← Back to **Author Console** (/group?id=NeurIPS.cc/2024/Conference/Authors#your-submissions)

Understanding and Enhancing the Robustness of Diffusion-Based Purification

PDF (/pdf?id=9YfV3i5qOZ)

LiuYiming (/profile?id=~LiuYiming1), Kezhao Liu (/profile?id=~Kezhao_Liu1), Yao Xiao (/profile?id=~Yao_Xiao4), ZiYi Dong (/profile?id=~ZiYi_Dong1), Xiaogang Xu (/profile?id=~Xiaogang_Xu2), Pengxu Wei (/profile?id=~Pengxu_Wei1), Liang Lin (/profile?id=~Liang_Lin1)

02 May 2024 (modified: 05 Jun 2024) NeurIPS 2024 Conference Submission Conference, Senior Area Chairs, Area Chairs, Reviewers, Authors Revisions (/revisions?id=9YfV3i5qOZ) CC BY-NC-ND 4.0 (https://creativecommons.org/licenses/by-nc-nd/4.0/)

Keywords: Adversarial Defense, Adversarial Purification, Diffusion Training

TL;DR: We empirically find that stochasticity is crucial for the robustness of diffusion-based purification, and propose a novel training method for diffusion models incorporating adversarial perturbations to enhance the robustness.

Abstract:

Diffusion-Based Purification (DBP) has emerged as an effective defense mechanism against adversarial attacks. Traditionally, the efficacy of DBP has been attributed to the forward diffusion process, which narrows the distribution gap between clean and adversarial images through the addition of Gaussian noise. Although theoretical studies support this explanation, to what extent it contributes to robustness remains unclear. In this paper, we argue that the inherent stochasticity in the DBP process is the primary driver of its robustness. To explore this, we introduce a novel Deterministic White-Box (DW-box) evaluation framework to assess robustness and analyze the attack trajectories and loss landscapes. Our findings suggest that DBP models primarily leverage stochasticity to evade effective attack directions, rather than directly neutralizing adversarial perturbations. To further enhance DBP robustness, we integrate adversarial perturbations into diffusion training, propose Rank-Based Gaussian Mapping (RBGM) to make perturbations more compatible with the diffusion models, and introduce Adversarial Denoising Diffusion Training (ADDT) to strengthen diffusion models with classifier-guided perturbations. Empirical evidence demonstrates the effectiveness of ADDT.

Supplementary Material: zip (/attachment?id=9YfV3i5qOZ&name=supplementary_material)
Corresponding Author: weipx3@mail.sysu.edu.cn
Reviewer Nomination: liuym225@mail2.sysu.edu.cn
Primary Area: Safety in machine learning
Submission Number: 1523

Search keywords... Sort: Newest First

- = ≡

Everyone Program Chairs Submission1523... Submission1523 Area... Submission1523 Authors

17 / 17 replies shown

Submission1523... Submission1523... Submission1523... Submission1523... Submission1523... Submission1523...

✖

Add: **Withdrawal**

Official Comment by Authors

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

14 Aug 2024, 20:00 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:

Dear ACs,

We sincerely appreciate the valuable comments from all reviewers. However, we would like to kindly raise your awareness that Reviewer **NqPk** and Reviewer **ZRj3** might have some misunderstandings about the contribution of this paper.

Specifically, they believe that the **scalability** to larger datasets is necessary for our proposed method ADDT and they tend to reject the paper based on insufficient evidence of the scalability (though we have added experiments on ImageNet-1k that validated the effectiveness of our method). Nonetheless, as argued in the last paragraph of our latest response to Reviewer **NqPk**, the contribution of this paper goes beyond the technical design of ADDT, and proposing a highly scalable algorithm or achieving state-of-the-art performance is not our focus. Instead, the **first two points of our contributions**, given in Lines 60-65, refer to the empirical studies on the root of DBP robustness that highlight stochasticity as the main contributor, which challenges the existing understanding of the mechanism of DBP (Lines 85-91). Besides, our study clarifies the role of stochasticity in DBP in contrast to that in **certified defense methods** like randomized smoothing (please refer to our response to **Q2.1** of Reviewer **ZRj3**). We believe that these new findings and perspectives could have a sustained impact on future research on DBP, which is a promising approach to adversarial defense and could be more valuable for real-world applications compared to AT, although existing studies on DBP are at an early stage.

We hope that our clarification will help you to assess the contribution of this paper.

Rebuttal Discussion

Official Comment Area Chair aodY 10 Aug 2024, 17:12

Program Chairs, Reviewer NqPk, Reviewer ZRj3, Reviewer hZmv, Reviewer qERM, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors

Comment:

Dear Reviewers,

The authors have provided a response to the comments. Please respond to the rebuttal actively.

