Consistency Purification: Effective and Efficient Diffusion Purification towards Certified Robustness

Yiquan Li[1*]; Zhongzhu Chen[2*]; Kun Jin[2*]; Jiongxiao Wang[1*]; Jiachen Lei[3]; Bo Li[4]; Chaowei Xiao[1]

[1]University of Wisconsin-Madison; [2]University of Michigan-Ann Arbor; [3]California Institute of Technology; [4]University of Illinois Urbana-Champaign

Presenter: Jiongxiao Wang

∗The first four authors contributed equally. Correspondence to: Jiongxiao Wang <jwang2929@wisc.edu>.

Background

Randomized Smoothing for Certified Robustness

Certify the robustness of a given classifier under ℓ_2 norm perturbations. Given the base classifier f and an input x, randomized smoothing defines the smoothed classifier by

$$g(x) = \arg max_c \, \mathbb{P}_{\epsilon \sim \mathcal{N}(0,\sigma^2 I)}(f(x+\epsilon) = c)$$

Then $g(x)$ induces the certified robustness for x with radius R by

$$R = \frac{\sigma}{2}(\Phi^{-1}(p_A) - \Phi^{-1}(p_B))$$

Diffusion Purification

For a larger certified radius, robust training of classifier f with Gaussian augmented noise is normally required to ensure the accurate classification of the smoothed classifier. Due to the inherent denoising capability of the diffusion model D, the base classifier f can be defined as the combination of any standard classifier f_{std} and the diffusion model by

$$f(x+\epsilon) := f_{std}(D(x+\epsilon))$$

Motivation

Efficiency and Effectiveness Trade-offs for Previous Diffusion Purification

- One-shot denoising with Denoising Diffusion Probabilistic Model (DDPM)

  Efficient but not effective. One-shot only generates the posterior mean of noisy data.

- DensePure, Local Smoothing, Noised Diffusion Classifier, Probability Flow Ordinary Differential Equation (PF-ODE)

  Effective but not efficient. Multiple network evaluations are required.

Consistency Model, the Potential Pareto Superior Solution

- Consistency model leans the trajectory of PF-ODE by mapping any point along this trajectory back to its start point. It allows images with any scale of Gaussian noise to be directly purified to clean images.
- Consistency models are primarily trained for image generation, further adaptations are needed for better purification performance.

Theoretical Analysis

Why consistency model is both effective and efficient?

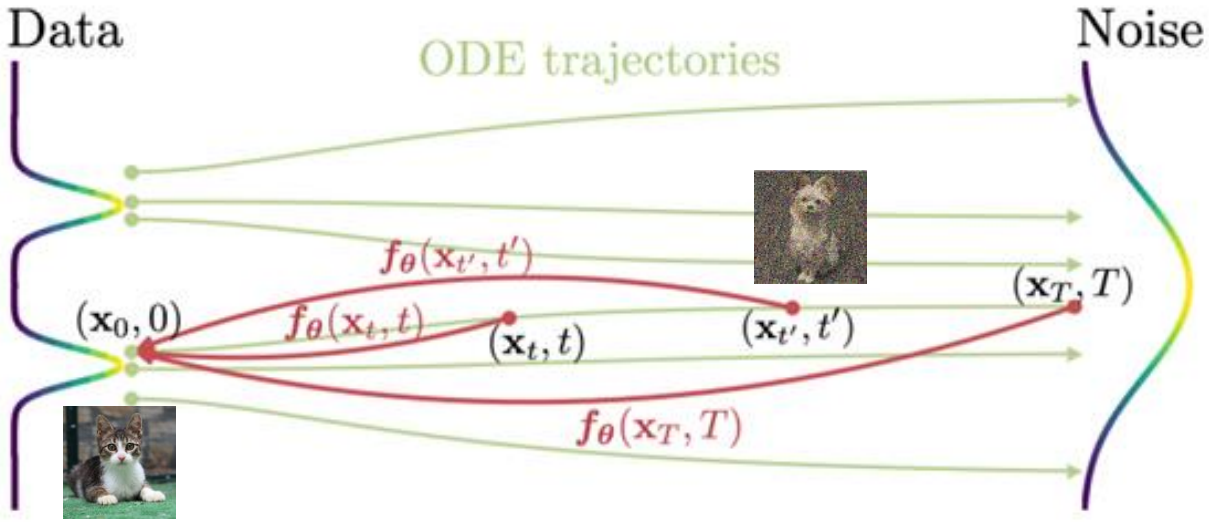
- Effective: in-distribution mapping
- Efficient: one-step purification

