Adversarial Purification via Super-Resolution and Diffusion

Mincheol Park[1,2], Cheonjun Park[5], Seungseop Lim[4], Mijin Koo[6],
Hyunwuk Lee[2], Won Woo Ro[3], Suhyun Kim[7*]
[1]Samsung Advanced Institute of Technology, [2]Samsung Electronics, [3]Yonsei University, [4]AITRICS,
[5]Hankuk University of Foreign Studies, [6]Seoul National University, [7]Kyung Hee University
{mincheol.park, cheonjun.park, hyunwuk.lee, wro}@yonsei.ac.kr,
ss.lim@aitrics.com, starmj09@snu.ac.kr, dr.suhyun.kim@gmail.com

Abstract

Deep neural networks are widely used in various computer vision tasks, but their vulnerability to adversarial perturbations remains a significant challenge for reliable decision-making. Adversarial purification, a test-time defense strategy, has shown potential in countering these threats by removing noise through diffusion models. This plug-and-play method, using off-the-shelf models, appears highly effective. However, the purified data from diffusion often deviates more from the original data than the adversarial examples, leading to missing critical information and causing misclassification. In this study, we propose that upsampling with Super-Resolution (SR), followed by downsampling, can also aid in eliminating adversarial noise, similar to the noise addition and removal process in diffusion models. While SR alone is not as effective as the diffusion process, it better restores the original features typically associated with the early layers of networks. By combining SR, which initially mitigates damage to early-layer information from adversarial attacks, with diffusion, we observe a synergistic effect, leading to enhanced performance over diffusion models alone. Our comprehensive evaluations demonstrate that this combined approach, PuriFlow, significantly improves accuracy and robustness, working synergistically with state-of-the-art methods.

1. Introduction

Deep Neural Networks (DNNs) have been widely accepted in various computer vision tasks. However, their susceptibility to small, human-imperceptible noises can lead to untrustworthy decision-making. In particular, the subtle, maliciously manipulated noises, known as adversarial perturbations, produce fatal images called adversarial examples [10]. This intentional threat appears across numerous types of attacks [3, 9, 13, 34], leaving many DNNs vulnerable and posing substantial challenges to their deployment in AI-powered systems.

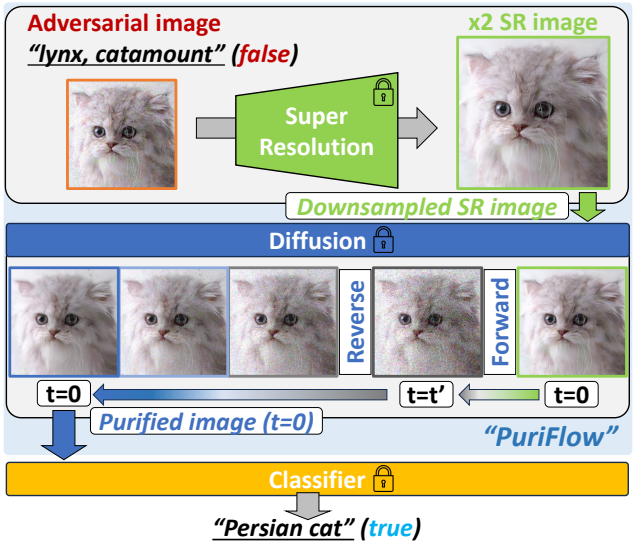


Figure 1. The overview of PuriFlow, which begins with restoring features of adversarial examples using the SR and drifting downsampled SR images within the forward diffusion. The perturbed images are then denoised in the reverse diffusion for classification.

To counter the adversarial threats, adversarial purification [38, 58] has emerged as a test-time defense strategy. This approach is powered by leveraging off-the-shelf diffusion models. At its core, noise removal through stochastic denoising effectively eliminates artifacts introduced by various attacks. Unlike adversarial training and certified methods [27, 30, 34], which require prior knowledge of target attacks, this simple application of generative power excels in protecting classifier without the need for training, pointing to a new direction.

While diffusion models appear effective in purification, a key problem remains widely discussed in studies [5, 11, 32,

*Corresponding author

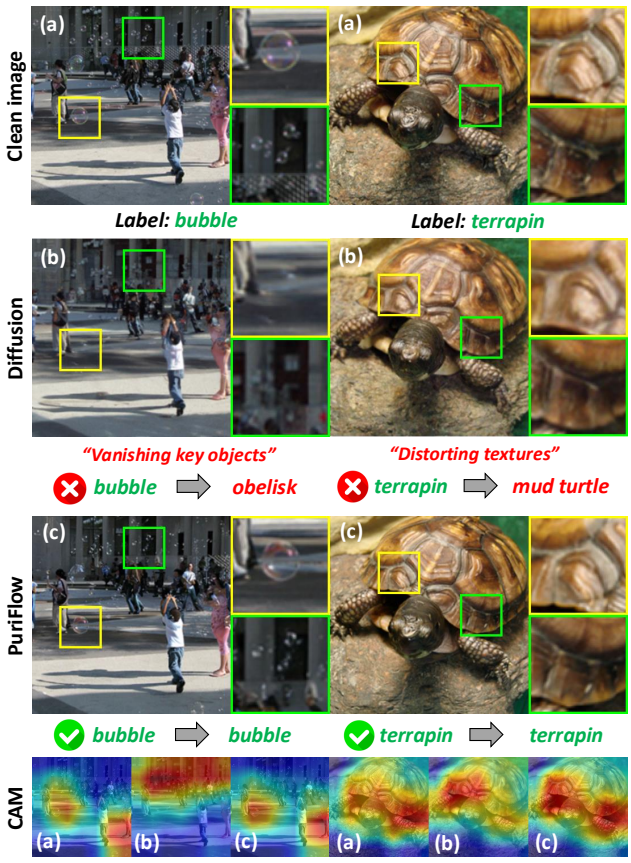


Figure 2. The example of purification on the clean images. The top row exhibits the ground truth images. The second and third rows depict purified images obtained using diffusion and our proposed PuriFlow. The fourth row visualizes Grad-CAM [44] as represented by ResNet-50. Compared to Diffpure, PuriFlow enhances visual fidelity and accuracy by preserving DNN's activation.