Best, AC

Author Rebuttal by Authors

Author Rebuttal

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

07 Aug 2024, 02:36 (modified: 07 Aug 2024, 23:22) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=VHPneq3KWY)

Rebuttal:

We sincerely thank all the reviewers for their valuable comments and appreciate their efforts in reviewing our work. Here we would like to clarify two issues commonly raised by **Reviewer NqPk** and **Reviewer hZmv**.

Q1: Evaluation with stronger PGD+EoT attack.

While increasing the iterations of PGD and EoT can provide a more reliable test of adversarial robustness (Appendix A), our main conclusions about the origin of DBP robustness and the comparative analysis of ADDT performance remain unaffected by the choice of PGD and EoT parameters. To balance computational feasibility and attack effectiveness, we chose PGD20+EoT10. As shown in the table below, increasing the number of iterations to **PGD200+EoT20** results in a moderate decrease in accuracy. However, testing DP_DDPM (the smallest model in our test) under this setting would require 50 hours of computation on eight 3090 GPUs, making it impractical to complete all experiments for this paper.

Method	PGD200+EoT20		PGD20+EoT10	
	Vanilla (l_∞)	Ours (l_∞)	Vanilla (l_∞)	Ours (l_∞)
DP_DDPM	41.02%	46.19%	47.27%	51.46%
DP_DDIM	36.23%	41.11%	43.16%	47.85%

We will add this table to the paper and better clarify the choice of attack parameters in the revision.

Q2: Experiments on a larger and higher-resolution dataset.

Evaluating on larger datasets like ImageNet-1k is important but costly. For instance, performing ADDT on ImageNet with ResNet-101 would require about 100 hours on 8*2080ti GPUs, and evaluating on 1024 images of 224×224 resolution with PGD20+EoT10 would take 460 hours. Consequently, many previous works on AT-based methods have also refrained from performing experiments on ImageNet [1, 2, 3, 4, 5, 6].

As a preliminary attempt, we conducted a limited experiment using a small UNet (same as DP_DDPM). This model was trained from scratch for 12 epochs and we performed ADDT for 3 epochs (not fully converged). It was tested under PGD20+EoT10 with (l_∞) 4/255 on the first 1024 images of the ImageNet-1k validation set, yielding the results below.

	Clean Accuracy	Robust Accuracy
Vanilla	80.31%	46.15%
Ours	80.30%	46.92%

We will continue training the model on ImageNet and update the results in the revision if the paper could be accepted.

References:

1. Wang, Zekai, et al. Better diffusion models further improve adversarial training. ICML, 2023.
2. Jin, G., et al. Randomized Adversarial Training via Taylor Expansion. CVPR, 2023.
3. Li, Boqi, and Weiwei Liu. WAT: improve the worst-class robustness in adversarial training. AAAI, 2023.
4. Fabian, Latorre, et al. Finding Actual Descent Directions for Adversarial Training. ICLR, 2023.
5. Nie W, Guo B, Huang Y, et al. Diffusion models for adversarial purification. ICML, 2022.
6. Kang M, Song D, Li B. DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification. NeurIPS, 2023.

PDF: pdf (/attachment?id=VHPneq3KWY&name=pdf)

Updated Experimental Results

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

12 Aug 2024, 16:29 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Dear Reviewers:

We have updated several experimental results that you might be most concerned about. We are sorry for the late update due to the limitation of our computing resources. These results will be integrated into the revision of the paper.

Update 1: Evaluation of DiffPure with stronger PGD+EoT attack.

In Q1 of our initial global response, we provide the results on **DP_DDPM** and **DP_DDIM**, and we have just finished the experiments on **DiffPure** (where the evaluation on PGD200+EoT20 takes ~4 days on 8*4090 GPUs for each model). The results are provided in the table below.

Method	PGD200+EoT20		PGD20+EoT10	
	Vanilla (ℓ_∞)	Ours (ℓ_∞)	Vanilla (ℓ_∞)	Ours (ℓ_∞)
DiffPure	48.93%	55.76%	55.96%	62.11%

In line with our previous results in Q1, our ADDT method consistently improves the robust accuracy of the model under PGD200 + EoT20 and PGD20 + EoT10.

Update 2: Experiments on ImageNet-1k.

In Q2 of our initial global response, we provide the preliminary results of performing ADDT on ImageNet, where the superiority of our method seems not significant due to insufficient training (**3 epochs** only). We have now trained the model for **8 epochs** and achieved an improved robust accuracy of 48.02%, as shown in the table below. In line with Q2, our experiment is conducted on ResNet-101, and the model is tested under PGD20+EoT10 with ℓ_∞ 4/255 bound. Unfortunately, we are unable to test the robust accuracy under PGD200+EoT20 for now, as it takes ~7 days on 8*4090 GPUs for each model. We are still training the model and will update the final results with stronger attacks in the revision.

