Table 1. **Comparison untargeted vs targeted attacks.** We report clean test accuracy and robust accuracy achieved by APGD_{DLR}, $\text{APGD}^{\text{T}}_{\text{DLR}}$, FAB and FAB^{T}. Moreover, we show the difference between the results of the targeted and untargeted attacks, and boldface it when is negative (that is the targeted attack is stronger). FAB does not scale to CIFAR-100/ImageNet due to the large number of classes.

#	paper	clean	APGD_{DLR}	$\text{APGD}^{\text{T}}_{\text{DLR}}$	diff.	FAB	FAB^{T}	diff.
CIFAR-10 - l_∞ - $\epsilon = 8/255$								
1	(Carmon et al., 2019)	89.69	60.64	59.54	**-1.10**	60.62	60.12	**-0.50**
2	(Alayrac et al., 2019)	86.46	62.03	56.27	**-5.76**	58.20	56.81	**-1.39**
3	(Hendrycks et al., 2019)	87.11	56.96	54.94	**-2.02**	55.40	55.27	**-0.13**
4	(Rice et al., 2020)	85.34	55.72	53.43	**-2.29**	54.13	53.83	**-0.30**
5	(Qin et al., 2019)	86.28	55.46	52.85	**-2.61**	53.77	53.28	**-0.49**
6	(Engstrom et al., 2019)	87.03	52.65	49.32	**-3.33**	50.37	49.81	**-0.56**
7	(Kumari et al., 2019)	87.80	51.68	49.15	**-2.53**	49.87	49.54	**-0.33**
8	(Mao et al., 2019)	86.21	50.33	47.44	**-2.89**	48.32	47.91	**-0.41**
9	(Zhang et al., 2019a)	87.20	47.33	44.85	**-2.48**	45.66	45.39	**-0.27**
10	(Madry et al., 2018)	87.14	46.03	44.28	**-1.75**	45.41	44.75	**-0.66**
11	(Pang et al., 2020)	80.89	44.56	43.50	**-1.06**	44.47	44.06	**-0.41**
12	(Wong et al., 2020)	83.34	46.64	43.22	**-3.42**	44.05	43.74	**-0.31**
13	(Shafahi et al., 2019)	86.11	44.56	41.64	**-2.92**	42.90	43.44	0.54
14	(Ding et al., 2020)	84.36	50.26	41.74	**-8.52**	47.18	42.47	**-4.71**
15	(Moosavi-Dezfooli et al., 2019)	83.11	40.29	38.50	**-1.79**	39.04	38.97	**-0.07**
16	(Zhang & Wang, 2019)	89.98	48.96	37.29	**-11.67**	40.84	38.48	**-2.36**
17	(Zhang & Xu, 2020)	90.25	49.40	37.54	**-11.86**	40.36	38.99	**-1.37**
18	(Jang et al., 2019)	78.91	37.01	34.96	**-2.05**	35.54	35.50	**-0.04**
19	(Kim & Wang, 2020)	91.51	48.41	35.93	**-12.48**	38.88	35.41	**-3.47**
20	(Moosavi-Dezfooli et al., 2019)	80.41	35.47	33.70	**-1.77**	34.08	34.08	**0.00**
21	(Wang & Zhang, 2019)	92.80	40.33	33.61	**-6.72**	33.51	31.19	**-2.32**
22	(Wang & Zhang, 2019)	92.82	36.58	29.73	**-6.85**	31.82	29.10	**-2.72**
23	(Mustafa et al., 2019)	89.16	4.54	1.13	**-3.41**	1.12	0.71	**-0.41**
24	(Pang et al., 2020)	93.52	0.49	0.00	**-0.49**	0.03	0.00	**-0.03**
CIFAR-10 - l_∞ - $\epsilon = 0.031$								
1	(Zhang et al., 2019b)	84.92	54.04	53.10	**-0.94**	53.79	53.45	**-0.34**
2	(Atzmon et al., 2019)	81.30	44.50	41.16	**-3.34**	40.92	40.73	**-0.19**
3	(Xiao et al., 2020)	79.28	31.27	32.34	1.07	79.28	79.28	**0.00**
CIFAR-100 - l_∞ - $\epsilon = 8/255$								
1	(Hendrycks et al., 2019)	59.23	31.66	28.48	**-3.18**	-	28.74	-
2	(Rice et al., 2020)	53.83	20.25	18.98	**-1.27**	-	19.24	-
MNIST - l_∞ - $\epsilon = 0.3$								
1	(Zhang et al., 2020)	98.38	94.77	94.88	0.11	95.60	96.84	1.24
2	(Gowal et al., 2019)	98.34	93.84	93.93	0.09	94.72	97.03	2.31
3	(Zhang et al., 2019b)	99.48	93.96	93.58	**-0.38**	94.12	94.62	0.50
4	(Ding et al., 2020)	98.95	94.03	94.62	0.59	94.33	95.37	1.04
5	(Atzmon et al., 2019)	99.35	94.54	94.16	**-0.38**	94.12	95.26	1.14
6	(Madry et al., 2018)	98.53	89.75	90.57	0.82	91.75	93.69	1.94
7	(Jang et al., 2019)	98.47	92.15	93.56	1.41	93.24	94.74	1.50
8	(Wong et al., 2020)	98.50	85.39	86.34	0.95	87.30	88.28	0.98
9	(Taghanaki et al., 2019)	98.86	0.00	0.00	**0.00**	0.02	0.01	**-0.01**
ImageNet - l_∞ - $\epsilon = 4/255$								
1	(Engstrom et al., 2019)	63.4	32.0	27.7	**-4.3**	-	28.4	-
CIFAR-10 - l_2 - $\epsilon = 0.5$								
1	(Augustin et al., 2020)	91.08	74.94	72.91	**-2.03**	74.13	73.18	**-0.95**
2	(Engstrom et al., 2019)	90.83	70.20	69.24	**-0.96**	69.54	69.46	**-0.08**
3	(Rice et al., 2020)	88.67	68.95	67.68	**-1.27**	68.03	67.97	**-0.06**
4	(Rony et al., 2019)	89.05	67.02	66.44	**-0.58**	66.81	66.74	**-0.07**
5	(Ding et al., 2020)	88.02	66.53	66.09	**-0.44**	66.43	66.33	**-0.10**
ImageNet - l_2 - $\epsilon = 3$								
1	(Engstrom et al., 2019)	55.3	30.9	28.3	**-2.6**	-	28.5	-