38, 60]: *how to lower the upper bound of differences between purified and original data, which can exceed those caused by adversarial attacks*. This concern also arises when clean data is input, as shown in Figure 2. However, resolving it solely within the diffusion model is not straightforward. The noise predictor in off-the-shelf diffusion models is optimized to predict only the mean under isotropic covariance [26], limiting its ability to fully restore subtle features tied to true data covariates. To improve the restoration of details, diffusion models may require retraining with a proper objective [37], which may not align well with test-time defenses. Recently, guidance during noise removal has been proposed for test-time defenses [5]; however, this approach, despite allowing extensive processing time, is limited to large-scale datasets, impacting its generality.

To improve purification without retraining, we note that adversarial attacks are designed to maximize changes in model output with minimal alterations to the input. This results in alterations in the feature map that can be substantial, while changes in the image might be minor. Diffusion-based purification processes, which lack prior knowledge of the original image, risk deviating in the wrong direction during the noise addition and recovery phases, especially if the starting image contains feature information significantly distorted by adversarial noises. Therefore, it is crucial to eliminate as much distorted information as possible before applying the diffusion process.

We hypothesize that using Super-Resolution (SR) to upsample and then downsample an image could help reduce noise. Our tests confirm that while SR does reduce adversarial noise, it is not as effective as diffusion models. However, an in-depth experiment reveals that SR performs better than diffusion models in restoring features to the early layers of a network. This insight suggests that initially, using SR to reduce adversarial noise before applying diffusion could enhance overall performance by restoring the features the diffusion process alone cannot recover.

This paper proposes *PuriFlow*, an enhanced purification flow that leverages the synergy of SR and the diffusion process. As shown in Figure 1, PuriFlow integrates SR before the forward diffusion process to restore distorted features of adversarial images early on. By boosting *content* similarity [28], SR generates enhanced samples with features closer to the originals, even when mapped into a high-dimensional space. Our findings show that content proximity increases when SR images are resized back to their original dimensions, positioning resized SR images in regions with a higher probability for correct labels. In other words, resizing SR effectively reduces the initial cross-entropy on adversarial examples, allowing better diffusion. Our theoretical analysis and comprehensive evaluations demonstrate that this initialization aids in restoring features by diffusion, further reduces cross-entropy, and improves accuracy.

Our contributions can be summarized as three-fold:

- We find the purification effectiveness of the SR process. While SR alone is less effective in fully eliminating adversarial noises, it excels at preserving original features in the early layers of the network.
- Our approach, PuriFlow, finds a synergistic effect by integrating SR with diffusion. Our insight into geometrical analysis and empirical study finds that this integration can reduce cross-entropy in adversarial examples.
- Our pipeline utilizes only off-the-shelf SR and diffusion models. With negligible execution time required for SR, this plug-and-play method demonstrates improved accuracy and robustness in comprehensive evaluations.

2. Related Work

Denoising defenses. Traditional test-time adversarial defenses often apply denoising techniques to input images. JPEG compression has been employed to remove high-

frequency noise [14, 16], but Guo et al. [22] found it insufficient, proposing total variance minimization with image quilting as an alternative. Likewise, Prakash et al. [40] use wavelet denoising combined with BayesShrink to mitigate artifacts leading to misclassification, which surpasses total variance minimization and Wiener filtering in adversarial defenses [54]. Xie et al. [55] denoise feature maps through non-local means, supporting adversarial training. While practical, these methods often distort image details and face a core challenge: the inverse problem, which results in non-unique solutions.

SR defenses. Modern SR models have gained attention to enhance perceptual quality in adversarial examples [36]. Recently, Bhardwaj et al. [7] raised concerns about the computational demands of the SR models, which can limit real-time applicability. They demonstrate the effectiveness of a highly efficient SR model [8], showing robust performance even on compact neural processing units. PuriFlow shares a similar purpose on rapid SR applications but adds weight to its synergistic use for restoring refined features with minimal overhead, assisting diffusion to enhance defense.

Adversarial purification. Purifying adversarial noise using generative models has shown promise in strengthening classifier robustness as a test-time defense. PixelDefend [47] utilizes an autoregressive generative model, while Defense-GAN [43] relies on GANs. Langevin Dynamics (LD) sampling has also been effective for defenses through energy-based models [15, 21, 25]. Yoon et al. [58] introduce a denoising score-based model with an LD variant for denoising processes. DiffPure [38] and Lee and Kim [32] employ forward and reverse-time Stochastic Differential Equations (SDEs) [26, 46] for purification, but the loss of class information remains a risk due to the dependence on diffusion time steps. Bai et al. [5] mitigate this problem with contrastive guidance, while ScoreOpt [60] uses a one-shot denoiser. This work attempts to tackle losing crucial information in purified images by combining SR with diffusion.

3. Background for Diffusion Models

The score-based diffusion model is widely accepted in adversarial purification. It performs with bidirectional processes: forward and reverse time transition of $\mathbf{x}_t \in \mathbb{R}^d$ over the time interval $t \in [0, T]$. Starting from an original sample $\mathbf{x}_0 \sim p_0(\mathbf{x})$ where $p_0(\mathbf{x})$ is the unknown, true data distribution, the diffusion model progressively transforms $p_0(\mathbf{x})$ into a nearly spherical Gaussian distribution $p_T(\mathbf{x}) \approx \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$. This converted distribution $p_T(\mathbf{x})$ retains no information about $p_0(\mathbf{x})$.

This forward process of $\{\mathbf{x}_t\}_{t=0}^T$, as defined by Itô SDE, is given with the positive increment of t:

$$d\mathbf{x} = \boldsymbol{f}(\mathbf{x}, t)dt + g(t)d\mathbf{w}, \tag{1}$$

where the initial data $\mathbf{x}_0 \sim p_0(\mathbf{x})$, $\boldsymbol{f}(\cdot, t) : \mathbb{R}^d \mapsto \mathbb{R}^d$ is the drift coefficient of $\mathbf{x}_t$ and $g(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ is the diffusion coefficient tied with Brownian motion $\mathbf{w} \in \mathbb{R}^d$. Here, when $\boldsymbol{f}(\mathbf{x}, t)$ is affine, the transition kernel always becomes a Gaussian distribution $p_{t't}(\mathbf{x}_t|\mathbf{x}_{t'})$ where $0 \leq t' < t \leq T$ [49]. Thus, the forward diffusion process allows for a direct transition with a closed form using the Gaussian kernel $p_{0t}(\mathbf{x}_t|\mathbf{x}_0)$ from $t = 0$ to a certain t, which prevents the need for neural network estimations.

