Robust Representation Consistency Model

ICLR 2025

By Jiachen Lei

Outline

- **Background**
  - What's Certified Robustness?
  - What're the problems with Certified Robustness methods?
- **Preliminaries**
- **Method**
  - A high-level perspective: How do we address the problems?
  - Details: Contrastive Denoising Pre-training & fine-tuning
- **Experiment**
  - SOTA performance & Fast Inference Speed
  - Scalable on ImageNet: More Training budget, better performance
- **Conclusion**

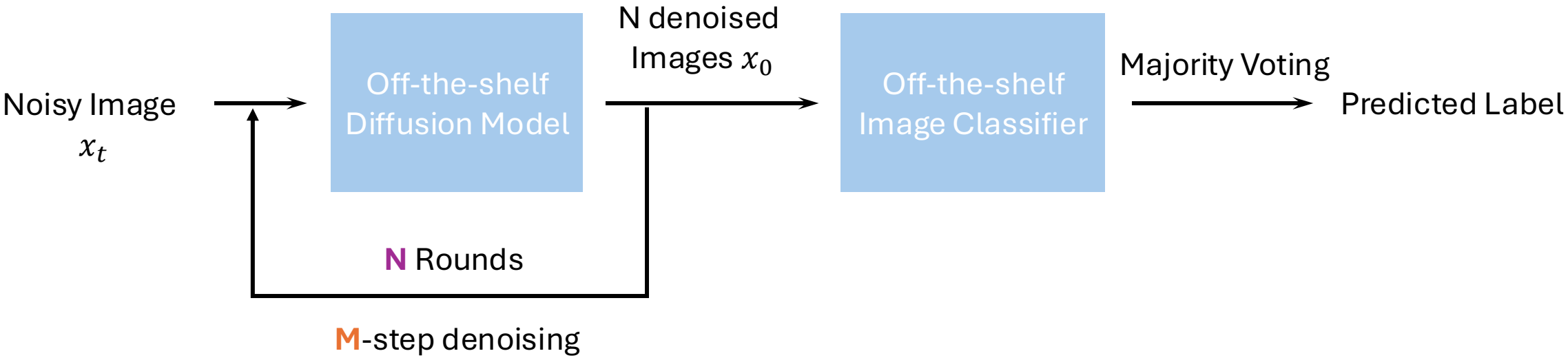
Background

- ***What is Certified Robustness? A narrow explanation***
- Focus on the robustness of DNN against perturbations (e.g. Gaussian perturbations)
- The model robustness is certified by utilizing Randomized Smoothing[6]

- ***What're the problems in the Certified Robustness area?***
- Existing methods: **Diffusion-based methods** and **Traditional methods**
- Problems with theses methods:
  - Diffusion-based methods (e.g. DiffSmooth[1], DensePure[2])
    - **Strong** performance + **Slow** inference speed
  - Traditional methods (e.g., Consistency[3], SmoothAdv[4]):
    - **Relatively Poor** performance + **High** inference speed

Preliminaries: Diffusion-based methods

- DensePure[2]: Denoise perturbed images **N** rounds and predict label by majority voting.
  - Off-the-shelf diffusion model and image classifier
  - Multi-step denoising & Majority Voting
- Inference Complexity on a single image: O(**N*****M**)



Preliminaries: Traditional methods

- Consistency[3]: align predictions output by the classifier

$$\boldsymbol{CE}(y, NN(x_t)_\theta) + \eta_1 \boldsymbol{CE}(NN(x_t^1)_\theta, NN(x_t^2)_\theta) + \eta_2 \boldsymbol{Entropy}(NN(x_t)_\theta),$$

$$with\ \boldsymbol{Entropy}(a) = -p(a)logp(a), \boldsymbol{CE}(a,b) = (1-a)logp(1-b) + alogp(b)$$

Here, y is ground truth label, $NN(\cdot)_\theta$ is the neural network, $x_t = (x_t^1, x_t^2)$, x_t^1 and x_t^2 are two noisy versions of the same input x.

- Inference Complexity on a single image: O(1)

Trained with Heuristic Training Method

Method: A High-level explanation

How should we close the gap of diffusion-based methods in terms of the tradeoff between performance and efficiency?