Table 2. **Robustness evaluation of adversarial defenses by *AutoAttack*.** We report clean test accuracy, the robust accuracy of the individual attacks as well as the combined one of *AutoAttack* (AA column). We also provide the robust accuracy reported in the original papers and compute the difference to the one of *AutoAttack*. If negative (in red) *AutoAttack* provides lower (better) robust accuracy.

#	paper	clean	APGD_{CE}	$\text{APGD}^{\text{T}}_{\text{DLR}}$	FAB^{T}	Square	AA	reported	reduct.
CIFAR-10 - l_∞ - $\epsilon = 8/255$									
1	(Carmon et al., 2019)	89.69	61.74	**59.54**	60.12	66.63	59.53	62.5	-2.97
2	(Alayrac et al., 2019)	86.46	60.17	**56.27**	56.81	66.37	56.03	56.30	-0.27
3	(Hendrycks et al., 2019)	87.11	57.23	**54.94**	55.27	61.99	54.92	57.4	-2.48
4	(Rice et al., 2020)	85.34	57.00	**53.43**	53.83	61.37	53.42	58	-4.58
5	(Qin et al., 2019)	86.28	55.70	52.85	53.28	60.01	52.84	52.81	0.03
6	(Engstrom et al., 2019)	87.03	51.72	**49.32**	49.81	58.12	49.25	53.29	-4.04
7	(Kumari et al., 2019)	87.80	51.80	**49.15**	49.54	58.20	49.12	53.04	-3.92
8	(Mao et al., 2019)	86.21	49.65	**47.44**	47.91	56.98	47.41	50.03	-2.62
9	(Zhang et al., 2019a)	87.20	46.15	**44.85**	45.39	55.08	44.83	47.98	-3.15
10	(Madry et al., 2018)	87.14	44.75	**44.28**	44.75	53.10	44.04	47.04	-3.00
11	(Pang et al., 2020)	80.89	57.07	**43.50**	44.06	49.73	43.48	55.0	-11.52
12	(Wong et al., 2020)	83.34	45.90	**43.22**	43.74	53.32	43.21	46.06	-2.85
13	(Shafahi et al., 2019)	86.11	43.66	**41.64**	43.44	51.95	41.47	46.19	-4.72
14	(Ding et al., 2020)	84.36	50.12	**41.74**	42.47	55.53	41.44	47.18	-5.74
15	(Moosavi-Dezfooli et al., 2019)	83.11	41.72	**38.50**	38.97	47.69	38.50	41.4	-2.90
16	(Zhang & Wang, 2019)	89.98	64.42	**37.29**	38.48	59.12	36.64	60.6	-23.96
17	(Zhang & Xu, 2020)	90.25	71.40	**37.54**	38.99	66.88	36.45	68.7	-32.25
18	(Jang et al., 2019)	78.91	37.76	**34.96**	35.50	44.33	34.95	37.40	-2.45
19	(Kim & Wang, 2020)	91.51	56.64	35.93	**35.41**	61.30	34.22	57.23	-23.01
20	(Moosavi-Dezfooli et al., 2019)	80.41	36.65	**33.70**	34.08	43.46	33.70	36.3	-2.60
21	(Wang & Zhang, 2019)	92.80	59.09	33.61	**31.19**	64.22	29.35	58.6	-29.25