Method	Clean Accuracy	Robust Accuracy
Adversarial Training [I]	69.52%	41.02%
DP_DDPM	80.31%	46.15%
DP_DDPM (ADDT)	80.20%	48.02%

As kindly suggested by **Reviewer ZRj3**, we also compare our results with those acquired by **adversarial training (AT)** with the same classifier architecture and perturbation constraint, which are reported in [I]. As shown in the table above, the robust accuracy of our model under PGD20+EoT10 is significantly higher than that of AT. We acknowledge that evaluating our model with PGD200+EoT20 may further decrease the robust accuracy as evidenced in Q1 and Update 1, but we speculate that the robust accuracy achieved by our method under PGD200+EoT20 should be **at least comparable** to that of AT, especially given that our training is still not fully converged. Besides, the results above show that our DBP-based method can achieve **significantly better clean accuracy** compared with AT, which is desired for real-world applications. In addition, we notice that recent works may achieve state-of-the-art results with AT [II, III, IV], but their experiments are based on more advanced network architectures, more parameters, or extra training data.

[I] Singh N D, Croce F, Hein M. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. NeurIPS, 2023.
[II] Wang Z, Li X, Zhu H, et al. Revisiting Adversarial Training at Scale. CVPR, 2024.
[III] Liu, Chang, et al. "A comprehensive study on robustness of image classification models: Benchmarking and rethinking." IJCV, 2024.
[IV] Amini S, Teymoorianfard M, Ma S, et al. MeanSparse: Post-Training Robustness Enhancement Through Mean-Centered Feature Sparsification. arXiv, 2024.

Official Review of Submission1523 by Reviewer NqPk

Official Review Reviewer NqPk 12 Jul 2024, 15:00 (modified: 31 Jul 2024, 04:19)
Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer NqPk Revisions (/revisions?id=dS3V7qTiig)

Summary:
This work emphasizes the critical role of stochasticity of diffusion-based purification in defending against adversarial noise. Furthermore, it proposes a method that incorporates adversarial training into diffusion-based purification, named ADDT. Empirical experiments demonstrate the effectiveness of the proposed method.

Soundness: 3: good
Presentation: 3: good
Contribution: 2: fair
Strengths:

1. This paper is well-written, with elaborative article organization and clear illustrations.
2. The analysis of the influence of stochasticity on robustness is somewhat interesting.
3. The enhancement of the robustness of DBP is effective.
4. Code is provided to help the reproducibility.

Weaknesses:

1. One of the main conclusions on the critical role of stochasticity is not that impressive. Many literatures [1, 2] have shown the importance of stochasticity before.
2. Over-estimated robustness on DBP. Both [3] and [1] reported lower adversarial robustness on DBP, which could attributed to the limited PGD and EOT steps used in this work. I recommend using the PGD200+EOT-20, as indicated in [3], to report the main results.
3. Following the above comment, besides experimental results (which may not be that reliable), providing a theoretical guarantee to show the reason why ADDT is better than DiffPure would enhance this work.

4. The robustness of DBP on CIFAR-100 is quite low (Table 3). Can you provide some explanations?
5. The experiments lack results on higher-dimensional images. While the results on the Tiny-ImageNet small image dataset are included, evaluating purification effects on more complex large-scale natural images, especially on the ImageNet-224 dataset, is crucial.

[1] Robust Evaluation of Diffusion-Based Adversarial Purification, ICCV, 2023.

[2] Randomness in ML Defenses Helps Persistent Attackers and Hinders Evaluators, arxiv, 2023.

[3] Robust Classification via a Single Diffusion Model, ICML, 2024

Questions:
Please see the weakness above. I will reconsider my score after the discussion.

Limitations:
N/A

Flag For Ethics Review: No ethics review needed.
Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

-
=
≡

Rebuttal by Authors

Rebuttal
✏ Authors (👁 ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))
📅 07 Aug 2024, 02:59 (modified: 07 Aug 2024, 20:40) 👁 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors
📄 Revisions (/revisions?id=ZDHVgLg8js)

Rebuttal:
W1: The importance of stochasticity has been shown in previous literature.

Thank you for your insightful observations concerning the role of stochasticity in DBP robustness. The studies in references [1] and [2] do acknowledge the influence of stochasticity. However, [1] focuses on better practical evaluation methods for DBP robustness, while our work dives into how stochasticity contributes to the robustness in DBP; [2] argues that *"the majority of the robustness from this defense (DBP) comes from the* ***deterministic*** *de-noising process that uses the pre-trained diffusion model*", while our conclusion that stochasticity is the primary factor for DBP robustness challenges this argument.