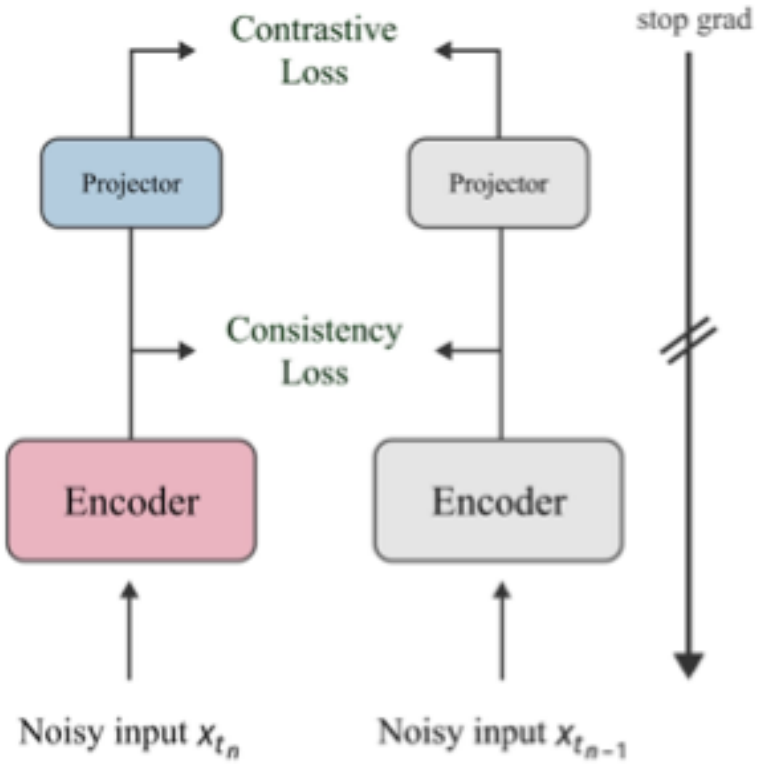
- **Robustness:** model predictions are **consistent** across clean and perturbed images.
- **Motivation:** Probability Flow Ordinary Differential Equation (PF ODE)[5]
  - It defines the **deterministic reverse sampling process** of a diffusion model.
  - It ensures that sampling trajectories remain **distinct and do not cross each other**

As a result:

- Any point uniquely belongs to a single sampling trajectory.
- Points on the same PF ODE sampling trajectory correspond to the same initial clean image
- Therefore, we propose to **align representation of points on the same PF ODE sampling trajectory**.

Method: Pre-training

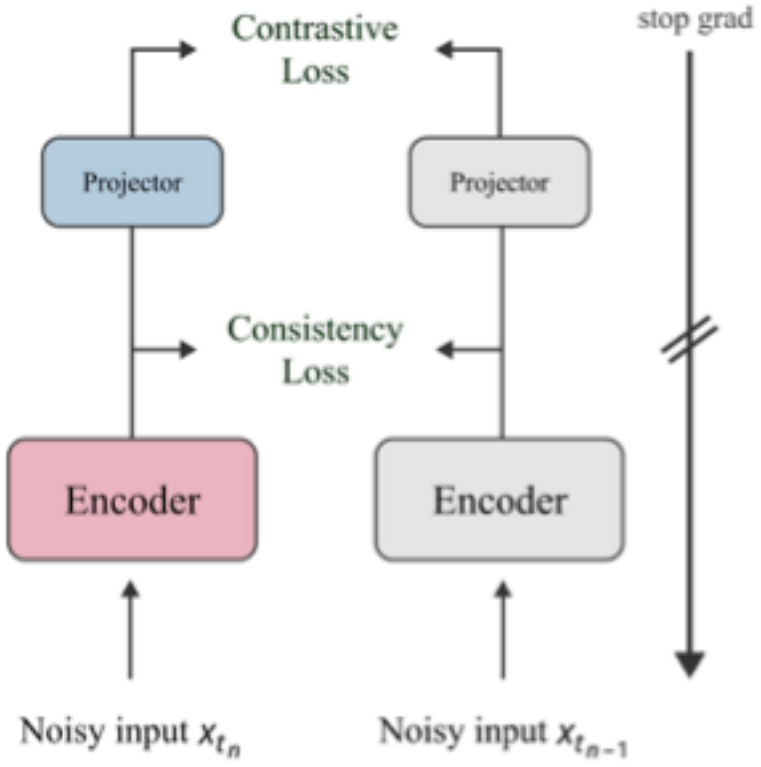
- Decompose training into: Pre-training & fine-tuning
- During Pre-training:
  - Training objective (both in the form of infoNCE loss): contrastive loss + consistency loss
  - Learn meaningful representations
  - Align representation along PF ODE trajectories
    - Positive pair: temporally adjacent points on the same PF ODE traj
    - Negative pair: points on different PF ODE trajctories