Given a perturbed sample $\mathbf{x}_T \sim p_T(\mathbf{x})$, the reverse-time SDE defines that it drifts backward in the time steps:

$$d\mathbf{x} = [\boldsymbol{f}(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]dt + g(t)d\bar{\mathbf{w}}, \tag{2}$$

where $\bar{\mathbf{w}}$ represents a standard Wiener process and the infinitesimal time step dt is negative. A key component in Eq. 2 is the score function $\nabla_{\mathbf{x}} \log p_t(\mathbf{x})$, which offers how the density $p_t(\mathbf{x})$ changes with respect to $\mathbf{x}$. Instead of directly knowing this function, one approach is to train a neural network, $\mathbf{s}_\theta(\mathbf{x}, t)$, to estimate it. This training is achieved by solving a denoising score-matching [51] problem:

$$\ell(\theta) = \sum_{t=1}^{T} \lambda(t)\mathbb{E}_{\mathbf{x},\tilde{\mathbf{x}}|\mathbf{x}}[\mathbf{s}_\theta(\tilde{\mathbf{x}}, t) - \nabla_{\tilde{\mathbf{x}}} \log p_{0t}(\tilde{\mathbf{x}}|\mathbf{x})||_2^2]. \tag{3}$$

Here, Gaussian kernel $p_{0t}(\tilde{\mathbf{x}}|\mathbf{x})$ depicts the transition from $\tilde{\mathbf{x}} = \mathbf{x}_t$ to $\mathbf{x} = \mathbf{x}_0$ and $\lambda(t)$ is weighting function ensuring that the contributions at different times are appropriately balanced. This optimized model serves as $\mathbf{s}_{\theta^*}(\mathbf{x}, t)$ where $\theta^* = \text{argmin}_\theta \ell(\theta)$ to replace the elusive score function, enabling an effective reverse diffusion process.

4. Methodology

In this section, we introduce our motivation, design methodology, and geometrical analysis of the proposed approach, *PuriFlow*. To begin, we briefly review adversarial attacks.

Adversarial attack. Given a clean image $\mathbf{x}_0$ with label y, adversarial attacks generally seek perturbations to mislead a classifier $\mathcal{F}_\phi(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^K$ into incorrectly classifying $\mathbf{x}_0$. This manipulated image $\mathbf{x}_0^a$ is generated by maximizing cross-entropy loss within a tiny region around $\mathbf{x}_0$:

$$\mathbf{x}_0^a = \lim_{i \to N} \prod_{\mathcal{B}(\mathbf{x}_0^i, \epsilon)} (\mathbf{x}_0^i + \mu \text{sign}(\nabla_{\mathbf{x}_0^i} \ell_{ce}(\phi; \mathbf{x}_0^i, y))), \tag{4}$$

where $\mathbf{x}_0^0$ is the initial clean image $\mathbf{x}_0$ and N is the number of attack iterations. $\mathcal{B}(\mathbf{x}_0^i, \epsilon)$ is the ℓ_p-norm ball around $\mathbf{x}_0^i$ within a radius of ϵ, $\Pi(\cdot)$ is the projection to the norm ball, and μ denotes the step size. The gradient can be obtained from surrogate networks when $\mathcal{F}_\phi(\cdot)$ is unknown.

4.1. PuriFlow: Enhanced Purification Flow

Given an adversarial example, $\mathbf{x}_0^a$, VP-SDE [48] is widely accepted approach for purification. VP-SDE controls noise

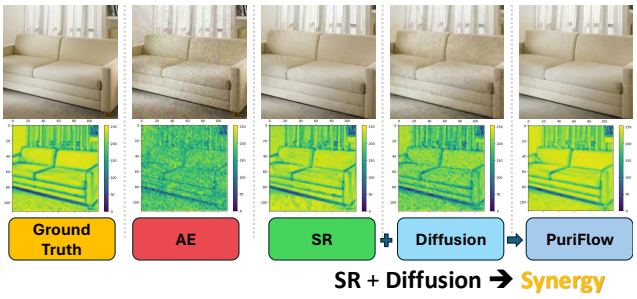


Figure 3. **Top**: visualization of different image types. AE denotes the adversarial example generated by a PGD attack on ResNet-50 using an ImageNet-1k image. SR indicates the downsampled high-resolution image. Diffusion and PuriFlow use the same diffusion method, VP-SDE [38]. **Bottom**: corresponding feature maps from the Conv2_2 layer of VGG-19, as described in [18].

addition through drift coefficients $\boldsymbol{f}(\mathbf{x}, t) = -\frac{1}{2}\beta_t \mathbf{x}$ and $g(t) = \sqrt{\beta_t}$, where $\beta_t = \frac{1}{t}\beta_{max} + (1 - \frac{1}{t})\beta_{min}$ [38]. This model then diffuses the data by solving the forward-time SDE in Eq. 1 with a closed form:

$$\mathbf{x}_t^a = \sqrt{\bar{\alpha}_t}\mathbf{x}_0^a + \sqrt{1 - \bar{\alpha}_t}\mathbf{z}_t, \tag{5}$$

where $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ and $\bar{\alpha}_t = e^{-\int_0^t \beta_t(s)ds}$.

The corresponding diffusion process is then performed by solving the reverse-time SDE in Eq. 2:

$$d\mathbf{x} = [\boldsymbol{f}(\mathbf{x}_t^a, t) - g^2(t)\nabla_{\mathbf{x}_t^a} \log p_t(\mathbf{x}_t^a)]dt + g(t)d\bar{\mathbf{w}}, \tag{6}$$

where the drift coefficients, $\boldsymbol{f}(\cdot, t)$ and $g^2(t)$, follows [5, 48].