Through the novel perspective of this paper, we emphasize that the robustness of DiffPure is not primarily attributable to a smooth landscape or mere purification effects but rather stems from the attacker's inability to pinpoint the specific random state ϵ. This prevents effective localization of local extrema, considerably contributing to the DBP robustness. Such insights also underscore the necessity to keep the randomness elements safe in practical DBP-based defense.

W2: Over-estimated robustness due to limited PGD and EOT steps.

Thanks for your suggestion on reliable robustness evaluation. Please refer to **Q1** of the global response (Author Rebuttal).

W3: Theoretical guarantee for the superiority of ADDT over DiffPure.

Thanks for your suggestion on better justifying our method. As the proof is lengthy, please refer to the global PDF.

W4: The robustness of DBP on CIFAR-100 is quite low.

Thanks for your careful examination of the experimental results. The robust accuracy of a model depends on both the robustness and the accuracy of the model. Since CIFAR-100 has a larger number of classes compared to CIFAR-10, models tend to exhibit significantly lower accuracy for both clean and adversarial samples. As a reference, we listed the results from [I] in the table below. The robust performance of DBP is quite similar to other defense methods.

	CIFAR-10		**CIFAR-100**	
Method	NAT	PGD20	NAT	PGD20
NT	**93.04**	0.00	**65.74**	0.02
AT (β = 1)	84.32	48.29	60.10	28.22
AT (β = 1/2)	87.84	44.51	60.84	22.64
TRADES (λ = 6)	83.91	**54.25**	59.93	**29.90**
TRADES (λ = 1)	87.88	45.58	60.18	28.93
FAT	87.72	46.69	61.71	22.93
IAT	84.60	40.83	57.04	21.40
RST	84.71	44.23	60.30	23.56
Generalist	**89.09**	**50.01**	**62.97**	**29.48**

[I] Wang, et al. Generalist: Decoupling natural and robust generalization, CVPR, 2023

W5: Lack of results on higher-dimensional images.

Thanks for your valuable suggestion. Please refer to **Q2** of the global response (Author Rebuttal).

Official Comment by Authors

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

13 Aug 2024, 15:55 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer NqPk

Comment:
Dear Reviewer NqPk,

Thank you again for your constructive comments, and we would like to kindly remind you that we have posted the **updated experimental results** in our global response that are related to W2 and W5 of your comments, which may further address your concerns. We would like to know if you have any additional concerns or questions, and we are looking forward to your reply.

Replying to Rebuttal by Authors

Official Comment by Reviewer NqPk

Official Comment Reviewer NqPk 14 Aug 2024, 10:55

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thank you for the rebuttal. My concerns still exist.

Q2. From the additional results you provide and Appendix A, the robustness is indeed over-estimated and EOT-20+PGD200 is a good choice to avoid such cases. This has aroused my further concerns about the main results of this work. Additionally, the high cost of rigorous evaluation is caused by the extremely high inference cost of this work. It could hinder the practical usage of the method.

Q4. The result on CIFAR-10 (51.46%, Table 3) is comparable to the result (50.01%) that you refer to, while on CIFAR-100, this work only obtained 23.73% accuracy, which is significantly lower than the 29.48% that you refer to. This raises concerns about the scalability of harder datasets.

Q5. The improvements become marginal on ImageNet, which validates my concern in Q4 again.

Based on these concerns, together with the reviews from other reviewers, I tend to keep my rating.

Replying to Official Comment by Reviewer NqPk

Official Comment by Authors

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

14 Aug 2024, 18:42 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:
Thank you for your valuable comments, and we would like to further clarify some points regarding your concerns.

Q2.1. Overestimated robustness.

We admit that using PGD20+EoT10 may lead to overestimated robustness, but this issue **does not impact our main conclusions** including that stochasticity is the major contributor to DBP robustness, and that ADDT can improve the robust accuracy. Specifically, our supplementary results have consistently and adequately validated that ADDT significantly enhances the robustness of the DBP model, even when subjected to more intensive attacks. The experimental settings in the paper are selected to balance computational feasibility with attack effectiveness as outlined in Appendix A.

Q2.2. High inference cost.

The inference cost of DBP is known to be higher than that of a bare classifier due to the overhead of the purification model, as discussed in Appendix C.6 of [I]. However, the notable expenses of the robustness evaluation of DBP models originate mainly from the **large number of attack steps required** (e.g., 4000 for PGD200+EoT20) rather than the inference costs, especially when compared to deterministic models that do not require EoT during the attack. Meanwhile, computing the exact gradients of DBP models requires **gradient checkpointing** for reducing memory consumption, which introduces redundant computations and also slows down the robustness evaluation, but it **does not affect the normal inference**.

In practice, this actually enhances the security by **increasing the cost for attackers**. Additionally, compared with DiffPure, **integrating ADDT reduces the inference cost**, as we can achieve comparable performance with fewer necessary denoising steps (as suggested by Table 3).