22	(Wang & Zhang, 2019)	92.82	69.62	29.73	**29.10**	66.77	26.93	66.9	-39.97
23	(Mustafa et al., 2019)	89.16	8.16	1.13	**0.71**	33.91	0.28	32.32	-32.04
24	(Chan et al., 2020)	93.79	2.06	**0.53**	58.13	71.43	0.26	15.5	-15.24
25	(Pang et al., 2020)	93.52	89.48	**0.00**	**0.00**	35.82	0.00	31.4	-31.40
CIFAR-10 - l_∞ - $\epsilon = 0.031$									
1	(Zhang et al., 2019b)	84.92	55.28	**53.10**	53.45	59.43	53.08	56.43	-3.35
2	(Atzmon et al., 2019)	81.30	79.67	41.16	**40.73**	47.99	40.22	43.17	-2.95
3	(Xiao et al., 2020)	79.28	39.99	32.34	79.28	**20.44**	18.50	52.4	-33.90
CIFAR-100 - l_∞ - $\epsilon = 8/255$									
1	(Hendrycks et al., 2019)	59.23	33.02	**28.48**	28.74	34.26	28.42	33.5	-5.08
2	(Rice et al., 2020)	53.83	20.57	**18.98**	19.24	23.57	18.95	28.1	-9.15
MNIST - l_∞ - $\epsilon = 0.3$									
1	(Zhang et al., 2020)	98.38	95.32	94.88	96.84	**93.97**	93.96	96.38	-2.42
2	(Gowal et al., 2019)	98.34	94.79	93.93	97.03	**92.88**	92.83	93.88	-1.05
3	(Zhang et al., 2019b)	99.48	93.60	93.58	94.62	**92.97**	92.81	95.60	-2.79
4	(Ding et al., 2020)	98.95	94.58	94.62	95.37	**91.42**	91.40	92.59	-1.19
5	(Atzmon et al., 2019)	99.35	99.10	94.16	95.26	**90.86**	90.85	97.35	-6.50
6	(Madry et al., 2018)	98.53	90.57	90.57	93.69	**88.56**	88.50	89.62	-1.12
7	(Jang et al., 2019)	98.47	94.05	93.56	94.74	**88.00**	87.99	94.61	-6.62
8	(Wong et al., 2020)	98.50	86.68	86.34	88.28	**83.07**	82.93	88.77	-5.84
9	(Taghanaki et al., 2019)	98.86	30.50	**0.00**	0.01	**0.00**	0.00	64.25	-64.25
ImageNet - l_∞ - $\epsilon = 4/255$									
1	(Engstrom et al., 2019)	63.4	31.0	**27.7**	28.4	46.8	27.6	33.38	-5.78
CIFAR-10 - l_2 - $\epsilon = 0.5$									
1	(Augustin et al., 2020)	91.08	74.70	**72.91**	73.18	83.10	72.91	73.27	-0.36
2	(Engstrom et al., 2019)	90.83	69.62	**69.24**	69.46	80.92	69.24	70.11	-0.87
3	(Rice et al., 2020)	88.67	68.58	**67.68**	67.97	79.01	67.68	71.6	-3.92
4	(Rony et al., 2019)	89.05	66.59	**66.44**	66.74	78.05	66.44	67.6	-1.16
5	(Ding et al., 2020)	88.02	66.21	**66.09**	66.33	76.99	66.09	66.18	-0.09
ImageNet - l_2 - $\epsilon = 3$									
1	(Engstrom et al., 2019)	55.3	31.5	**28.3**	28.5	46.6	28.3	35.09	-6.79

