Your Classifier Can Do More: Towards Balancing the Gaps in Classification, Robustness, and Generation

Supplementary Material

Table 1. Robustness (%) of EB-JDAT-JEM++ on CIFAR-10.

Model	APGD-CE	APGD-DLR	FAB	SQUARE
Ours	64.46	65.29	90.52	70.72

Table 2. Comparative robustness (%) of EB-JDAT-JEM++ on CIFAR-10 under transfer attacks from a PGD-AT surrogate.

Model	PGD	MI-FGSM	VMI-FGSM	VNI-FGSM
TRADES	22.57	21.68	21.54	21.97
Ours	**39.06**	**38.33**	**37.72**	**39.19**

Table 3. Robustness (%) of EB-JDAT-JEM++ on CIFAR-10 under APGD-CE (varying distortion/iterations) and EOT-PGD.

	APGD-CE: distortion bound				APGD-CE: iterations				EOT-PGD: distortion bound			
Model	4/255	8/255	16/255	32/255	1	10	20	50	4/255	8/255	16/255	32/255
Ours	80.28	64.46	30.63	2.29	68.43	64.97	64.46	64.32	80.97	64.68	34.24	7.11

1. Gradient obfuscation analysis

Tab. 1 reports a per-attack AA breakdown, without using EOT, Tab. 2 evaluates transfer attacks using PGD-AT as surrogate model, where EB-JDAT-JEM++ consistently outperforms a strong AT baseline. Following [1], we check three masking indicators: (i) robustness not decreasing with larger distortion bound, (ii) one-step attacks outperforming iterative ones, and (iii) robustness relying on stochastic gradients. Tab. 3 shows that robustness decreases with larger distortion bound, iterative attacks outperform one-step, and EOT-PGD is at least as strong as standard PGD, jointly contradicting gradient-masking indicators.

2. Initialization for generation

For fairness, we use the default initialization of each JEM variant, as mismatched training and evaluation initializations may lead to generation failure. JEM++/SADAJEM use informative init, while JEM (FID 38.40) uses random init, under which EB-JDAT-JEM achieves competitive generation (FID 39.43).

3. Limitation And Discussion

Training EB-JDAT on complex, high-dimensional data remains challenging. This challenge is also encountered by JEMs,including JEM [2], JEM++ [3] and SADAJEM [4]. This instability arises from the typically sharp probability distribution of real data in high-dimensional space, which leads to inaccurate guidance for image sampling in regions with low data density. Although the approach we propose is a general and flexible optimization framework for all JEMs, considering training stability, we recommend training within faster and more stable JEM variants, such as JEM++ and SADAJEM. Nevertheless, our method significantly enhances the robustness of JEMs, surpassing SOTA AT [5], while incurring only a slight degradation in accuracy and generative performance, thereby achieving the **best overall trade-off** among robustness (68.76%) , accuracy (90.39%) , and generation (FID=27.42) .

Scaling EB-JDAT to large-scale datasets remains challenging due to the high computational cost of adversarial training and the additional overhead of EBM sampling, we will clarify this limitation and explore latent-space optimization on larger datasets in future work.

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] Will Grathwohl, Kuan-Chieh Wang, Joern-Henrik Jacobsen, David Duvenaud, Mohammad Norouzi, and Kevin Swersky. Your classifier is secretly an energy based model and you should treat it like one. In *International Conference on Learning Representations(ICLR)*, 2020. 1

[3] Xiulong Yang and Shihao Ji. Jem++: Improved techniques for training jem. In *International Conference on Computer Vision(ICCV)*, pages 6494–6503, 2021. 1

[4] Xiulong Yang, Qing Su, and Shihao Ji. Towards bridging the performance gaps of joint energy-based models. In *Conference on Computer Vision and Pattern Recognition(CVPR)*, pages 15732–15741, 2023. 1

[5] Kejia Zhang, Juanjuan Weng, Shaozi Li, and Zhiming Luo. Towards adversarial robustness via debiased high-confidence logit alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2783–2792, 2025. 1