During this reverse process, the crucial step is computing the score, defined as $\nabla_{\mathbf{x}_t^a} \log p_t(\mathbf{x}_t^a)$. Here, we raise a key question: *Can the neural network $\boldsymbol{s}_{\theta^*}(\cdot, t)$, trained for score estimation, precisely predict $\boldsymbol{x}_0$ from $\boldsymbol{x}_t^a$?* Even if $\mathbf{s}_{\theta^*}(\cdot, t)$ is well-trained, accurately estimating the direction toward $\mathbf{x}_0$ is not straightforward when dealing with the noised data $\mathbf{x}_t^a$ that has been perturbed from $\mathbf{x}_0^a$, since $\mathbf{s}_{\theta^*}(\cdot, t)$ has not been exposed to such perturbed data during training with Eq. 3. In other words, the trained neural network lacks prior knowledge of guiding $\mathbf{x}_t^a$, risking deviations in the wrong direction during the noise addition and recovery phases.

This concern is seemingly explicit. Figure 3 shows that the perturbations introduced by Eq. 4 appear subtle in pixel values but cause significant damage to the underlying features of $\mathbf{x}_0$. These observations indicate that diffusion alone struggles to fully reconstruct the original features. To address this, we seek an existing method that positions the perturbed data closer to its ground-truth image, aiming to reduce feature discrepancies early on and thus ensure a more precise projection of the score function.

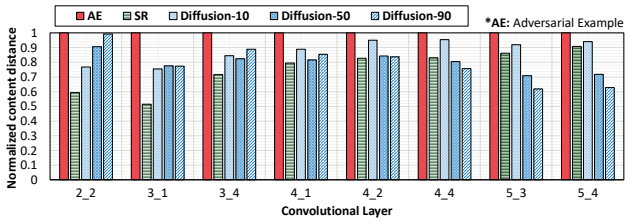
To this end, we focus on a model that can effectively reduce *content* distance [18, 28] to capture feature proximity. Reducing feature-level shifts that lead to increased cross-entropy aligns with the goal of purification, ultimately improving classification accuracy. This approach also benefits clean data in preserving features. In this context, the SR model suits this purpose. The SR model, $\mathcal{H}_{\omega^*}(\cdot) : \mathbb{R}^d \mapsto \mathbb{R}^D$ such that $D > d$, is designed to upsample and interpolate low-resolution images into high-resolution counterparts. To optimize its parameters ω, various SR methods [1, 28, 31, 45, 52] minimize both pixel distance and VGG loss [31], which captures mutual similarity in feature maps across different layers of the VGG model:

$$\omega^* = \min_\omega \frac{1}{W^l H^l}\Big(\sum_{u=1}^{W^l}\sum_{v=1}^{H^l}[\Phi^l(\mathbf{X})_{u,v} - \Phi^l(\mathcal{H}_\omega(\mathbf{x}))_{u,v}]^2\Big), \tag{7}$$

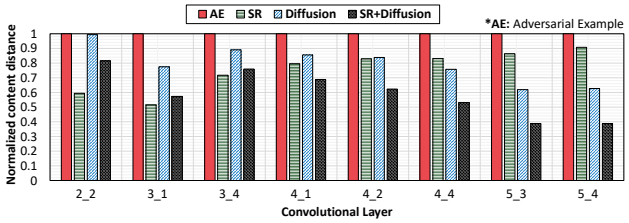
where $\mathbf{X} \in \mathbb{R}^D$ and $\mathbf{x} \in \mathbb{R}^d$, with W^l and H^l denoting the dimensions of extracted features $\Phi^l(\cdot)$ in the l-th convolution layer. Thus, SR models can naturally reduce the content distance by enhancing feature similarity while estimating unknown pixels and their covariates in $\mathbb{R}^D$.

Our findings, as illustrated in Figure 4, show that SR images downsampled from $\mathbb{R}^D$ to $\mathbb{R}^d$ can also reduce the content distance. This reduction covers all layers of the VGG network, demonstrating that even a one-shot application can effectively mitigate feature discrepancies present in adversarial examples. Notably, as observed in Figure 4a, increasing t' when applying diffusion alone cannot sufficiently aid in restoring features tied to the early layers of the CNN. This observation suggests that a single application of SR can be more effective than diffusion in recovering detailed features [59] captured by the early layers.

In this paper, we propose an enhanced purification flow,



(a) Impact of SR compared to diffusion with increments of t'.



(b) Synergistic effect in feature restoration of SR combined with diffusion. Diffusion and SR+Diffusion use the same diffusion time $t' = 90$.

Figure 4. Normalized content distance from the ground truth for each image type, measured across specific CONV layers of VGG-19 as shown in Figure 5. Real values are provided in the Appendix.

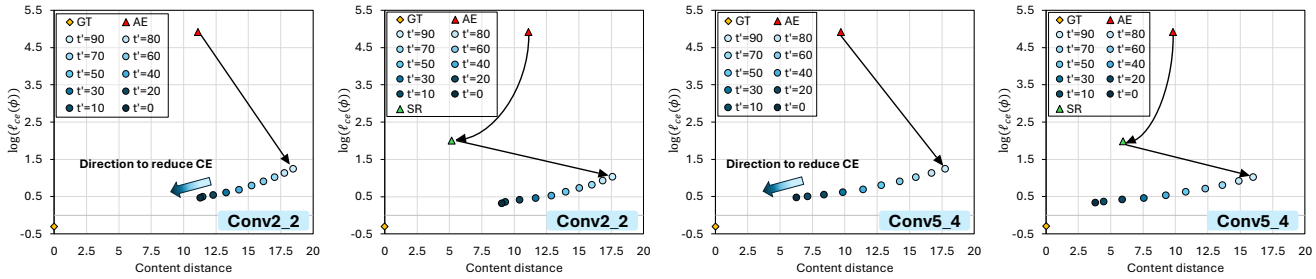


Figure 5. Transition of diffusion only and SR-integrated diffusion with the same diffusion time of $t' = 90$ under a PGD attack ($\epsilon = 8/255$, $N = 100$) on ResNet-50 using ImageNet-1k. "SR" represents downsampled high-dimensional images, while "AE" denotes adversarial examples. Content distance [18] from the ground truth is measured in two convolutional layers of VGG-19 and plotted with cross-entropy.