Table 3. **Robustness evaluation of randomized l_∞-adversarial defenses by *AutoAttack*.** We report the clean test accuracy (mean and standard deviation over 5 runs) and the robust accuracy of the individual attacks as well as the combined on of *AutoAttack* (again over 5 runs). We also provide the robust accuracy reported in the respective papers and compute the difference to the one of *AutoAttack* (negative means that *AutoAttack* is better). The statistics of our attack are computed on the whole test set except for the ones of (Yang et al., 2019), which are on 1000 test points due to the computational cost of this defense. The ϵ is the same as used in the papers.

#	paper	model	clean	APGD_{CE}	APGD_{DLR}	FAB	Square	*AutoAttack*	report.	reduct.
CIFAR-10 - $\epsilon = 8/255$										
1	(Wang et al., 2019)	En_5RN	82.39 (0.14)	**48.81**	49.37	-	78.61	45.56 (0.20)	51.48	-5.9
2	(Yang et al., 2019)	with AT	84.9 (0.6)	**30.1**	31.9	-	-	26.3 (0.85)	52.8	-26.5
3	(Yang et al., 2019)	pure	87.2 (0.3)	**21.5**	24.3	-	-	18.2 (0.82)	40.8	-22.6
4	(Grathwohl et al., 2020)	JEM-10	90.99 (0.03)	**11.69**	15.88	63.07	79.32	9.92 (0.03)	47.6	-37.7
5	(Grathwohl et al., 2020)	JEM-1	92.31 (0.04)	**9.15**	13.85	62.71	79.25	8.15 (0.05)	41.8	-33.6
6	(Grathwohl et al., 2020)	JEM-0	92.82 (0.05)	**7.19**	12.63	66.48	73.12	6.36 (0.06)	19.8	-13.4
CIFAR-10 - $\epsilon = 4/255$										
1	(Grathwohl et al., 2020)	JEM-10	91.03 (0.05)	**49.10**	52.55	78.87	89.32	47.97 (0.05)	72.6	-24.6
2	(Grathwohl et al., 2020)	JEM-1	92.34 (0.04)	**46.08**	49.71	78.93	90.17	45.49 (0.04)	67.1	-21.6
3	(Grathwohl et al., 2020)	JEM-0	92.82 (0.02)	**42.98**	47.74	82.92	89.52	42.55 (0.07)	50.8	-8.2

