Details on the attack's failure caused by floating-point underflow errors. To analyze the reasons for the failure of the attack caused by floating-point underflow errors, we selected $\mathbb{Z} = \mathbf{z}_y - \max_{i \neq y} \mathbf{z}_i$, which indicates the success or failure of the attack based on its sign. We compared the change of $\mathbb{Z}$ for samples attacked by PGD-100 with CE loss but failed, and PGD-100 with MIFPE loss but succeeded. Figure 1 shows that for samples attacked with CE loss, $\mathbb{Z}$ remains constant throughout the attack. In contrast, for samples attacked with MIFPE loss, $\mathbb{Z}$ smoothly drops below 0 after approaching 0. This phenomenon reveals that floating-point rounding errors cause the calculated gradient to be 0, resulting in a null perturbation added to the sample, which keeps the example constant throughout the iteration and causes the attack to fail.

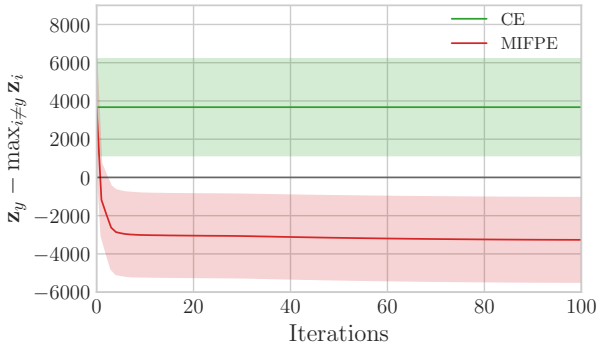


Figure 1. The changing process of the value of $\mathbb{Z} = \mathbf{z}_y - \max_{i \neq y} \mathbf{z}_i$ with the number of iterations during the attack on the CIFAR10 dataset using the model from Neural level sets [1] and single-precision arithmetic. The horizontal axes show the number of iterations used so far, and the vertical axes show the value of $\mathbb{Z}$.

Indirectly controls the values of $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$ can also reduce the impact of floating point errors. We know that MIFPE controls $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$, but this is just one of the combinations in $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi i}, i \in \{2, 3, ..., K\}$. Therefore, we asked what would happen if we used $i \in \{3, ..., K\}$ instead of $i = 2$ in MIFPE. To answer this, we designed an experiment where we tested the model's robustness using different combinations of $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi i}, i \in \{2, 3, ..., K\}$ in MIFPE. Figure 2 shows that all combinations reduce the overestimation of model robustness caused by floating-point errors to varying degrees, but the optimal result is achieved for $i = 2$. This is because controlling $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi i}, i \in \{3, ..., K\}$ values also indirectly controls the values of $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$.

Floating-point rounding errors account for the majority of overestimation of robustness. To understand the distribution of $\Delta = \mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$ on models with floating-point errors leading to overestimation of robustness and to analyze whether floating-point rounding errors or floating-point

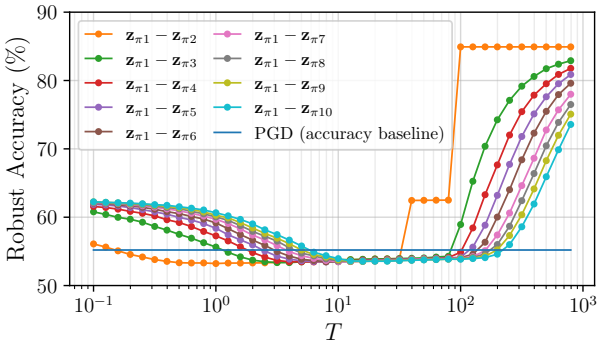
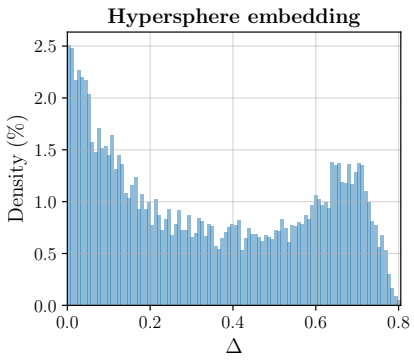