Why consistency model only is not enough?

- Semantic inconsistency of ODE trajectory

Consistency models are primarily trained for image generation, which is not enough for purification

The data point in the middle of the trajectory (dog) may not have the same semantic meaning as the data point in the start of the trajectory (cat).



Theoretical Analysis

How to further improve?

- Consistency fine-tuning by minimizing the LPIPS loss between of the clean sample and the purified sample

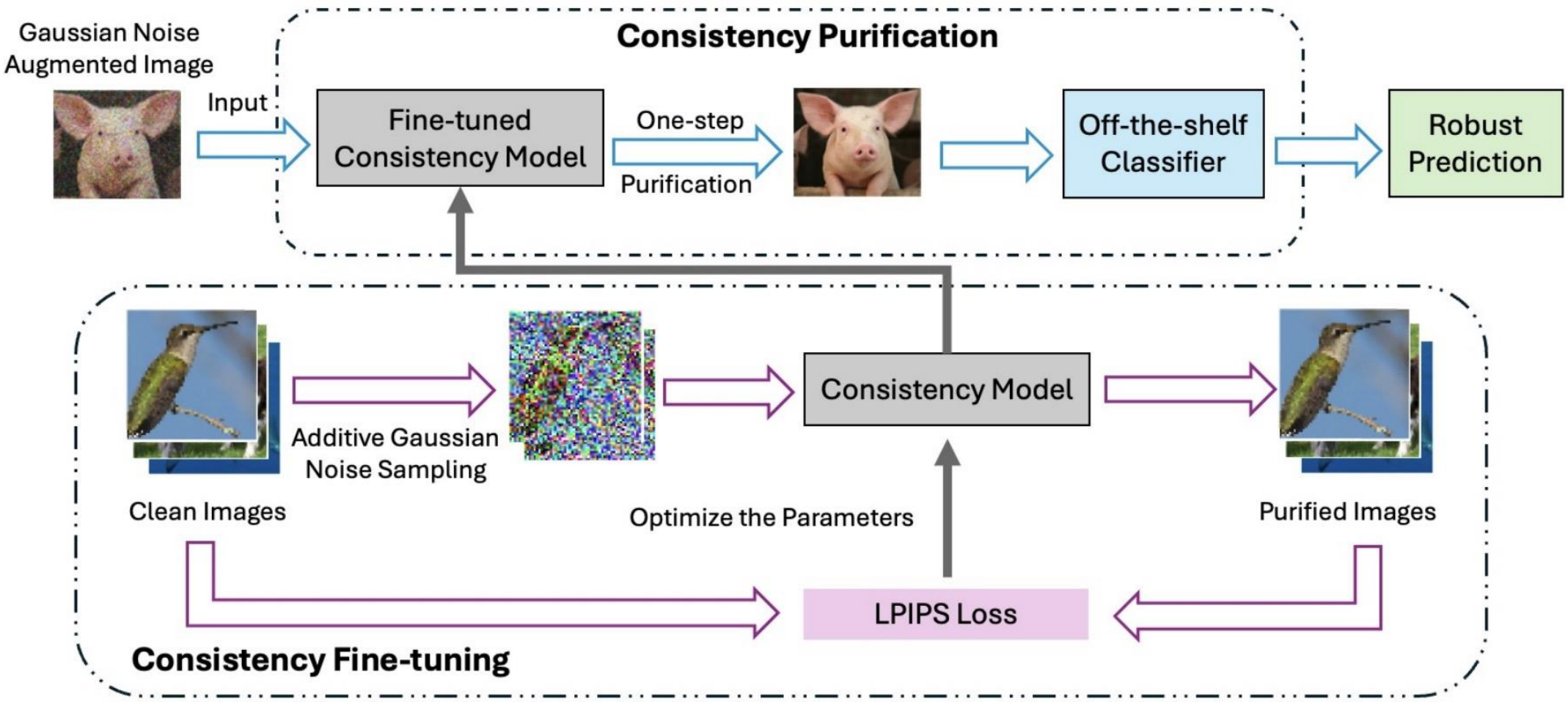
Why not ℓ_1 or ℓ_2 loss?