PuriFlow, where SR works before the start of the forward-time SDE. As shown in Figure 4b, this approach demonstrates the effectiveness of using SR early before solving both SDEs. By correcting feature distortions that diffusion alone struggles to recover, SR allows for a synergistic effect, enhancing feature restoration across all layers through diffusion. As observed in Figure 5, the early reduction in content distance leads to a notable decrease in cross-entropy. Thus, the subsequent diffusion aligns the data more closely with the ground truth. To delve into this synergistic effect, we provide a theoretical study from a manifold perspective. We examine how SR influences data positioning and allows for more precise use of the score function within diffusion.

4.2. Geometrical Analysis on PuriFlow

Suppose $\mathcal{M} \subset \mathbb{R}^d$ is the set of all data points, i.e., $\forall \mathbf{x} \in \mathcal{M}$, defining $\mathcal{M}$ as the data manifold. That is, the data distribution p_0 is uniform on the data manifold $\mathcal{M}$. Then, within a local neighborhood, the manifold coincides with its tangent space of dimension $k \ll d$, expressed as follows:

$$\mathcal{M} \cap \mathcal{B}(\mathbf{x}_0, dr) = T_{\mathbf{x}_0}\mathcal{M} \cap \mathcal{B}(\mathbf{x}_0, dr), \quad T_{\mathbf{x}_0}\mathcal{M} \simeq \mathbb{R}^k, \quad (8)$$

where $T_{\mathbf{x}_0}\mathcal{M}$ denotes the tangent space to the manifold $\mathcal{M}$ at $\mathbf{x}_0$ and dr is an infinitesimal radius within which the manifold can be locally approximated by its tangent space. Here, we take note of a property of the trained score function in diffusion models [12].

Remark 1. *The noisy data become increasingly concentrated on spherical manifolds as diffusion time t increases. The score function is trained via denoising score matching (Eq. 3), using data on $\mathcal{M}$. Let $\mathcal{P}_t(\cdot)$ denote a mapping from noisy data $\boldsymbol{x}_t$ to its estimated clean counterpart $\hat{\boldsymbol{x}}_0$:*

$$\mathcal{P}_t(\cdot) : \boldsymbol{x}_t \mapsto \hat{\boldsymbol{x}}_0 = -\frac{1}{2}(1-\alpha_t)(\boldsymbol{x}_t + 2\boldsymbol{s}_{\theta^*}(\boldsymbol{x}_t, t)), \quad (9)$$

where $\mathcal{P}_t(\boldsymbol{x}_t) \in \mathcal{M}$ and it satisfies $\boldsymbol{J}^2_{\mathcal{P}_t} = \boldsymbol{J}_{\mathcal{P}_t} = \boldsymbol{J}^T_{\mathcal{P}_t} : \mathbb{R}^d \to T_{\mathcal{P}_t(\boldsymbol{x}_t)}\mathcal{M}$. That is, $\mathcal{P}_t(\cdot)$ acts as an orthonormal projection onto $\mathcal{M}$.

Remark 2. *The score function is trained solely on the data concentrated on both the noisy and data manifolds. Therefore, applying it to data outside these manifolds may lead to inaccurate inference.*

Both premises suggest that when the score function is applied to off-manifold data, such as adversarial examples [7], its inaccurate inference may lead to misaligned orthonormal projections onto an unintended tangent space. Thus, the off-manifold data should be corrected to better align with its underlying manifold. Now, we show that the applying SR followed by downsampling can typically position off-manifold data, degraded from $\mathbf{x}_0$, closer to the local manifold, which is the tangent space of $\mathbf{x}_0$.

Proposition 1. *Suppose $\mathcal{H}_{\omega^*}$ is an ideal SR and smooth continuity, and let $\mathcal{R}$ be a linear downsampler. Given an off-manifold data $\mathfrak{D}(\boldsymbol{x}_0)$ degraded from $\boldsymbol{x}_0$, the processed data $\mathcal{R}\mathcal{H}_{\omega^*}(\mathfrak{D}(\boldsymbol{x}_0))$ is projected near the tangent space $T_{\boldsymbol{x}_0}\mathcal{M}$ of the data manifold $\mathcal{M}$ at $\boldsymbol{x}_0$ with a tiny $\delta > 0$:*

$$\|\mathcal{R}\mathcal{H}_{\omega^*}(\mathfrak{D}(\boldsymbol{x}_0)) - \Pi_{T_{\boldsymbol{x}_0}\mathcal{M}}(\mathfrak{D}(\boldsymbol{x}_0))\|_F < \delta. \quad (10)$$

In Section 5, we evaluate the effectiveness of using the score function more accurately, supported by Proposition 1, against various adversarial attacks.

5. Experiment

We evaluate the performance of PuriFlow in two configurations: integrating SR with solving SDEs [38] and combining SR with a recent alternative, iterative One-Shot Denoising (OSD) [60]. OSD is defined by Tweedie's formula [50], expressed as $\mathbb{E}[\mathbf{x}|\tilde{\mathbf{x}}] = \tilde{\mathbf{x}} + \sigma^2 \nabla_{\tilde{\mathbf{x}}} \log p_\sigma(\tilde{\mathbf{x}})$, where $\tilde{\mathbf{x}}$ denotes perturbed data and $p_\sigma(\tilde{\mathbf{x}})$ represents marginal Gaussian perturbation of $p(\mathbf{x})$. This formulation avoids repeated noise removal steps for varying time t.