In most of the cases more than one of the attacks included in *AutoAttack* achieves a lower robust accuracy than reported (APGD_{CE} improves the reported evaluation in 21/49 cases, $\text{APGD}^{\text{T}}_{\text{DLR}}$ in 45/49, FAB^{T} in 39/49 and Square Attack in 17/49, but 9/9 on MNIST). $\text{APGD}^{\text{T}}_{\text{DLR}}$ most often attains the best result for CIFAR-10, CIFAR-100 and ImageNet, Square Attack on MNIST. Also, $\text{APGD}^{\text{T}}_{\text{DLR}}$ is the most reliable one as it has the least severe failure which we define as the largest difference in robust accuracy to the best performing attack (maximal difference less than 12%, compared to 89% for APGD_{CE}, 59% for FAB^{T} and 70% for Square Attack). Thus our new DLR loss is able to resist gradient masking.

Randomized defenses: Another line of adversarial defenses relies on adding to a classifier some stochastic component. In this case the output (hence the decision) of the model might change across different runs for the same input. Thus we compute in Table 3 the mean (standard deviation in brackets) of our statistics over 5 runs. Moreover the results of *AutoAttack* are given considering, for each point, the attack performing better on average across 5 runs. To counter the randomness of the classifiers, for APGD we compute the direction for the update step as the average of 20 computations of the gradient at the same point (known as Expectation over Transformation (Athalye et al., 2018)) and use the untargeted losses (1 run). We do not run FAB here since it returns points on or very close to the decision boundary, so that even a small variation in the classifier is likely to undo the adversarial change. We modify Square Attack to accept an update if it reduces the target loss on average over 20 forward passes and, as this costs more time we use only 1000 iterations. For the models from (Grathwohl et al., 2020), we attack that named JEM-0 with 5 restarts with the deterministic versions (i.e. without averaging across multiple passes of the networks), since the stochastic component has little influence, and then reuse the same adversarial examples on JEM-1 and JEM-10 (the results of FAB confirm that it is not suitable to test randomized defenses). Table 3 shows that *AutoAttack* achieves always lower robust accuracy than reported in the respective papers, with APGD_{CE} being the best performing attack, closely followed by APGD_{DLR}. In 7 out of 9 cases the improvement is significant, larger than 10% (and in 3/9 cases larger than 25%). Thus *AutoAttack* is also suitable for the evaluation of randomized defenses.

6.1. Analysis of SOTA of adversarial defenses

While the main goal of the evaluation is to show the effectiveness of *AutoAttack*, at the same time it provides an assessment of the SOTA of adversarial defenses. The most robust defenses rely on variations or fine-tuning of adversarial training introduced in (Madry et al., 2018). One step forward has been made by methods which use additional data for training, like (Carmon et al., 2019) and (Alayrac et al., 2019). Moreover, several defenses which claim SOTA robustness turn out to be significantly less robust than (Madry et al., 2018). Interestingly, the most (empirically) resistant model on MNIST is one trained for obtaining provable certificates on the exact robust accuracy, and comes with a verified lower bound on it of 93.32% (Zhang et al., 2020).

While this paper contains up to our knowledge the largest independent evaluation of current adversarial defenses, this is by no means an exhaustive survey. Several authors did not reply to our request or were not able to provide models (or at least code). We thank all the authors who helped us in this evaluation. We hope that *AutoAttack* will contribute to a faster development of adversarial defenses and recommend it as part of a standard evaluation pipeline as it is quick and requires no hyperparameter tuninig.

Acknowledgements

We are very grateful to Alvin Chan, Chengzhi Mao, Seyed-Mohsen Moosavi-Dezfooli, Chongli Qin, Saeid Asgari Taghanaki, Bao Wang and Zhi Xu for providing code, models and answering questions on their papers. We also thank Maksym Andriushchenko for insightful discussions about this work. We acknowledge support from the German Federal Ministry of Education and Research (BMBF) through the Tübingen AI Center (FKZ: 01IS18039A). This work was also supported by the DFG Cluster of Excellence "Machine Learning – New Perspectives for Science", EXC 2064/1, project number 390727645, and by DFG grant 389792660 as part of TRR 248.