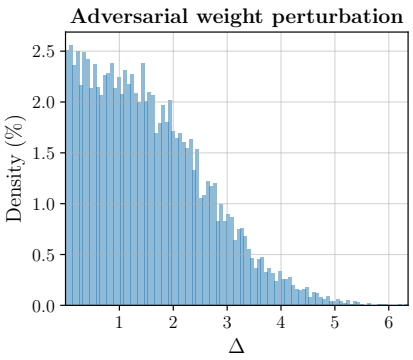
Figure 2. We assess model robustness among different combination for $\mathbf{z}_{\pi 1} - \mathbf{z}_{\pi i}$, $(i \in 2, 3, ..., K)$ using 100 iterations of FT-PGD with CE loss on the CIFAR-10 dataset under half-precision floating-point arithmetic. The model is obtained from [15].

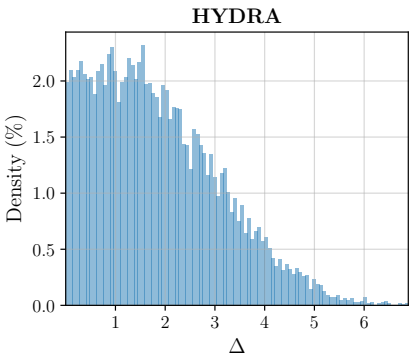
downflow errors account for the majority of floating-point errors, we plotted the distribution of Δ on twelve models from [1–5, 8–13, 15], respectively. To further understand how the T in the MIFPE loss function impacts the attack effectiveness, we varied T from 10^{-1} to 10^{3} on each model. Figure 3 illustrates the $\Delta = \mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$ distribution and robust accuracy with different T for the models.

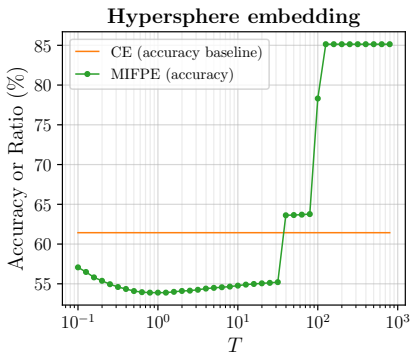
We found that the range of Δ varies dramatically for the different defence models. Among them, none of the Δ in Figure 3 [(a)-(k)] exceeds 20, which is much smaller than the $\lambda \approx 103.28$ for the single-precision floating-point arithmetic, and the floating-point rounding errors are the main reason for overestimating model robustness under single-precision floating-point arithmetic. While only in Figure 3 (l) most of the Δ exceeds the $\lambda \approx 103.28$ for the single-precision floating-point arithmetic, the floating-point underflow errors are the main reason for overestimating the model's robustness under single-precision floating-point arithmetic. In summary, Floating-point rounding errors are the main reason for most of the overestimation of model robustness caused by floating-point errors. In contrast, the overestimation of model robustness caused by floating-point underflow errors is severe but rarely occurs. We found that the best T values for all models are usually close to 1. So we used the factor $T = 1$ in all our experiments.

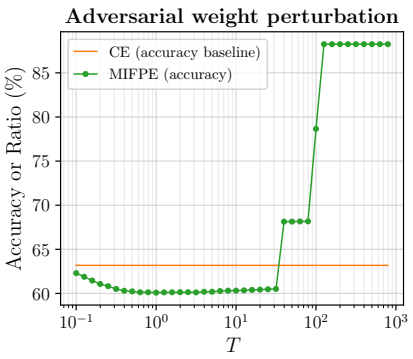
When floating-point errors are not the primary cause of overestimation. We evaluated the performance of MIFPE on rare models that suffer from overestimation due to a typical gradient masking problem that the flat loss surface in the input space, rather than floating-point errors. The results, presented in Figure Figure 4, demonstrate that MIFPE can significantly reduce the problem of overestimation of model robustness by adjusting the T value, even when floating-point errors are not the primary cause of overestimation.