[I] Nie, Weili, et al. Diffusion models for adversarial purification. ICML, 2022.

Q4. Inferior results on CIFAR-100 compared with AT.

We acknowledge that the CIFAR-100 results in this paper are inferior to those achieved by AT in the referenced table. We believe that this is due to the use of a **smaller UNet** compared with DiffPure (rather than ADDT itself), which may not have sufficient capacity for a more complex dataset. Our experiments on CIFAR-10 (Table 2) suggest that employing the larger UNet (DDPM++) could notably enhance performance. Unfortunately, limited by the remaining time of the discussion period (only 9 hours left), we could not provide the results on CIFAR-100 using a larger UNet. We commit to complementing the corresponding results in the revision.

In addition, we would like to clarify that the goal of proposing ADDT in this paper is to **enhance the DBP robustness**, instead of achieving state-of-the-art performance compared with other defense methods. While the performance of DBP may be inferior to AT under certain experimental settings, DBP has been regarded as a promising approach to practical adversarial defense in recent years, given its strengths in portability to different models and capability of defending unseen attacks, as compared with AT.

Q5. Marginal improvement on ImageNet.

We conducted a preliminary experiment using the smallest UNet and achieved an improvement on robust accuracy of **~2%** in our **updated experimental results** attached to the global response, which should not be marginal in our opinions. Besides, as we argued in Q4, using a larger UNet is expected to yield even more significant improvement.

Q4 & Q5. Scalability of the proposed method.

We understand that scalability is important for practical deep-learning algorithms, and we will be dedicated to complementing the experiments with necessary evaluations on harder or larger-scale datasets. However, as the title of this paper suggests, our contributions not only include the proposed ADDT for enhancing DBP, but also lie in the **new understanding of DBP robustness**, which highlights the stochasticity in DBP as the major contributor to the robustness, challenging previous views on the mechanism of DBP. Hence, **implementing a highly scalable and application-ready algorithm is not our main focus**. Instead, we hope that the new perspectives in this paper and the validated effectiveness of ADDT in improving the DBP robustness can motivate future research on DBP and the design of practical stochasticity-based defense methods.

Official Review of Submission1523 by Reviewer hZmv

Official Review Reviewer hZmv 12 Jul 2024, 05:33 (modified: 12 Aug 2024, 23:02)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer hZmv Revisions (/revisions?id=woQARUycKC)

Summary:
Diffusion-Based Purification (DBP) has emerged as a new adversarial defense framework against adversarial attacks. However, there is no reliable robustness evaluation of such defenses yet. The paper stated and justified that the model robustness by DBP is mainly attributed to the inherent stochasticity within the DBP models, which breaks the previous conventional understanding of reducing the distribution gap between clean and adversarial examples by the forward diffusion process. This paper introduced Deterministic White-box attack, a new adversarial attack setting that includes additional stochastic element information besides white-box attack. The new attack evaluation suggests that DBP models primarily leverage stochasticity to circumvent effective attack directions. Hence, this paper proposed Adversarial Denoising Diffusion Training (ADDT), a fine-tuning technique to further enhance the robustness of current state-of-the-art DBP methods and effectively mitigate the stochasticity issues within the iterative diffusion process.

Soundness: 3: good
Presentation: 3: good
Contribution: 3: good
Strengths:

1. This paper is well-written and easy to read.
2. The paper has done solid work by critically justifying that the effectiveness of DBP methods to existing strong adaptive attacks can be attributed to the intrinsic stochastic property of DBP.
3. The proposed fine-tuning method enhances the robustness of existing DBP methods by a notable margin and empirically shows that the stochasticity issue has been mitigated.

Weaknesses:

1. The experiments are conducted under small-scale datasets such as CIFAR-10, CIFAR-100 and Tiny-ImageNet datasets. It would be better if the authors could provide further experimental results on large high-resolution datasets like ImageNet, CelebA-HQ etc.
2. The number of attack iterations for PGD is neither sufficient nor convincing. I would encourage the authors to increase the attack iterations for PGD.
3. I would encourage the authors to keep the subset of the dataset consistent with the DiffPure experiment settings as DiffPure has been used for evaluation.

Questions:

1. In Appendix B, the authors provide different levels of deterministic white-box setting. We then know that the stochastic element corresponds to the added Gaussian noise ϵ during the forward process and the estimated noise ϵ_θ during the reverse process. It is hard to follow how you achieve the deterministic white-box attack to incorporate this stochastic information into the attack. Could you further explain the details of implementing such an attack setting?
2. In Figure 2, DP_DDPM has shown that the robust accuracy decreases within both DW_Fwd and DW_Both settings. We know that both the forward and reverse diffusion processes contain stochastic information. The robust accuracy slightly decreases from 47.27% to 45.41% when forward information is included, but drops instantly from 45.41% to 16.80% after utilising DW_Both. It seemed that the two diffusion processes bring different extents of influence on robust accuracy. Could you further explain this?

