A APPENDIX

A.1 MULTIPLE RUNS

We run RobMask for three times with different random seed and present the mean in Table 8. We report the more detailed result in Table 8:

Model	#Epochs	CIFAR-10	CIFAR-100
ResNet-18	20	94.54±0.22	77.41± 0.20%
	100	95.90±0.24	8.59±0.18%
DenseNet-121	20	95.10± 0.20%	75.94± 0.17%
	100	96.29± 0.18%	79.99± 0.18%
Preact-18	20	94.92± 0.12%	73.43± 0.16%
	100	95.83± 0.14%	78.01±0.18%
ResNeXt-29	20	94.83± 0.21%	74.43± 0.22%
	100	96.84± 0.22%	79.31± 0.23%

Table 8: RobMask results on CIFAR-10/100 over ResNet-18, DenseNet-121, Preact-18, and ResNeXt-29. Models are trained for 20 and 100 epochs.

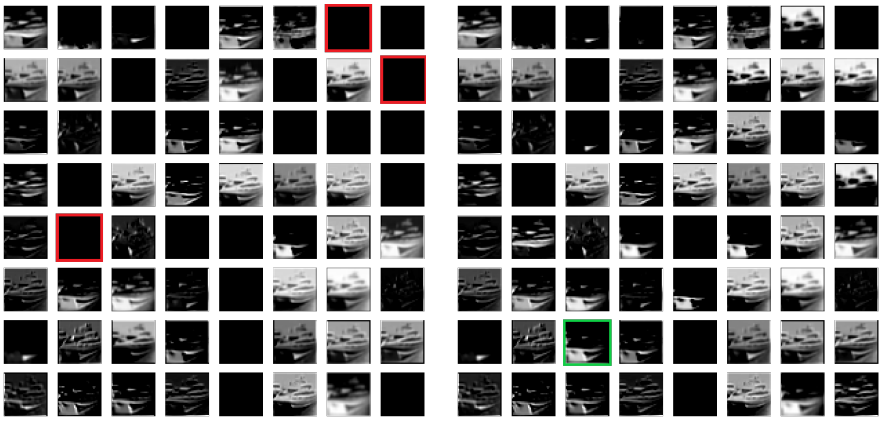
A.2 COMPARISON ON BATCH NORMALIZATIONS

In this section, we extend our experiment in Figure 1 to show the batch statistics in the deeper layers across the neural networks. It clearly shows that the rescaling weight has more effect than other parameters in the batch normalization. To be noted, since the deep layer's mean and variance would be affected by the shallow layers' rescaling weight parameter, the result on the deeper layer couldn't disentangle the effect between normalization and rescaling because it is mixed.

	Mean	Variance	Weight	Bias
Layer 0	1.0	1.0	0.7620	1.0
Layer 1	0.9842	0.9718	0.7883	1.0
Layer 2	0.9530	0.9199	0.7544	1.0
Layer 3	0.9743	0.9691	0.8594	1.0
Layer 4	0.8894	0.9340	0.8126	1.0
Layer 5	0.9555	0.9516	0.8813	1.0
Layer 6	0.9853	0.9452	0.7141	1.0
Layer 7	0.9554	0.9169	0.8609	1.0
Layer 8	0.9903	0.9646	0.8961	1.0
Layer 9	0.9635	0.9755	0.8046	1.0
Layer 10	0.9823	0.9522	0.9396	1.0
Layer 11	0.9823	0.9769	0.7906	1.0
Layer 12	0.9753	0.9593	0.7839	1.0
Layer 13	0.9914	0.9874	0.8891	1.0
Layer 14	0.9699	0.9898	0.6593	1.0
Layer 15	0.9902	0.9870	0.8605	1.0
Layer 16	0.9603	0.9889	0.7423	1.0
Layer 17	0.9736	0.9742	0.6809	1.0
Layer 18	0.9772	0.9838	0.9573	1.0

Table 9: Cosine similarity on under every batch normalization layer under standard fine-tuned training on adversarial trained model

A.3 ADVERSARIAL MASKING VISULIAZATION



(a) Adversarial training with further standard finetuning of BN

(b) Adversarial training

Figure 4: Illustration of the Adversarial Masking effect. We mark several feature maps (red and green boxes) are blocked out or magnified when comparing (a) and (b), which can be viewed as a selection mask on "non-robust" and "robust" features.