Supplementary Material for *Subspace Adversarial Training*

1. Detailed Sampling Strategy

We present a detailed sampling strategy of model checkpoints for DLDR [6] in Tab. 2, where we consider the times of uniform sampling in each epoch, the number of sampling epochs, the total number of samplings, and the dimension of the subspace extracted from the samplings. The general sampling strategy is quite simple: we uniformly sample a few checkpoints in every training epoch before the overfitting occurs. For single-step AT, the exact time when catastrophic overfitting occurs may be slightly different for different runs. Our sampling strategy is a conservative and robust one. We suggest sampling as more as model checkpoints before the overfitting occurs to estimate the subspace more accurately and to achieve better results. Note that the total number of samplings t is small (around 200), and thus the computational overhead on corresponding decomposition (a $t \times t$ matrix) is small. This explains why the computational cost of the decomposition is negligible compared to the total computational cost. A more delicate sampling strategy design may improve the performance and could be an interesting topic for future works.

Ablation study for the dimension d. We provide an ablation study for the dimension of subspace in Fig. 1, where we vary the dimension of subspace for Fast Sub-AT from 50 to 100 and record the best robust accuracy obtained in corresponding subspaces. We observe that the best robust accuracy varies very little across such a wide range of dimensions (<1%). Thus we conclude that our Sub-AT is robust to the exact choice of the dimension d.

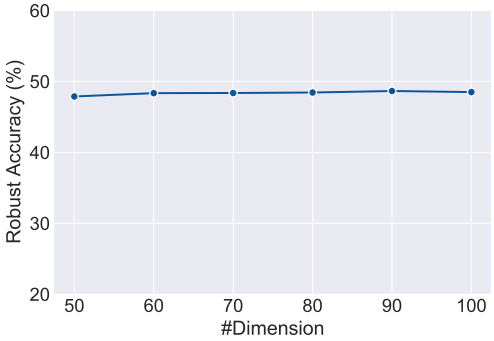

Figure 1. Ablation study for the dimension of subspace on CIFAR-10. Fast Sub-AT achieves similar best robustness performance across a wide range of dimensions. The robust accuracy is evaluated under PGD-20 attack.

2. Additional Results with Auto-Attack

We further evaluate the performance of Sub-AT on mitigating robust overfitting in multi-step AT with Auto-Attack, a stronger and more reliable attack proposed by Croce *et al*. [3]. From the results in Tab. 1, we observe that indeed, Sub-AT effectively mitigates the robust overfitting and consistently improves the robust accuracy. Hence, training in subspace genuinely overcomes the overfitting and improves the model robustness rather than as a result of gradient masking [1, 7].

Table 1. Robust accuracy of multi-step AT and Sub-AT on CIFAR-10/100 against ℓ_2 and ℓ_∞ adversarial perturbations. The robust accuracy is evaluated under Auto-Attack [3]. Our Sub-AT achieves consistent improvements on the robust accuracy while effectively mitigating the robust overfitting. The best robustness performances and the smallest difference between the best and the final are marked in **bold**.

Dataset	Norm	Settings	Auto-Attack		
			Best	Final	Diff.
CIFAR-10	ℓ_∞	AT	47.29	41.08	6.21
		Sub-AT	**48.37**	**47.88**	**0.49**
	ℓ_2	AT	65.66	63.93	1.73
		Sub-AT	**66.99**	**67.15**	**-0.16**
CIFAR-100	ℓ_∞	AT	22.83	18.12	4.71
		Sub-AT	**23.89**	**23.83**	**0.06**
	ℓ_2	AT	36.91	32.70	4.21
		Sub-AT	**38.14**	**37.84**	**0.30**

3. Results on Tiny-ImageNet

The results on Tiny-ImageNet [4] is presented in Tab. 3, where we use PreAct ResNet18 model [5] and train it for 100 epochs with learning rate decay at 50 and 80 following [2]. Sub-AT is trained for 20 epochs with a constant learning rate 1, accordingly. For single-step Fast AT, we observe that the catastrophic overfitting occurs at the 17th epoch, and thus we only sample 16 epochs for DLDR. Even with such few samplings, we obtain an 18.79% single-step robust accuracy, while standard PGD-10 AT achieves a slightly better 19.84% robust accuracy, but with around **12x** training time overhead and, even worse, serious robust overfitting problem. Within the better subspace extracted from PGD-10 AT, our Fast Sub-AT achieves **21.87**% robust accuracy, which is significantly better than base PGD-10 AT and also enjoys computational benefits.