References

Alayrac, J.-B., Uesato, J., Huang, P.-S., Fawzi, A., Stanforth, R., and Kohli, P. Are labels required for improving adversarial robustness? In *NeurIPS*, 2019.

Andriushchenko, M., Croce, F., Flammarion, N., and Hein, M. Square attack: a query-efficient black-box adversarial attack via random search. In *ECCV*, 2020.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *ICML*, 2018.

Atzmon, M., Haim, N., Yariv, L., Israelov, O., Maron, H., and Lipman, Y. Controlling neural level sets. In *NeurIPS*, 2019.

Augustin, M., Meinke, A., and Hein, M. Adversarial robustness on in-and out-distribution improves explainability. In *ECCV*, 2020.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *IEEE Symposium on Security and Privacy*, 2017a.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, 2017b.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., and Kurakin, A. On evaluating adversarial robustness. version from May 14, 2019, 2019. URL https://github.com/evaluating-adversarial-robustness/adv-eval-paper.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. S. Unlabeled data improves adversarial robustness. In *NeurIPS*, pp. 11190–11201, 2019.

Chan, A., Tay, Y., Ong, Y. S., and Fu, J. Jacobian adversarially regularized networks for robustness. In *ICLR*, 2020.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *NeurIPS*, 2019.

Croce, F. and Hein, M. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *ICML*, 2020.

Croce, F., Andriushchenko, M., and Hein, M. Provable robustness of relu networks via maximization of linear regions. In *AISTATS*, 2019a.

Croce, F., Rauber, J., and Hein, M. Scaling up the randomized gradient-free adversarial attack reveals overestimation of robustness using established attacks. *International J. of Computer Vision (IJCV)*, 2019b.

Ding, G. W., Wang, L., and Jin, X. AdverTorch v0.1: An adversarial robustness toolbox based on pytorch. arXiv preprint arXiv:1902.07623, 2019.

Ding, G. W., Sharma, Y., Lui, K. Y. C., and Huang, R. Mma training: Direct input space margin maximization through adversarial training. In *ICLR*, 2020.

Engstrom, L., Ilyas, A., Santurkar, S., and Tsipras, D. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *ICLR*, 2015.

Gowal, S., Dvijotham, K., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T. A., and Kohli, P. On the effectiveness of interval bound propagation for training verifiably robust models. In *ICCV*, 2019.

Grathwohl, W., Wang, K.-C., Jacobsen, J.-H., Duvenaud, D., Norouzi, M., and Swersky, K. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*, 2020.

Hendrycks, D., Lee, K., and Mazeika, M. Using pre-training can improve model robustness and uncertainty. In *ICML*, pp. 2712–2721, 2019.

Jang, Y., Zhao, T., Hong, S., and Lee, H. Adversarial defense via learning to generate diverse attacks. In *ICCV*, 2019.

Kim, J. and Wang, X. Sensible adversarial learning, 2020. URL https://openreview.net/forum?id=rJlf_RVKwr.

Kumari, N., Singh, M., Sinha, A., Machiraju, H., Krishnamurthy, B., and Balasubramanian, V. N. Harnessing the vulnerability of latent layers in adversarially trained models. In *IJCAI*, pp. 2779–2785, 7 2019.

Kurakin, A., Goodfellow, I. J., and Bengio, S. Adversarial examples in the physical world. In *ICLR Workshop*, 2017.

Lecuyer, M., Atlidakis, V., Geambasu, R., Hsu, D., and Jana, S. Certified robustness to adversarial examples with differential privacy. In *IEEE Symposium on Security and Privacy (SP)*, 2019.

Li, B., Chen, C., Wang, W., and Carin, L. Certified adversarial robustness with additive noise. In *NeurIPS*, 2019.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Valdu, A. Towards deep learning models resistant to adversarial attacks. In *ICLR*, 2018.