In this context, the denoising score-matching problem defined by Eq. 3 optimizes the neural network $\mathbf{s}_{\theta^*}(\cdot, t)$, given $\mathbf{x}_t$, to predict a fully denoised $\mathbf{x}_0$ per a specific time t. Thus, Tweedie's formula can be represented as:

$$\mathbb{E}[\mathbf{x}_0|\mathbf{x}_{t'}] = \mathbf{x}_{t'} + \sigma^2 \nabla_{\mathbf{x}_{t'}} \log p_\sigma(\mathbf{x}_{t'}). \quad (11)$$

Method	Criterion	Standard (%)	Robust (%)
Yang et al. [56]	Gibbs Update	94.80	40.80
Song et al. [47]	Mask+Recon.	95.00	9.00
Hill et al. [24]	EBM+LD	84.12	54.90
Yoon et al. [58]	DSM+LD*	86.14	70.01
Nie et al. [38]	SDEs	89.02	81.40
Bai et al. [5]	Guide+SDEs	92.61	81.94
PuriFlow([38])	**SR+SDEs**	**90.06**	**83.00**
Zhang et al. [60]	iOSD	93.75	88.08
PuriFlow([60])	**SR+iOSD**	**94.14**	**91.01**

Table 1. Comparison of different adversarial purification methods against BPDA+EOT attacks with ℓ_∞ perturbations. Our evaluations uses WRN-28-10 on CIFAR-10, with consistent experimental settings as outlined in [38, 60] for $\epsilon = 8/255$. (* This purification utilizes a variant of the LD sampling.)

The alternative proposed by [60] involves iteratively applying OSD using Eq. 11 during noise removal. Therefore, it can serve as a suitable baseline to confirm synergies with SR. We assess these configurations, detailed as algorithms in the Appendix, across key test-time defense scenarios, including adaptive white-box, preprocessor-blind [58], and black-box attacks. This section also includes an ablation study to demonstrate the impact of incorporating SR. Detailed experimental settings and additional evaluations for certified robustness are provided in the Appendix. For adaptive attacks, we adopt MDSR [33] for efficiency and ESRGAN [52] otherwise. We denote t' on a discrete scale [26] from 1 to 1000 to represent the continuous interval [0, 1].

5.1. Defense on Adaptive Attacks

In this section, we evaluate the performance of PuriFlow against four strong adaptive attacks: BPDA+EOT [3, 4], an adaptive ensemble white-box attack; AutoAttack [13], an adaptive white-box attack; PGD+EOT [32], and DiffAttack [29], a diffusion-based purification targeted attack. Here, the evaluations focus on projections within an ℓ_∞-norm ball, as the projection in ℓ_2-norm ball is less effective.

BPDA+EOT. Backward Pass Differentiable Approximation (BPDA) combined with Expectation over Transformation (EOT) is used to evaluate randomized adversarial purification methods. Table 1 demonstrates that PuriFlow outperforms standalone diffusion on CIFAR-10 with WRN-28-10, achieving a more balanced standard and robust accuracy than the guiding method. In addition, integrating SR with iterative OSD further enhances performance, emphasizing its critical role in diffusion-based purification.

AutoAttack. Tables 2 demonstrates the effectiveness of PuriFlow against AutoAttack, on ImageNet-1k. PuriFlow outperforms adversarial training methods and a diffusion-only approach across ResNet-50 and DeiT-S. Compared to diffusion only, it achieves notable robust accuracy improvements, including 11.99% (ResNet-50) and 9.94% (DeiT-S), and standard accuracy. PuriFlow also shows efficiency with reduced diffusion time t' on this large-scale dataset.

Type	Method	Standard (%)	Robust (%)
ResNet-50		76.13	0.00
AT	Engstrom et al. [17]	62.56	31.06
	Wong et al. [53]	55.62	26.95
	Salman et al. [42]	64.02	37.89
	Bai et al. [6]	67.38	35.51
AP	Nie et al. [38]	67.79	40.93
	PuriFlow([38])	**71.68**	**52.92**
DeiT-S		79.90	0.00
AT	Bai et al. [6]	66.50	35.50
AP	Nie et al. [38]	73.63	43.18
	PuriFlow([38])	**75.69**	**53.12**

Table 2. Evaluation comparing adversarial training (AT) and purification (AP) methods against AutoAttack with ℓ_∞ perturbations at $\epsilon = 4/255$ on ImageNet-1k. Diffusion time is set to $t' = 150$ for [38] as specified its study and $t' = 130$ for PuriFlow.

Type	Method	Standard (%)	Robust (%)
WRN-28-10		95.63	0.00
AT	Pang et al. [39]	88.62	64.95
	Gowal et al. [19]	88.54	65.93
	Gowal et al. [20]	87.51	66.01
AP	Yoon et al. [58]	85.66	33.48
	Zhang et al. [60]*	93.75	59.37
	PuriFlow([60])*	**94.14**	**60.76**
WRN-70-16		95.79	0.00
AT	Gowal et al. [19]	91.10	68.66
	Gowal et al. [20]	88.74	69.03
	Rebuffi et al. [41]	92.22	69.97
AP	Yoon et al. [58]	86.76	37.11
	Zhang et al. [60]*	95.11	63.28
	PuriFlow([60])*	**95.31**	**64.86**

Table 3. Evaluation comparing adversarial training (AT) and purification (AP) methods against PGD+EOT within an ℓ_∞-norm ball of radius $\epsilon = 8/255$ on CIFAR-10. Results are sourced from [32]. (* For a fair comparison, PuriFlow and [60] are directly evaluated on the same settings with $t' = 250$, as stated in [60].)

PGD+EOT. Motivated by [32], we assess PuriFlow's robustness against PGD+EOT that utilizes gradient signs within the entire randomized defense framework. Table 3 shows PuriFlow's strength on CIFAR-10, achieving the highest standard accuracy for WRN-28-10 and WRN-70-16. Despite the substantial impact of PGD+EOT on randomized purification, PuriFlow achieves superior robust accuracy among purification methods, demonstrating a synergistic effect in which SR effectively acts as the initialization before starting the iterative one-shot denoising process.

Type	Method	Standard (%)	Robust (%)
WRN-70-16		95.79	0.00
AP	Nie et al. [38]	90.07	45.31
	PuriFlow([38])	**90.21**	**59.47**
ResNet-50		76.13	0.00
AP	Nie et al. [38]	67.79	28.13
	PuriFlow([38])	**71.68**	**43.75**

Table 4. Evaluation against DiffAttack, targeting WRN-70-16 on CIFAR-10 and ResNet-50 on ImageNet-1k. Each diffusion time t' is the same as those used for other attacks on PuriFlow.