Table 2. Detailed sampling strategy for DLDR. We report the times of uniformly sampling in each epoch of training, the number of sampling epochs, the total number of samplings t, and the dimension of the subspace extracted from the samplings of parameters. Note that there is an additional sampling on the parameter initialization as we start the Sub-AT from the initialization.

Type	Datasets	Method	#Times/epoch	#Epochs	t	#Dimension
Single-step	CIFAR-10	Fast AT	2	65	131	80
	CIFAR-10	GradAlign	2	65	131	100
	CIFAR-10	GAT	2	100	201	100
	CIFAR-100	Fast AT	2	65	131	80
	CIFAR-100	GradAlign	2	65	131	100
	CIFAR-100	GAT	1	140	141	100
	Tiny-ImageNet	Fast AT	4	16	65	50
Multi-step	CIFAR-10	PGD-10 AT	2	100	201	120
	CIFAR-100	PGD-10 AT	2	100	201	120
	Tiny-ImageNet	PGD-10 AT	4	50	201	120

Table 3. Results on Tiny-ImageNet. We use PreAct ResNet18 model and consider both single-step and multi-step AT. Sub-AT demonstrates its superior performance in both robustness performance and computational overhead. The time consumption is evaluated on an Nvidia Tesla V100.

	Method	Subspace	Best		Final		Time
			Natural	PGD-20	Natural	PGD-20	
Single-step	Fast AT	–	28.79	11.90	42.54	0.00	3.5h
	Fast Sub-AT ($\epsilon = 8/255$)	Fast AT	39.38	16.75	39.19	16.17	1.3h
	Fast Sub-AT ($\epsilon = 12/255$)	Fast AT	38.11	18.52	38.306	18.13	1.3h
	Fast Sub-AT ($\epsilon = 16/255$)	Fast AT	37.32	**18.79**	37.20	**18.22**	1.3h
Multi-step	PGD-10 AT	–	42.76	19.84	46.57	14.18	15.7h
	PGD-10 Sub-AT	PGD-10 AT	40.82	20.52	40.78	19.55	12.1h
	PGD-10 AT ($\epsilon = 12/255$)	PGD-10 AT	40.88	21.42	42.27	21.41	8.6h
	PGD-10 AT ($\epsilon = 16/255$)	PGD-10 AT	40.99	**21.87**	41.27	**21.60**	8.6h

Table 4. Results on multi-step PGD-10 AT with WideResNet-28-10 model against PGD-20 attack (ℓ_∞ norm, $\epsilon = 8/255$).

Dataset	Settings	Robust Accuracy			Natural Accuracy		
		Best	Final	Diff.	Best	Final	Diff.
CIFAR-10	AT	53.26	46.70	6.56	84.69	85.81	-1.12
	Sub-AT	**55.14**	**54.75**	**0.39**	84.79	84.71	0.08
CIFAR-100	AT	29.44	23.26	6.18	57.64	56.08	1.56
	Sub-AT	**31.07**	**30.69**	**0.38**	57.24	57.40	-0.16

4. Results on Wide-ResNet

We conduct further experiments on Wide-ResNet [8] architectures. Specifically, we use the WideResNet-28-10 model and consider ℓ_∞ adversarial perturbations with radius $8/255$ on CIFAR-10 and CIFAR-100. From the results in Tab. 4, we observe that similarly, robust overfitting can be successfully mitigated by Sub-AT, meanwhile with significant improvements in robustness. We achieve **+1.88**% on CIFAR-10 and **+1.63**% on CIFAR-100, and successfully control the robust accuracy gap (between the best and the final) under 0.4%. Thus we conclude that Sub-AT can be easily applied to other architectures and obtain similar enhancements in robustness.

References

[1] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference*

on Machine Learning (ICML), pages 274–283. PMLR, 2018. 1

[2] Tianlong Chen, Zhenyu Zhang, Sijia Liu, Shiyu Chang, and Zhangyang Wang. Robust overfitting may be mitigated by properly learned smoothening. In *International Conference on Learning Representations (ICLR)*, 2020. 1

[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 (ICML)*, pages 2206–2216. PMLR, 2020. 1

[4] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255, 2009. 1

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016. 1

[6] Tao Li, Lei Tan, Qinghua Tao, Yipeng Liu, and Xiaolin Huang. Low dimensional landscape hypothesis is true: DNNs can be trained in tiny subspaces. *arXiv preprint arXiv:2103.11154*, 2021. 1

[7] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519, 2017. 1

[8] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference (BMVC)*, 2016. 2