Limitations:
Yes.

Flag For Ethics Review: No ethics review needed.
Rating: 6: Weak Accept: Technically solid, moderate-to-high impact paper, with no major concerns with respect to evaluation, resources, reproducibility, ethical considerations.
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

07 Aug 2024, 04:51 (modified: 07 Aug 2024, 20:40) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=Vn5xZ2Ykap)

Rebuttal:

W1: Lack of results on higher-dimensional images.

Thanks for your valuable suggestion. Please refer to Q2 of the global response (Author Rebuttal).

W2: The number of attack iterations for PGD is insufficient.

Thanks for your suggestion on improving the evaluation of robustness. Please refer to Q1 of the global response (Author Rebuttal).

W3: The subset of experiment data should be consistent with DiffPure.

Thanks for your advice on keeping the experiment setting consistent. DiffPure conducts the robustness evaluation on a fixed random subset of 512 images, while we test the models on the first 1024 images. However, as reported in their paper [I] (Section 5, under the subsection "Evaluation Metrics"), whether evaluating on the sampled subset or the whole test set makes minor differences. Our preliminary attempt to re-evaluate on the same 512 images also supports this finding that the specific choice of subset is not significant.

The reason why we did not follow the experiment settings of DiffPure is that they rely on the adjoint method for robustness evaluation, which is less effective and leads to incomparable results with overestimated robustness. However, as you suggested, we will consider re-evaluating the models on the same subset if it is essential and affordable for us.

[I] Nie W, Guo B, Huang Y, et al. Diffusion models for adversarial purification. ICML, 2022.

Q1: Details of implementing deterministic white-box attack.

We appreciate your interest in the implementation of our deterministic attack methods. Firstly, we would like to clarify the stochastic elements in the diffusion process. Specifically, the *forward diffusion process* involves the noise variable ϵ, whereas the reverse diffusion process refers to the noise added when predicting x_{t-1} from x_t with ϵ instead of the predicted noise variable ϵ_θ. For DDPM, the *reverse diffusion process* is depicted in Equation (5), where the stochastic component corresponds to ϵ in the last term, rather than ϵ_θ. For DDIM, the reverse process is deterministic (i.e., an ODE process), so there is no such ϵ.

During each inference time of the diffusion process, a set of random Gaussian noise ϵ is sampled and utilized. In the Deterministic White-box setting, the **same set of noise** is replicated during the attack.

Formally, the forward process of DBP can be simplified as:

$x_t = F(x, \epsilon_f)$

The stochastic reverse process (e.g., DDPM, EDM, etc.) can be simplified as:

$x'_{t-1} = R_s(x_t, t, \epsilon_t)$

...

$x'_0 = R_s(x'_1, 1, \epsilon_1)$

The deterministic reverse process (e.g., DDIM) can be simplified as:

$x'_{t-1} = R_d(x_t, t)$

...

$x'_0 = R_d(x'_1, 1)$

Here, F and R represent the forward and reverse processes, respectively. The noise variables $(\epsilon_1, \ldots \epsilon_t)$ and ϵ_f are sampled from a Gaussian distribution to recreate x'_0 from x. If an attacker is aware of all the ϵ used in DBP, the process becomes deterministic for them. For example, in a stochastic reverse process, to control the noise, we sample a set of $\epsilon_1, \ldots, \epsilon_t$ and ϵ_f and use these during both the attack and evaluation phases. We can also partially control the process by fixing some noise components and sampling the others randomly. This approach does not affect the evaluation as all noise remains randomly sampled from a Gaussian distribution; the only difference is the attacker's knowledge.

Q2: How forward and reverse diffusion processes affect the robustness differently.

Thank you for your insightful questions. First, we would like to clarify that the striking difference between the drop in robust accuracy in Figure 2 as you point out does not reflect the comparison between the forward and reverse processes. Specifically, for DDPM, switching from conventional white-box to DW_Fwd results in a slight decrease in robust accuracy from 47.27% to 45.41%. In this case, **the model is still stochastic** to the attacker since the stochasticity in the reverse process is not accessible to the attacker. However, the drastic decrease in robust accuracy from 45.41% to 16.80% when switching from DW_Fwd to DW_Both can be explained by the **complete knowledge of the stochastic elements** of the attacker in the DW_Both setting. As argued in this paper, the existence of stochasticity to the attacker is the main contributor to the robustness of DBP.