Type	Method	Standard (%)	Robust (%)
ResNet-50		76.13	0.00
AT	Engstrom et al. [17]	62.56	38.97
	Wong et al. [53]	55.62	29.15
	Salman et al. [42]	64.02	38.01
	Bai et al. [6]	67.38	40.27
AP	Nie et al. [38]	74.52	38.87
	PuriFlow([38])	**74.05**	**49.92**
DeiT-S		79.90	0.00
AT	Bai et al. [6]	66.50	40.32
AP	Nie et al. [38]	78.94	34.59
	PuriFlow([38])	**78.44**	**49.85**

Table 5. Comparison under a preprocessor-blind PGD attack targeting only the classifier, within an ℓ_∞ perturbations at $\epsilon = 4/255$ on ImageNet-1k. The time t' for both AP methods is set as 10.

DiffAttack. DiffAttack neutralizes vanishing and exploding gradients, high memory costs, and increased randomness—natural defensive effects from lengthy t' of diffusion-based purification. Table 4 demonstrates that while purification methods using $t' = 100$ for CIFAR-10 and $t' = 150$ for ImageNet-1k exhibit limitations in robustness by this intentional attack, integrating SR demonstrates a notable improvement by 14.16% and 15.62%. We hypothesize that SR's early enhancement of feature proximity helps mitigate diverging feature maps from the ground truth, even under the *reconstruction-deviating* impacts of DiffAttack.

5.2. Defense on Preprocessor-Blind Attack

We evaluate PuriFlow against the PGD ($N = 100$) attack, targeting only the classifier on ImageNet-1k. Table 5 shows that PuriFlow, leveraging only off-the-shelf models, surpasses adversarial training methods in both standard and robust accuracy for ResNet-50 and DeiT-S. Notably, PuriFlow significantly outperforms diffusion only [38] at the same practical diffusion time $t' = 10$, achieving comparable standard accuracy while enhancing robust accuracy by 11.05% for ResNet-50 and 15.26% for DeiT-S. This improvement signifies the synergistic importance of SR in adversarial purification utilizing diffusion models.

Type	Method	Standard (%)	Robust (%)
ResNet-50		76.13	9.25
AP	Nie et al. [38]	67.79	62.88
	PuriFlow [38]	**74.00**	**68.36**

Table 6. Evaluation of ResNet-50 under a black-box attack using Square Attack ($N = 5000$) within an ℓ_∞-norm ball of radius $\epsilon = 4/255$ on ImageNet-1k. Diffusion time is set to $t' = 150$ for [38] as specified in their study and $t' = 10$ for PuriFlow. Evaluation settings follow those in [38].

Method	Type	Standard (%)	Robust (%)
ResNet-50		76.13	0.00
Wavelet	Denoising	71.15	68.63
TVM	Denoising	57.80	56.96
NL-means	Denoising	64.32	62.67
DRLN×2 [2]	SR	71.66	68.72
EDSR×2 [33]	SR	71.77	68.73
ESRGAN×2 [52]	SR	72.97	68.75
MDSR×2 [33]	**SR**	**73.32**	**69.47**

Table 7. Comparison of SR with other denoising methods as an initialization approach for diffusion [38] ($t' = 90$) under a PGD attack on ImageNet-1k. "TVM" refers to total variance minimization, while "NL-means" denotes the non-local means algorithm.

5.3. Defense on Black-Box Attack

Table 6 shows that PuriFlow, using $t' = 10$—approximately 6.67% of the diffusion time required by diffusion-only method [38]—achieves notable improvements in both standard and robust accuracy for ResNet-50. Specifically, it shows enhancements of 6.21% and 5.48% in each accuracy on ImageNet-1k. These results highlight the synergy of initialization for diffusion, indicating better performance and greater efficiency in this black-box attack.

5.4. Ablation Study

Impact of SR compared to various candidates. PuriFlow can be configured by combining various initialization methods. We test several SR models and traditional denoising techniques by selectively replacing each method used in diffusion-based purification. This evaluation, conducted with PGD targeting the classifier on ImageNet-1k, is shown in Table 7. While Wavelet denoising is practical for PGD, it stands less effective than SR techniques. Among SR models tested, MDSR×2 outperforms DRLN×2, EDSR×2, and ESRGAN×2, achieving the highest standard and robust accuracies. Furthermore, we explore different upsample ratios (×2, ×4, ×8) using ESRGAN. Table 10 indicates that ×2 represents the best balance between efficiency and performance, making it the optimal choice for PuriFlow.

SR	Diffusion	Classifier	Standard (%)	Robust (%)	Wall time (s/img)
		✓	76.13	00.00	≈0.00
✓		✓	73.68	30.70	0.01
	✓	✓	74.52	38.37	1.08
✓	✓	✓	**74.00**	**49.92**	**1.09**

Model	Type	Standard (%)	Robust (%)
ResNet-50	VP-SDE	72.97	68.75
	VE-SDE	73.39	66.51
DeiT-S	VP-SDE	77.29	72.37
	VE-SDE	77.72	69.33

Table 8. Evaluation of PuriFlow under a preprocessor-blind PGD attack on ImageNet-1k selectively employing off-the-shelf models. **Left**: ResNet-50, ESRGAN×2, and VP-SDE are the classifier, SR model, and diffusion [38] model, respectively, with diffusion time $t' = 10$. **Right**: Impact of diffusion models within PuriFlow, evaluated at diffusion time $t' = 90$ using ResNet-50 and DeiT-S on ImageNet-1k.

Off-the-shelf models	ESRGAN×2	EDSR×2	**MDSR×2**
SR (each column)	0.01	0.01	**≈0.00**
Diffusion [38]	1.08	0.81	**0.73**
ResNet-50	≈0.00	≈0.00	**≈0.00**
Total wall time (s/img)	1.09	0.82	**0.73**

Configuration	Criterion	Iterations	Purif. time
Diffusion [38]	SDEs	$t' = 100$	5.27
SR+Diffusion	**SR+SDEs**	$t' = 100$	**0.005+5.27**
ScoreOpt-O [60]	iOSD	$M = 5$	0.51
SR+ScoreOpt-O	**SR+iOSD**	$M = 5$	**0.005+0.51**

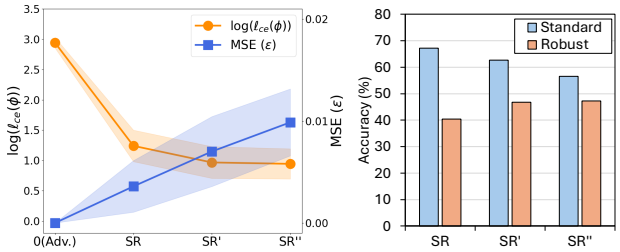
Table 9. **Top**: Wall time for PuriFlow with different SR models. **Bottom**: Examples of purification times (s/img) for configurations using WRN-70-16 [38] and WRN-28-10 [60] on CIFAR-10.