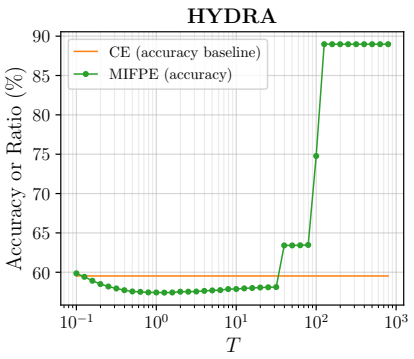




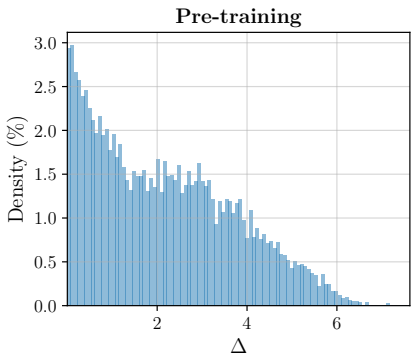


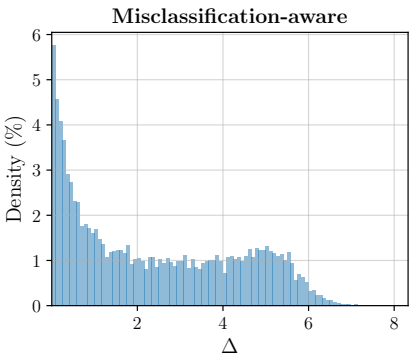


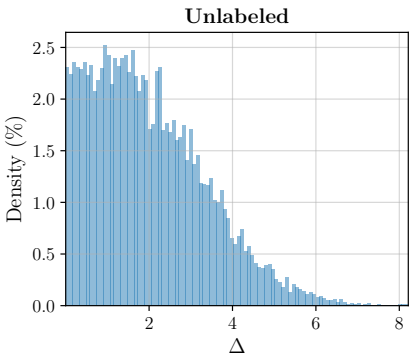


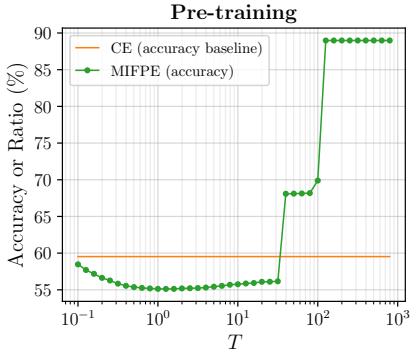


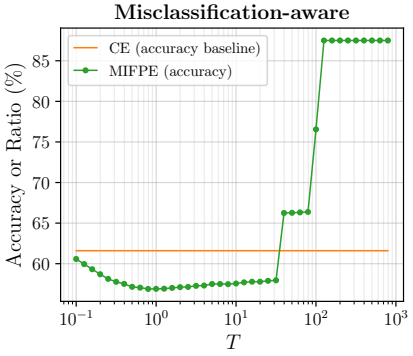
(a) [8] (b) [13] (c) [10]