To investigate the difference between forward and reverse diffusion processes in their contribution to the robustness, we test the model under an additional **DW_Rev** setting, where the attacker only knows the stochastic elements in the reverse process. The robust accuracy on DW_Rev is 35.25%, which is lower than that on DW_Fwd (i.e., 45.41%) and suggests that the stochasticity in the reverse process contributes more to the DBP robustness. This may be explained by the fact that for the reverse process of DDPM, the random noise is sampled independently for each iteration (see Q1 for illustration), while the forward process is not iterated and only samples the random noise once.

Replying to Rebuttal by Authors

Official Comment by Reviewer hZmv

Official Comment Reviewer hZmv 12 Aug 2024, 23:01

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thanks for your further explanation and answering all the questions. For W1 and W2, I would accept the result provided. For W3, I agree with your justification. If it is affordable, I hope you could align with the original settings in the future. For Q1, thanks for your further clarification, it makes the DW white-box attack implementation more clear. For Q2, after adding DW_Rev, it becomes more comparable with the results shown in DW_Fwd and DW_both. Overall, the author has thoroughly replied to my questions and I will increase my score to 6.

Replying to Official Comment by Reviewer hZmv

Official Comment by Authors

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

13 Aug 2024, 11:28 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thank you for your supportive and insightful review, which helps us enrich our experiments and refine the paper. We will carefully revise the paper according to the reviews and hope that our revision will satisfy you more.

Official Review of Submission1523 by Reviewer ZRj3

Official Review Reviewer ZRj3 01 Jul 2024, 12:01 (modified: 12 Aug 2024, 11:06)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer ZRj3 Revisions (/revisions?id=BI8j2QuZmh)

Summary:

This paper first analyzes the robustness of existing diffusion-based purification methods, and argues that the robustness of these methods stems from the stochasticity of diffusion process. After that, the authors propose to combine adversarial training with diffusion-based purification. Specifically, the authors propose Rank-based Gaussian Mapping to map the adversarial perturbations to a random variable in Gaussian distribution for more stable training, and design a pipeline for generating the adversarial examples for the whole model and train the diffusion by the generated adversarial examples.

Soundness: 1: poor

Presentation: 1: poor

Contribution: 2: fair

Strengths:

The authors propose Rank-based Gaussian Mapping to map the adversarial perturbations to a random variable in Gaussian distribution for more stable training, which is quite interesting.

Weaknesses:

- In lines 209-309, the authors claim "To improve the robustness evaluation of DBP, **we implement several modifications**". However, these modifications are **exactly the same as [1, 2]**. I suggest an ethnic review for such plagiarism.
- For Sec. 4.2. about the robustness of DBP, the authors claim that "we suggest that the loss landscape of DBP models is not inherently smooth". I admit that in DW-box setting it's not smooth since there is no stochasticity. However, in normal settings (except DW-box), the random noise in DBP effectively helps smooth out the local extrema and thus creates certified robustness [3, 4]. Also, since there are lot of stochasticity when attacking a model, even if this model is deterministic (optimization may induce stochasticity, and a small truncation error will be enlarged during iterative optimization), Fig. (3) is not trustworthy. It only measures optimization trajectory for one time. The authors should perform this experiment multiple times to average out the stochasticity during optimization. Currently, **this experiment has a Type I error of 50%**.

Reference

[1] Kang, et al. "DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification." Advances in Neural Information Processing Systems 36 (2023).

[2] Chen, et al. "Robust Classification via a Single Diffusion Model." arXiv preprint arXiv:2305.15241 (2023).

[3] Xiao, et al. "Densepure: Understanding diffusion models towards adversarial robustness." arXiv preprint arXiv:2211.00322 (2022).

[4] Carlini, et al. "(certified!!) Adversarial robustness for free!." arXiv preprint arXiv:2206.10550 (2022).

Questions:

no.

Limitations:

n/a

Flag For Ethics Review: No ethics review needed.

Rating: 4: Borderline reject: Technically solid paper where reasons to reject, e.g., limited evaluation, outweigh reasons to accept, e.g., good evaluation. Please use sparingly.

Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.

Code Of Conduct: Yes

Rebuttal by Authors

Rebuttal

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

06 Aug 2024, 13:29 (modified: 07 Aug 2024, 20:40) Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Revisions (/revisions?id=2AyF3dvDW7)

Rebuttal:

Q1: The robustness evaluation method is the same as previous works.:

Thanks for your comment. We would like to clarify that we do not claim the implementation of the robustness evaluation method as a main contribution of this paper (not listed in the Contribution part of Section 1), and we apologize for the relevant statements that seem to claim it as a novel technical contribution. We will properly cite the related works you provided (e.g., "following [1,2], we implement ...") and better clarify the minor technical differences between our evaluation protocol and those adopted in these previous works.

Q2.1: The relationship between DiffPure and Randomized Smoothing [IV]:

Thank you for your insightful comments. We appreciate this opportunity to clarify our position regarding the robustness of DBP models. While we discuss the role of randomness in robustness, our focus diverges from that of randomized smoothing and certified robustness.

