	**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* [22]	84.42	56.59	54.91	51.91	85.58	56.73
TRADES + TRAIN	**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* [4]	81.98	**57.78**	55.53	53.14	84.57	57.79
TRADES + LBGAT + TRAIN	**87.62**(↑ 5.64)	57.73(↓ 0.05)	**58.08**(↑ 2.55)	**53.64**(↑ 0.50)	**89.50**(↑ 5.00)	57.98(↑ 0.19)

Table A1: Results on CIFAR-10. When added to the existing baseline under most settings, our method achieves both natural accuracy and robust accuracy improvements, particularly in terms of C&W-20 Acc. "*" are the results directly quoted from LBGAT.

Defense	Natural Acc.	Robust Acc. PGD-20 Acc.	Robust Acc. C&W-20 Acc.	Robust Acc. AA Acc.	Topology Score Natural	Topology Score Robust
Standard Training	77.39	0.00	0.00	0.00	77.07	-
Vanilla AT [22]	60.44	28.06	27.85	24.81	57.17	31.32
Vanilla AT + TRAIN	**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 [4]	61.01	**30.10**	28.09	25.63	61.28	30.47
Vanilla AT + LBGAT + TRAIN	**68.20**(↑ 7.19)	29.83(↓ 0.27)	**30.84**(↑ 2.75)	**25.88**(↑ 0.25)	**66.08**(↑ 4.80)	**32.48**(↑ 2.01)
TRADES* [22]	56.50	30.93	28.43	26.87	52.57	32.17
TRADES + TRAIN	**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* [4]	60.43	35.50	31.50	**29.34**	61.06	37.52
TRADES + LBGAT + TRAIN	**62.62**(↑ 2.19)	**36.27**(↑ 0.77)	**31.72**(↑ 0.22)	29.19(↓ 0.15)	**64.84** (↑ 3.78)	**38.25**(↑ 0.73)

Table A2: Results on CIFAR-100. Similar to Table A1, our method can improve the natural accuracy (up to 8.78%), robust accuracy (up to 3.04%), and topology score (up to 13.21%) of baselines. "*" are the results directly quoted from LBGAT.

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 λ. 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.

D Sensitivity of different learning rate

Experimental settings. For Table A1, Table A2, qualitative experiments, and all ablation experiments, we keep the same super-parameter configuration as LBGAT [4]. 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 a momentum of 0.9 and weight decay of 2×10^{-4}. Moreover, all our experimental results are reproducible with a random seed of 1.

Our method exhibits superior performance when applied with the new hyperparameters. According to Tables A1 and A2, TRAIN can effectively increase both natural and robust accuracy, and contribute to the topology preservation of both natural and adversarial samples.

In Table A1, TRAIN gets an improvement by 2.16% compared to vanilla AT baseline on natural data. It surpasses vanilla AT on PGD-20, C&W, and AA accuracy by 2.19%, 2.46%, and 1.94% respectively, indicating its high robustness. Our method also has improvements on

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(↑ 2.61)	16.18(↑ 2.70)
TRADES (β = 6.0) + LBGAT+ Ours	**41.53**(↑ 2.27)	**17.09**(↑ 0.67)

Table A3: Quantitative experiment on Tiny ImageNet. "*" are the results directly quoted from LBGAT.

LBGAT by 4.50% to 1.87% in all aspects. For another common baseline, TRADES, TRAIN also gets competitive results on both natural and adversarial data. Note that natural accuracy decreases when applying LBGAT to TRADES, so it also brings a large enhancement when combined with our method. For the topology score which is measured by kNN accuracy, TRAIN could boost the performance by a large margin. Since kNN classification is based only on inter-sample relationships, such results prove that TRAIN could mitigate topology disruptions of both natural and adversarial samples from adversarial training.

The overall results on CIFAR-100 are similar to CIFAR-10. As shown in Table A2, TRAIN performs better than vanilla AT and LBGAT and gets a further improvement when deployed with LBGAT simultaneously. For TRADES, our method surpasses it by a large margin (8.78%) on natural data and improves the robust accuracy by 3.04%. Adding LBGAT to TRAIN causes a decrease in natural accuracy but achieves the best accuracy in PGD-20 and C&W-20. The above results show that the proposed TRAIN could be applied to popular adversarial training pipelines for achieving SOTA performance on both natural accuracy and robust accuracy. Despite a slight decrease in individual robust metrics, we have achieved a better balance between natural accuracy and adversarial robustness overall. For the topology score, we can find that combining the baseline with TRAIN can further enhance the quality of topology for both natural and adversarial samples in the representation space.

E Quantitative results on Tiny ImageNet.

To demonstrate the effectiveness of our approach on a highly demanding dataset, we performed rigorous experiments on the Tiny Imagenet dataset. The results, as depicted in Table A3, clearly demonstrate that the combination of our algorithm with TRADES and LBGAT techniques leads to substantial improvements in both natural accuracy and adversarial robustness. When combined with Trades, our approach achieves a 2.61% improvement in natural accuracy and a 2.70% improvement in robust accuracy. When combined with TRADES+LBGAT, our method achieves a 2.27% improvement in natural accuracy and a 0.67% improvement in robust accuracy.