Method: Fine-tuning

- During Fine-tuning:
  - Remove the *projector*
  - Load pre-trained encoder weights
  - Training objective (adopt from Consistency[3]):

$$\boldsymbol{CE}(y, NN(x_t)_\theta) + \eta_1 \boldsymbol{CE}(NN(x_t^1)_\theta, NN(x_t^2)_\theta) + \eta_2 \boldsymbol{Entrop}$$



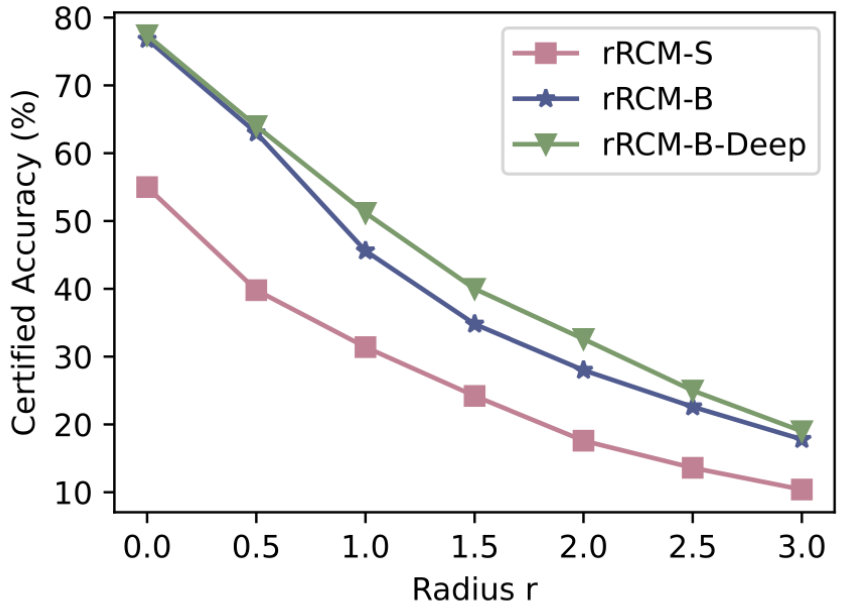
Experiment: Performance on ImageNet

Method	Latency[1]	Certified Accuracy at r (%) 0.0	0.5	1.0	1.5	2.0	2.5
Gaussian (Salman et al., 2019a)	1min 20s	67.0	49.0	37.0	29.0	19.0	15.0
Consistency (Jeong & Shin, 2020)	1min 20s	55.0	50.0	44.0	34.0	24.0	21.0
SmoothAdv (Salman et al., 2019a)	1min 20s	67.0	56.0	43.0	37.0	27.0	**25.0**
Boosting (Horváth et al., 2021)	4min	65.6	57.0	44.6	38.4	28.6	24.6
MACER (Zhai et al., 2020)	1min 20s	68.0	57.0	43.0	31.0	25.0	18.0
SmoothMix (Jeong et al., 2021)[2]	1min 20s	55.0	50.0	43.0	38.0	26.0	24.0
Denoised (Salman et al., 2020)	-	60.0	33.0	14.0	6.0	-	-
DDS‡ (Carlini et al., 2022)	3min 52s	76.2	61.0	41.4	28.0	21.2	17.2
DensePure‡ (Xiao et al., 2022) K=1	17min 8s	76.6	57.0	38.0	22.2	17.0	13.2
K=5	52min 20s	77.8	64.6	38.4	23.0	18.4	14.0
DiffSmooth‡ (Zhang et al., 2023) m = 5	4min 41s	70.1	59.7	34.7	24.8	18.0	13.8
m = 10	5min 10s	70.0	61.4	36.0	26.4	20.8	18.0
m = 15	5min 35s	69.8	62.2	36.4	28.2	21.6	19.2
rRCM-B‡	6s	76.6	62.6	45.2	33.8	27.0	22.0
rRCM-B	53s[4]	76.8	63.0	45.6	34.8	28.0	22.6
rRCM-B-Deep	1min 41s	**77.4**	**64.0**	**51.2**	**40.0**	**32.6**	**25.0**