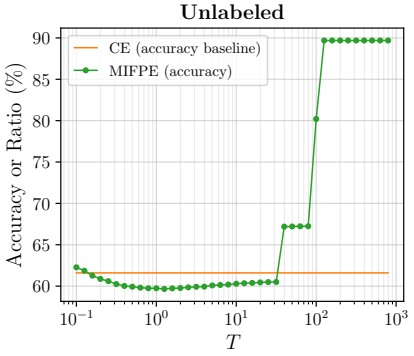




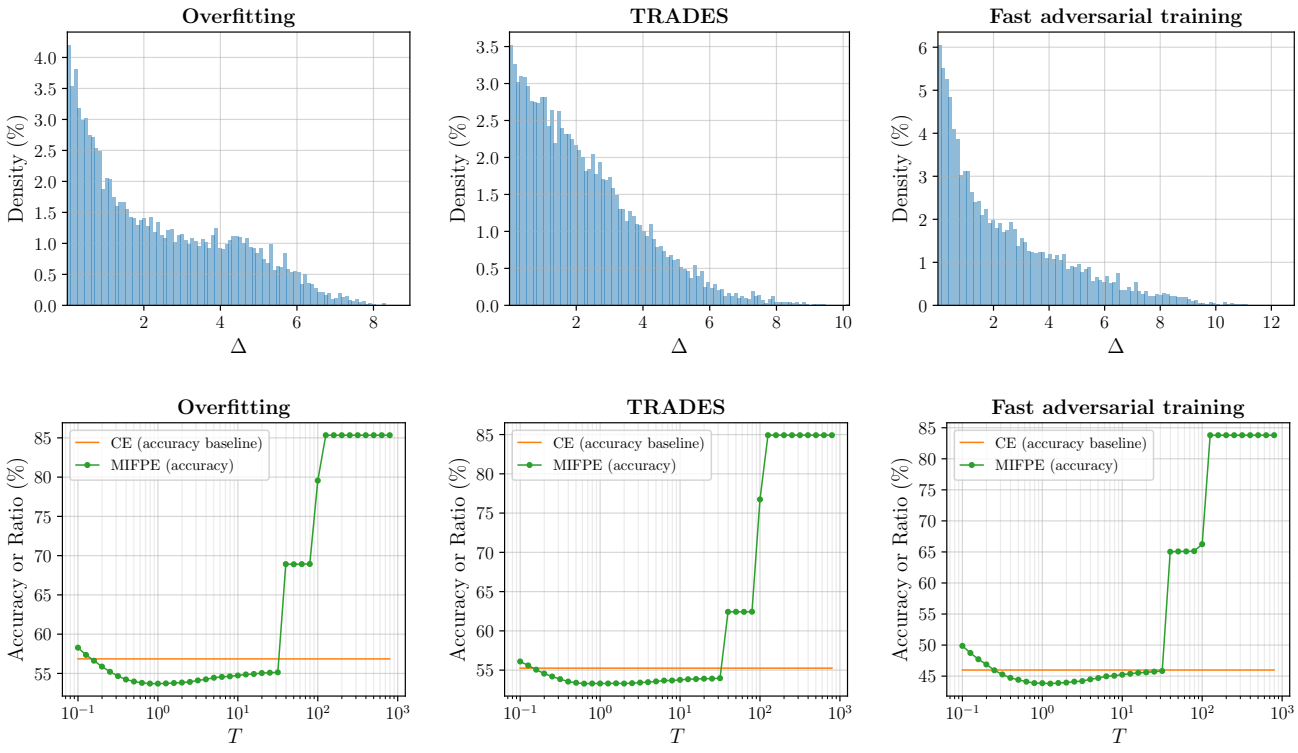