F More Ablation Studies

In this section, we delve into TRAIN to study its effectiveness in batch size, hyper-parameter λ, and model architectures. We also analyze the time complexity and training time of our method. All the ablation experiments are based on the CIFAR-100 dataset and combined with TRADES. All ablation experimental settings **(including ablation on different relationship preservation methods in our paper)** are the same as Tables A1 and A2.

Backbone of M'	Training Strategy	Backbone of M	Clean Acc.	Robust Acc. PGD-20 Acc.	C&W-20 Acc.	AA Acc.
None	Standard Training	ResNet-18	77.39	0	0	0
WideResNet34-10	TRADES + Ours	ResNet-18	62.62	36.27	31.72	29.19
None	Standard Training	WideResNet34-10	78.11	0	0	0
WideResNet34-10	TRADES + Ours	WideResNet34-10	63.09	35.54	30.41	28.76

Table A4: The ablation experiment about different backbones of the standard model.

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

Table A5: The ablation experiment about different batch sizes.

Impact of batch size. As shown in Table A5, 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 batch sizes is not large. Overall our method is not sensitive to different batch sizes.

To ensure fair comparisons with other methods, we chose a batch size of 128 for our other experiments.

	0	$5a$	$10a$	$20a$	$50a$
Natural Acc.	57.99	61.52	63.21	65.28	**66.40**
PGD-20 Acc.	31.53	32.31	33.47	**33.90**	33.62
L_{TP}	0.66	0.35	0.32	0.27	**0.24**

Table A6: Sensitivity analysis of hyper-parameter λ.

Sensitivity analysis of hyper-parameter λ. As Table A6 shows, with the increase of λ in Eq. (3), natural accuracy always gets higher; L_{TP} (calculated from the test set) gets lower; while the PGD-20 accuracy rises at first and then remains stable. It is reasonable because a large λ forces the topology of clean samples to be highly close to that of standard models. Finally, we set λ as $20a$ according to the PGD-20 accuracy following [10].

Impact of the different standard models. As depicted in Table A4, our approach exhibits robustness to variations in the backbones of standard models. Specifically, we observe that ResNet18 achieves a comparable trade-off between natural accuracy and adversarial robustness to WideresNet34-10 on the CIFAR-100 datasets while incurring lower training costs.

Table A7 shows the results of using different standard training strategies. To expedite the training process, a pre-trained standard model can be used in TRAIN (vanilla AT+TRAIN*). However, training the standard model and adversarial model jointly achieves superior results. This is attributed to the fact that the representation spaces of the two joint models are closer, facilitating optimization procedures.

Methods	Clean Acc	Robust Acc		
		PGD-20 Acc	C&W-20 Acc	AA Acc
Vanilla AT	60.44	28.06	27.85	24.81
Vanilla AT + TRAIN*	65.15	28.00	27.90	24.91
Vanilla AT + TRAIN	**66.39**	**29.88**	**29.84**	**25.81**

Table A7: Ablation experiment about different standard models on CIFAR-100. TRAIN* means using a well-trained standard model, and TRAIN means training two models jointly.

Methods	Clean Acc	Robust Acc		
		PGD-20 Acc	C&W-20 Acc	AA Acc
Vanilla AT	35.10	18.89	16.19	14.63
Vanilla AT+ TRAIN	38.58	20.25	17.64	15.39
TRADES	38.39	17.90	14.36	13.38
TRADES +TRAIN	43.64	18.52	14.86	13.51

Table A8: Experiments using MobileNetv3 on CIFAR100.

Impact of different backbones of M'. We conduct experiments on MobileNetv3, and the results reinforce the effectiveness of our approach across different backbones. As shown in Table A8, our method can further improve the baseline, especially in natural accuracy. We achieve a maximum improvement of 5.25% in natural accuracy and a maximum improvement of 1.45% in robust accuracy.

We can also find that the experimental results on MobileNet v3 are inferior compared to WideResNet34-10, both in terms of robustness and natural sample accuracy. This observation can be attributed to the positive correlation between the effectiveness of adversarial training algorithms and model capacity [1], and to reduce inference speed, MobileNet v3 has a significantly smaller model capacity compared to WideResNet34-10.

References

[1] Tao Bai, Jinqi Luo, Jun Zhao, Bihan Wen, and Qian Wang. Recent advances in adversarial training for adversarial robustness. *arXiv preprint arXiv:2102.01356*, 2021.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] Harini Kannan, Alexey Kurakin, and Ian Goodfellow. Adversarial logit pairing. *arXiv preprint arXiv:1803.06373*, 2018.

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

[9] 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.

[10] 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.

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

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

[13] 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.

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

[15] 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.

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

[17] 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.

[18] 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.

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