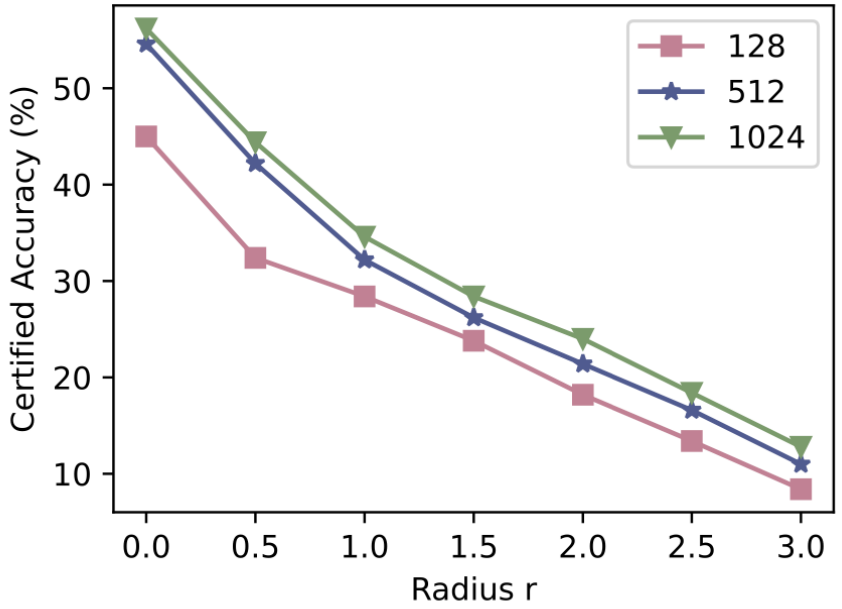
Experiment: Performance on CIFAR10

Method	Latency[1]	Certified Accuracy at r (%)				
		0.0	0.25	0.5	0.75	1.0
Gaussian (Cohen et al., 2019)	4s	83.0	61.0	43.0	32.0	22.0
Consistency (Jeong & Shin, 2020)	4s	77.8	68.8	58.1	48.5	37.8
SmoothAdv (Salman et al., 2019a)	4s	82.0	68.0	54.0	41.0	32.0
Boosting (Horváth et al., 2021)	40s	83.4	70.6	60.4	**52.4**	38.8
MACER (Zhai et al., 2020)	4s	81.0	71.0	59.0	46.0	38.0
SmoothMix (Jeong et al., 2021)	4s	77.1	67.9	57.9	47.7	37.2
DDS Carlini et al. (2022)	52s	79.8	69.9	55.0	47.6	37.4
DiffSmooth Zhang et al. (2023)	3min 34s	78.2	67.2	59.2	47.0	37.4
rRCM-B	16s	**83.6**	**73.4**	**61.4**	48.0	**39.2**

Experiment: Scalability



Scale up **model size** improves performance.



Increase **training batch size** improves performance.

Conclusion

- Establishes a unique connection between perturbed and clean samples along the trajectories of the probability flow (PF) of the denoising process.
- We close the gap of diffusion-based methods in terms of the tradeoff between performance and efficiency
- The method exhibits strong scalability on ImageNet

References

- [1] Zhang, Jiawei, et al. "DiffSmooth: Certifiably robust learning via diffusion models and local smoothing." *32nd USENIX Security Symposium (USENIX Security 23)*. 2023.
- [2] Xiao, Chaowei, et al. "Densepure: Understanding diffusion models for adversarial robustness." *The Eleventh International Conference on Learning Representations*. 2023.
- [3] Salman, Hadi, et al. "Provably robust deep learning via adversarially trained smoothed classifiers." *Advances in neural information processing systems* 32 (2019).
- [4] Jeong, Jongheon, and Jinwoo Shin. "Consistency regularization for certified robustness of smoothed classifiers." *Advances in Neural Information Processing Systems* 33 (2020): 10558-10570.
- [5] Song, Yang, et al. "Score-based generative modeling through stochastic differential equations." *arXiv preprint arXiv:2011.13456* (2020).
- [6] Cohen, Jeremy, Elan Rosenfeld, and Zico Kolter. "Certified adversarial robustness via randomized smoothing." *international conference on machine learning*. PMLR, 2019.