Mao, C., Zhong, Z., Yang, J., Vondrick, C., and Ray, B. Metric learning for adversarial robustness. In *NeurIPS*, pp. 478–489, 2019.

Mirman, M., Gehr, T., and Vechev, M. Differentiable abstract interpretation for provably robust neural networks. In *ICML*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *CVPR*, 2019.

Mosbach, M., Andriushchenko, M., Trost, T., Hein, M., and Klakow, D. Logit pairing methods can fool gradient-based attacks. In *NeurIPS 2018 Workshop on Security in Machine Learning*, 2018.

Mustafa, A., Khan, S., Hayat, M., Goecke, R., Shen, J., and Shao, L. Adversarial defense by restricting the hidden space of deep neural networks. In *ICCV*, 2019.

Pang, T., Xu, K., Du, C., Chen, N., and Zhu, J. Improving adversarial robustness via promoting ensemble diversity. In *ICML*, 2019.

Pang, T., Xu, K., Dong, Y., Du, C., Chen, N., and Zhu, J. Rethinking softmax cross-entropy loss for adversarial robustness. In *ICLR*, 2020.

Qin, C., Martens, J., Gowal, S., Krishnan, D., Dvijotham, K., Fawzi, A., De, S., Stanforth, R., and Kohli, P. Adversarial robustness through local linearization. In *NeurIPS*, 2019.

Rice, L., Wong, E., and Kolter, J. Z. Overfitting in adversarially robust deep learning. In *ICML*, 2020.

Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., and Granger, E. Decoupling direction and norm for efficient gradient-based l2 adversarial attacks and defenses. In *CVPR*, 2019.

Shafahi, A., Najibi, M., Ghiasi, M. A., Xu, Z., Dickerson, J., Studer, C., Davis, L. S., Taylor, G., and Goldstein, T. Adversarial training for free! In *NeurIPS*, pp. 3353–3364, 2019.

Song, C., He, K., Wang, L., and Hopcroft, J. E. Improving the generalization of adversarial training with domain adaptation. In *ICLR*, 2019.

Taghanaki, S. A., Abhishek, K., Azizi, S., and Hamarneh, G. A kernelized manifold mapping to diminish the effect of adversarial perturbations. In *CVPR*, 2019.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *ICLR*, 2019.

Wang, B., Shi, Z., and Osher, S. Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies. In *NeurIPS*, 2019.

Wang, J. and Zhang, H. Bilateral adversarial training: Towards fast training of more robust models against adversarial attacks. In *ICCV*, 2019.

Wong, E., Schmidt, F., Metzen, J. H., and Kolter, J. Z. Scaling provable adversarial defenses. In *NeurIPS*, 2018.

Wong, E., Rice, L., and Kolter, J. Z. Fast is better than free: Revisiting adversarial training. In *ICLR*, 2020.

Xiao, C., Zhong, P., and Zheng, C. Enhancing adversarial defense by k-winners-take-all. In *ICLR*, 2020.

Yang, Y., Zhang, G., Katabi, D., and Xu, Z. ME-net: Towards effective adversarial robustness with matrix estimation. In *ICML*, 2019.

Zhang, D., Zhang, T., Lu, Y., Zhu, Z., and Dong, B. You only propagate once: Accelerating adversarial training via maximal principle. In *NeurIPS*, pp. 227–238, 2019a.

Zhang, H. and Wang, J. Defense against adversarial attacks using feature scattering-based adversarial training. In *NeurIPS*, pp. 1829–1839, 2019.

Zhang, H. and Xu, W. Adversarial interpolation training: A simple approach for improving model robustness, 2020. URL https://openreview.net/forum?id=Syejj0NYvr.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019b.

Zhang, H., Chen, H., Xiao, C., Gowal, S., Stanforth, R., Li, B., Boning, D., and Hsieh, C.-J. Towards stable and efficient training of verifiably robust neural networks. In *ICLR*, 2020.