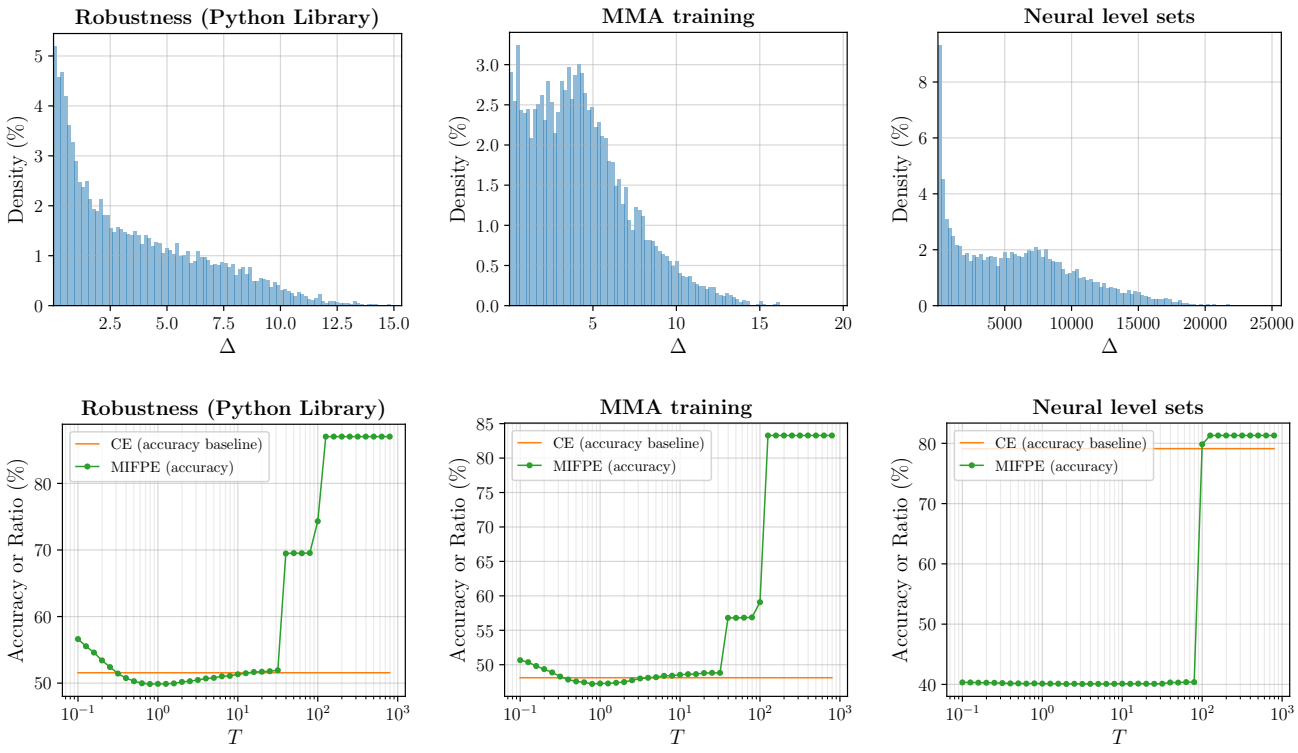




(d) [5] (e) [11] (f) [2]