- ℓ_1 or ℓ_2 loss does not represent semantic consistency
- minimizing ℓ_1 or ℓ_2 loss is essentially training DDPM, destroying the trajectory mapping

Theoretical guarantee:

- **Sketch of Theorem 3.3 in the paper:** the purification efficacy is higher when the distance between the data distribution and the purified distribution is lower.
- Consistency fine-tuning aims to decrease the distance between the two distributions

Our Approach: Consistency Purification



Overview of Consistency Purification Framework

Our Approach: Consistency Purification

Consistency Purification

For given consistency model purifier D_θ, any noisy input $x_t \sim \mathcal{N}(x, t^2 I)$ can be recovered to the trajectory's start by $x_\epsilon = D_\theta(x_t, t)$.

For the Gaussian noise augmented image $x_{rs} \sim \mathcal{N}(x, \sigma^2 I)$ with variance σ used in randomized smoothing, we need to compute the time step by $t_\sigma^* = \{t_i | \sigma \in (\frac{t_{i-1}+t_i}{2}, \frac{t_i+t_{i+1}}{2}]\}$, where t_i is the time schedule used during training with $t_i = \left(\epsilon^{\frac{1}{\rho}} + \frac{i-1}{N-1}\left(T^{\frac{1}{\rho}} - \epsilon^{\frac{1}{\rho}}\right)\right)^\rho$, $\rho = 7$.

Consistency Fine-tuning

We fine-tune the purifier D_θ by minimizing the following loss function:

$$\mathcal{L}_\theta = \mathbb{E}\left|\left|x - D_\theta(x_\sigma, t_\sigma^*)\right|\right|_{LPIPS}$$

where the expectation is taken with $x \sim p_{data}$, $\sigma \sim \mathcal{U}\{\sigma_i\}_{i=1}^m$, $x_\sigma \sim \mathcal{N}(x, \sigma^2 I)$.

The fine-tuned consistency model purifier D_{θ^*} results in the final purified image by:

$$x_p = D_{\theta^*}(x_{rs}, t_\sigma^*)$$

Experimental Results

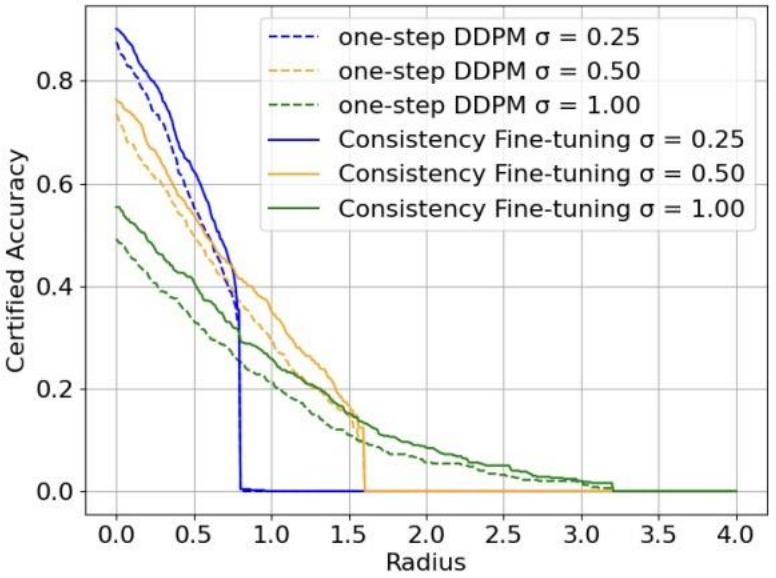
Table 2: Certified Accuracy of Consistency Purification for CIFAR-10 and ImageNet-64.