- We would like to first clarify the foundational concepts. Conventionally, the classification models discussed in the studies of adversarial robustness can be viewed as mappings from input space X to the label space Y. However, DBP additionally involves a random variable $\epsilon \in E$ that determines the random sampling in the forward and reverse processes (which can be the random seed in implementation). Hence, a DBP model f can be viewed as the mapping $f : (X, E) \rightarrow Y$.
- Previous studies on randomized smoothing treat the randomized model f as a mapping $f : X \rightarrow P_Y$, where P_Y is the space of label distribution. Typically, the final prediction can be formulated as $F(x) = \arg\max_c [f(x)]_c$, i.e., the class c with the highest probability in the output distribution $f(x)$. Apparently, F deterministically maps X to Y, consistent with the conventional models.
- Recent studies on DBP also regard the model as $f : X \rightarrow P_Y$, without explicitly studying the role of ϵ. **The key difference between DBP and randomized smoothing is that the final prediction for an input x is directly sampled from the distribution $f(x)$ for once, instead of sampling multiple times to approximate $F(x)$ as in randomized smoothing.**
- In this paper, we revisit DBP by treating the randomized model f as the mapping $f : (X, E) \rightarrow Y$ and studying the role of $\epsilon \in E$ as an input of f. From this perspective, the conventional adversarial setting assuming full knowledge of the model parameters (but not ϵ) is not a complete white box, which motivates us to study the DW-box setting.
- From our perspective, we can clearly point out the difference between DBP and randomized smoothing in terms of the loss landscape. Given an input x_0, the local loss landscape for a DBP model f is not deterministic as it also depends on ϵ. **Although the expected loss landscape over $\epsilon \in E$ may be smooth, it does not suggest the robustness of DBP, as ϵ is fixed during a single inference run of DBP**. Indeed, our study suggests that given x_0 and a fixed ϵ_0, the local landscape of DBP is likely not smooth. In contrast, the loss landscape of a randomized smoothing model F may be smooth as it is the average landscape over multiple ϵ. **To conclude, we argue that the random noise itself may not smooth the loss landscape, but the average over random noises may.**

We hope this explanation clarifies our position and deepens your understanding of our approach. We will try our best to make these points clearer in our revision.

Q2.2: Clarification on Figure 3:

Thank you for your feedback. As detailed in Appendix D, Figure 3 is based on the first 128 images of CIFAR-10, and the optimization trajectory is averaged over these images. We apologize for any inconvenience caused by not including these details in the main text due to space constraints.

[I] Chen, et al. Training deep nets with sublinear memory cost. arXiv preprint arXiv:1604.06174, 2016.

[II] Kang, et al. DiffAttack: Evasion Attacks Against Diffusion-Based Adversarial Purification. NIPS, 2023.

[III] Chen, et al. Robust Classification via a Single Diffusion Model. ICML, 2024.

[IV] Cohen, et al. Certified adversarial robustness via randomized smoothing. ICML, 2019.

Official Comment by Reviewer ZRj3

Official Comment Reviewer ZRj3 12 Aug 2024, 11:06 (modified: 12 Aug 2024, 11:07)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors Revisions (/revisions?id=bm76oRkRHf)

Comment:
Thank you for your reply, which effectively addresses my concerns. However, the lack of scalability to large datasets limits its real-world application. I would be willing to accept this paper if the authors complement it with an experiment on ImageNet that is comparable to adversarial training (AT). Also, please remember to integrate these modifications in your next revision.

Replying to Official Comment by Reviewer ZRj3

Official Comment by Authors

Official Comment

Authors (ZiYi Dong (/profile?id=~ZiYi_Dong1), LiuYiming (/profile?id=~LiuYiming1), Yao Xiao (/profile?id=~Yao_Xiao4), Liang Lin (/profile?id=~Liang_Lin1), +3 more (/group/info?id=NeurIPS.cc/2024/Conference/Submission1523/Authors))

12 Aug 2024, 16:45 Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors

Comment:

Thanks for your suggestion on comparing with AT in the ImageNet experiments. We have updated the results on ImageNet in the **Update 2 in the comment to our global response**. In summary, our experiments suggest that ADDT can achieve at least comparable robust accuracy and significantly better clean accuracy compared with AT. Please refer to the global comment for details.

We are happy to continue the discussion if you have any further concerns.

About OpenReview (/about)
Hosting a Venue (/group?id=OpenReview.net/Support)
All Venues (/venues)
Sponsors (/sponsors)

Frequently Asked Questions (https://docs.openreview.net/getting-started/frequently-asked-questions)
Contact (/contact)
Feedback
Terms of Use (/legal/terms)
Privacy Policy (/legal/privacy)