Additional results for the rebuttal

Table A1: **Ablation for data augmentations.** We show how varying levels of data augmentations influences robustness across all threat models in our training paradigm.

Model	Augmentation	clean	ℓ_∞	ℓ_2	ℓ_1
ViT-S+ConvStem	basic	66.5	38.5	41.3	19.2
ViT-S+ConvStem	3-Aug [1]	70.1	43.6	46.1	23.4
ViT-S+ConvStem	heavy	72.5	48.5	50.4	26.7

Table A2: **PGD vs APGD.** Accuracy difference after adversarial training of a ConvNeXt-T for 50 epochs using either PGD-2 or APGD-2.

Train-Attack	clean	ℓ_∞	ℓ_2	ℓ_1
PGD-2	72.8	45.4	39.1	16.5
APGD-2	71.0	46.8	38.1	14.9

Table A3: **Evaluation at an increased resolution.** For ConvNeXt-B+ConvStem, we see that both best clean and ℓ_∞ robust accuracies are attained a a higher resolution than the one trained for (224) across all epsilon values. The difference from base number at the resolution of 224 is shown in color.

ϵ_∞	Input resolution								
	192		224*	256		288		320	
clean	74.1	-1.8	75.9	76.9	+1.0	**77.7**	+1.8	77.2	+1.3
2/255	64.6	-2.3	66.9	67.9	+1.0	**68.6**	+1.7	68.4	+1.5
4/255	53.0	-3.2	56.1	**57.3**	+1.2	57.2	+1.1	56.6	+0.5
6/255	41.0	-2.8	43.8	44.4	+0.6	**44.5**	+0.7	43.0	-0.8
8/255	29.5	-0.9	30.4	**31.0**	+0.6	29.8	-0.6	27.9	-2.5

Table A4: **Non-ConvStem models for medium sized architectures added.** Comparing ViT-M+ConvStem and ConvNeXt-S+ConvStem with their standard non-ConvStem counterparts. All models are trained with the same setup for 50 epochs. The change on adding ConvStem is shown in color.

Model	Adversarial Tr. w.r.t. ℓ_∞							
	clean		ℓ_∞		ℓ_2		ℓ_1	
ViT-M	71.7		47.2		49.0		29.2	
ViT-M + ConvStem	72.4	+0.7	48.8	+1.6	50.6	+1.6	28.1	-1.1
ConvNeXt-S	74.1		52.3		43.8		19.5	
ConvNeXt-S + ConvStem	74.1		52.4	+0.1	50.9	+7.1	25.6	+5.1

References

[1] Edoardo Debenedetti, Vikash Sehwag, and Prateek Mittal. A light recipe to train robust vision transformers. In *First IEEE Conference on Secure and Trustworthy Machine Learning*, 2023.