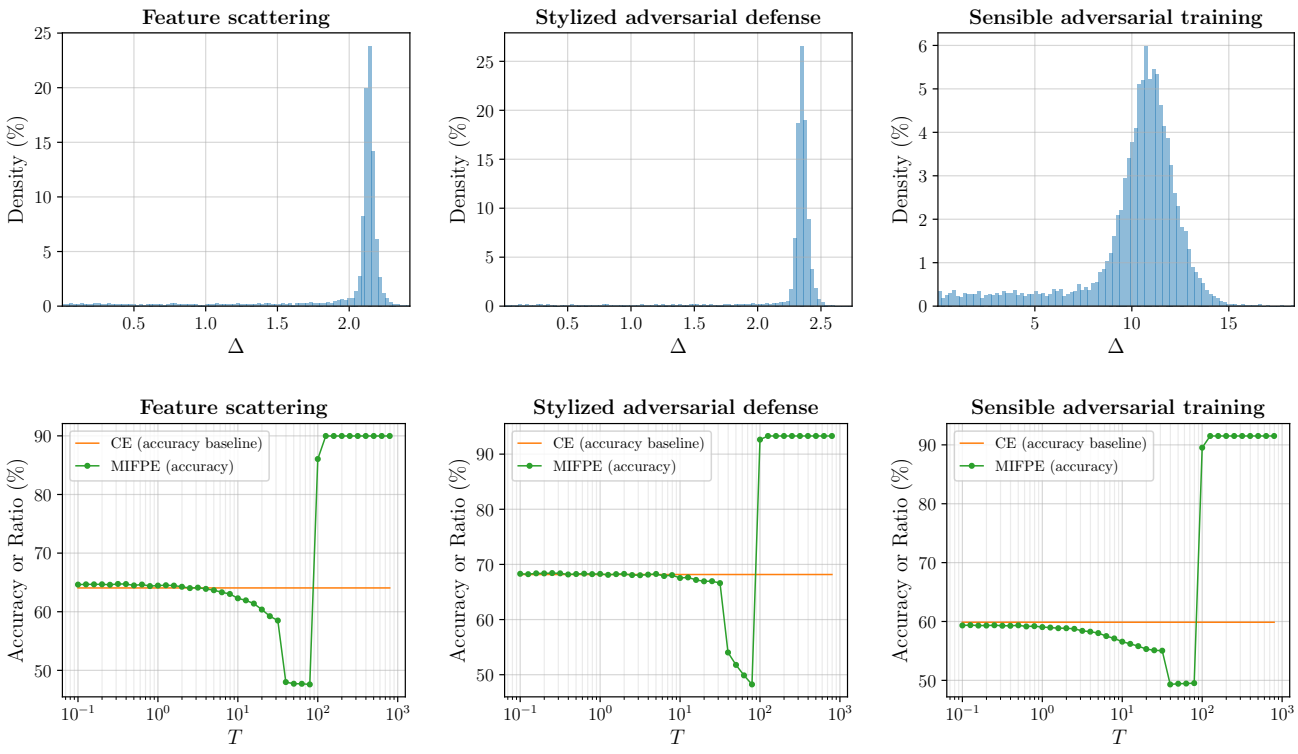


(g) [9] (h) [15] (i) [12]



(j) [4] (k) [3] (l) [1]

Figure 3. The $\Delta = \mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$ distribution (top) and robust accuracy with different T(bottom) for the models of [1–5, 8–13, 15]. The distribution is averaged over 100 bins. The model's robustness is tested under single-precision floating-point arithmetic using PGD with 100 iterations and the CE loss and MIFPE loss, respectively.



(a) [14] (b) [7] (c) [6]

Figure 4. The $\Delta = \mathbf{z}_{\pi 1} - \mathbf{z}_{\pi 2}$ distribution (top) and robust accuracy with different T(bottom) for the models of [6, 7, 14]. The distribution is averaged over 100 bins. The model's robustness is tested under single-precision floating-point arithmetic using PGD with 100 iterations and the CE loss and MIFPE loss, respectively.

References

[1] Matan Atzmon, Niv Haim, Lior Yariv, Ofer Israelov, Haggai Maron, and Yaron Lipman. Controlling neural level sets. In *Advances in Neural Information Processing Systems*, pages 2032–2041, 2019. 1, 3

[2] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems*, volume 32, pages 11192–11203, 2019. 1, 2, 3

[3] Gavin Weiguang Ding, Yash Sharma, Kry Yik Chau Lui, and Ruitong Huang. Mma training: Direct input space margin maximization through adversarial training. In *International Conference on Learning Representations*, 2020. 1, 3

[4] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. Available at: https://github.com/MadryLab/robustness. 1, 3

[5] Dan Hendrycks, Kimin Lee, and Mantas Mazeika. Using pre-training can improve model robustness and uncertainty. In *International Conference on Machine Learning*, pages 2712–2721. PMLR, 2019. 1, 2, 3

[6] Jungeum Kim and Xiao Wang. Sensible adversarial learning. *OpenReview*, 2020. Available at: https://openreview.net/forum?id=rJlf_RVKwr. 4

[7] Muzammal Naseer, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Fatih Porikli. Stylized adversarial defense. *arXiv 2007.14672*, 2020. 4

[8] Tianyu Pang, Xiao Yang, Yinpeng Dong, Kun Xu, Hang Su, and Jun Zhu. Boosting adversarial training with hypersphere embedding. In *NeurIPS*, 2020. 1, 2, 3

[9] Leslie Rice, Eric Wong, and J. Zico Kolter. Overfitting in adversarially robust deep learning. In *ICML*, 2020. 1, 3

[10] Vikash Sehwag, Shiqi Wang, Prateek Mittal, and Suman Jana. Hydra: Pruning adversarially robust neural networks. *arXiv preprint arXiv:2002.10509*, 2020. 1, 2, 3

[11] 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*, 2020. 1, 2, 3

[12] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 1, 3

[13] Boxi Wu, Jinghui Chen, Deng Cai, Xiaofei He, and Quanquan Gu. Do wider neural networks really help adversarial robustness? In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021. 1, 2, 3

[14] Haichao Zhang and Jianyu Wang. Defense against adversarial attacks using feature scattering-based adversarial training. In *Advances in Neural Information Processing Systems*, volume 32, pages 1831–1841, 2019. 4

[15] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *ICML*, 2019. 1, 3