CIFAR-10		Certified Accuracy at ϵ (%)				
Method	Purification Steps	0.0	0.25	0.5	0.75	1.0
onestep-DDPM[25]	One Step	87.6	73.6	55.6	39.2	29.6
onestep-EDM	One Step	87.4	76.2	58.8	40.8	32.4
PF-ODE EDM	Multi Steps	89.6	77.0	60.4	42.6	34.0
Diffusion Calibration[38]	One Step	90.2	76.4	57.2	42.6	32.4
Consistency Purification	One Step	**90.4**	77.2	59.8	42.8	33.2
+ Consistency Fine-tuning	One Step	90.2	**79.4**	**62.4**	**43.8**	**35.4**

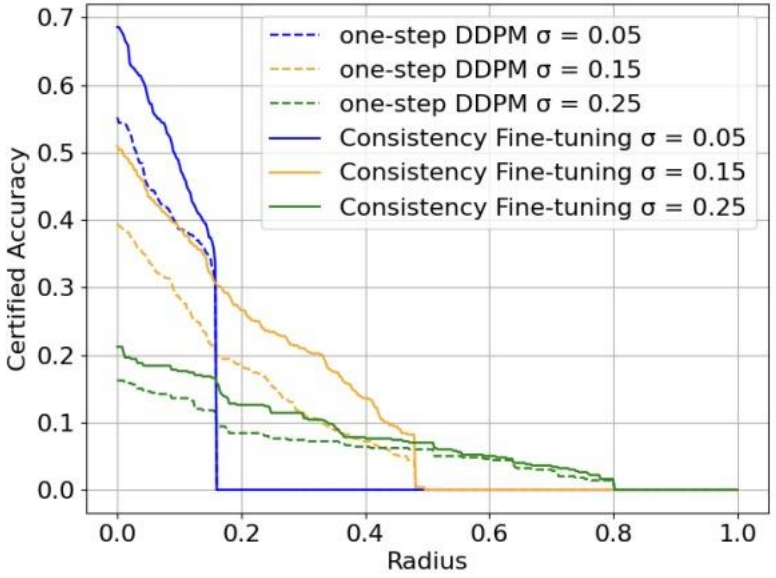
ImageNet-64		Certified Accuracy at ϵ (%)				
Method	Purification Steps	0.0	0.05	0.15	0.25	0.35
onestep-DDPM [25]	One Step	55.2	44.8	33.4	15.2	8.8
Consistency Purification	One Step	62.4	54.2	35.2	19.8	13.0
+ Consistency Fine-tuning	One Step	**68.6**	**58.0**	**37.4**	**23.4**	**17.4**

Experimental results have demonstrated that Consistency Purification could certify robustness with both efficiency and effectiveness compared with various baseline purification methods.

CIFAR-10



ImageNet-64



Visualization Examples

Visualization of purified images* after the diffusion purification by applying onestep-DDPM and Consistency Purification on CIFAR-10 with $\sigma = 0.5$ noise level.

(a) Purified images by onestep-DDPM

(b) Purified images by Consistency Purification

Visualization of purified images* after the diffusion purification by applying onestep-DDPM and Consistency Purification on ImageNet-64 with $\sigma = 0.25$ noise level.

(a) Purified images by onestep-DDPM

(b) Purified images by Consistency Purification

*Identical noise patterns are applied to images at corresponding locations.

Thank You for Listening!