Ratio	Standard (%)	Robust (%)	FLOPs	Params
×8	71.26%	66.80%	1110G	16.7M
×4	72.41%	67.32%	899G	16.7M
×2	**72.97%**	**68.75%**	**224G**	**16.7M**

Table 10. Impact of SR upsampling ratios from ESRGAN×2 combined with diffusion [38] ($t' = 90$) using ResNet-50 on ImageNet-1k. FLOPs represent computational complexity [23], and Params refer to the number of trainable parameters.

Off-the-shelf model adoption and latency analysis. Table 8 shows the enhanced defensive capabilities achieved by integrating off-the-shelf models into PuriFlow's framework. The table highlights the individual and combined effects of SR and diffusion. When used together, SR and diffusion improve robust accuracy by 19.22% and 11.55%, respectively, illustrating their strong synergistic effect on adversarial examples. Although combining SR with diffusion slightly decreases standard accuracy compared to diffusion alone, the robustness gains make this trade-off useful, with a tolerable latency of 1.09 s/img when executed on a single NVIDIA RTX 4090 GPU. Table 9 further confirms that integrating SR models adds minimal latency to the overall purification process. Specifically, SR's execution time is 1,054× faster than the diffusion's, and its overhead is only $1/102$ compared to one-shot denoising methods. This result emphasizes SR's role as an effective and efficient initialization.

Impact of diffusion types. We explore VE-SDE [46], which can act as an alternative to diffusion. Table 8 demonstrates that VP-SDE slightly reduces standard accuracy (approximately 0.42% for ResNet-50 and 0.43% for DeiT-S) but enhances robust accuracy by 2.24% and 3.04%, respectively. Despite the minor drop in standard accuracy, integrating VP-SDE shows overall robustness, synergizing a balanced performance with SR.



(a) Repeated use of SR (b) Trade-off on accuracy

Figure 6. Impact of iterative SR on cross-entropy and MSEs under a PGD ($N = 100$) attack targeting ResNet-50 within an ℓ_∞-norm ball ($\epsilon = 4/255$) on ImageNet-1k. SR, SR', and SR" indicate one, two, and three rounds of SR using ESRGAN×2, respectively.

Iterative applications of SR without diffusion. Figure 6 shows the impact of repeated SR applications for adversarial examples relative to their originals, with changes in MSE and cross-entropy and their impact on standard and robust accuracy. Up to three rounds of SR demonstrate a trade-off: MSE increases while cross-entropy decreases and eventually saturates. Here, robust accuracy rises but converges while standard accuracy decreases. We conjecture that repeated SR may behave as seen in model inversion [35, 57]: Images are generated to minimize cross-entropy loss, but this steady direction moves them farther from the originals.

6. Conclusion

We introduce PuriFlow, a novel purification flow integrating Super-Resolution (SR) with diffusion. While SR alone is less effective at mitigating adversarial noise, its image upsampling and downsampling excel at restoring subtle features captured by early layers. Diffusion alone lacks this ability. However, their mixture synergistically reduces overall feature proximity and cross-entropy. PuriFlow outperforms leading methods in extensive evaluations, demonstrating efficiency for minimal SR overhead.

Acknowledgment

This research was partly supported by the IITP(Institute of Information & Communications Technology Planning & Evaluation)-ITRC(Information Technology Research Center) grant funded by the Korea government(MSIT)(IITP-2025-RS-2023-00258649, 50%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00562437, 40%), and Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University), 10%).

References

[1] Yasin Almalioglu, Kutsev Bengisu Ozyoruk, Abdulkadir Gokce, Kagan Incetan, Guliz Irem Gokceler, Muhammed Ali Simsek, Kivanc Ararat, Richard J Chen, Nicholas J Durr, Faisal Mahmood, et al. Endol2h: deep super-resolution for capsule endoscopy. *IEEE Transactions on Medical Imaging*, 2020. 4

[2] Saeed Anwar and Nick Barnes. Densely residual laplacian super-resolution. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020. 7

[3] 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*, 2018. 1, 6

[4] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International Conference on Machine Learning*, 2018. 6

[5] Mingyuan Bai, Wei Huang, Tenghui Li, Andong Wang, Junbin Gao, Cesar F Caiafa, and Qibin Zhao. Diffusion models demand contrastive guidance for adversarial purification to advance. In *International Conference on Machine Learning*, 2024. 1, 2, 3, 4, 6

[6] Yutong Bai, Jieru Mei, Alan L Yuille, and Cihang Xie. Are transformers more robust than cnns? *Advances in Neural Information Processing Systems*, 2021. 6, 7

[7] Kartikeya Bhardwaj, Dibakar Gope, James Ward, Paul Whatmough, and Danny Loh. Super-efficient super resolution for fast adversarial defense at the edge. In *2022 Design, Automation & Test in Europe Conference & Exhibition*, 2022. 3, 5

[8] Kartikeya Bhardwaj, Milos Milosavljevic, Liam O'Neil, Dibakar Gope, Ramon Matas, Alex Chalfin, Naveen Suda, Lingchuan Meng, and Danny Loh. Collapsible linear blocks for super-efficient super resolution. *Proceedings of Machine Learning and Systems*, 2022. 3

[9] Junyoung Byun, Seungju Cho, Myung-Joon Kwon, Hee-Seon Kim, and Changick Kim. Improving the transferability of targeted adversarial examples through object-based diverse input. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1

[10] Junyoung Byun, Myung-Joon Kwon, Seungju Cho, Yoonji Kim, and Changick Kim. Introducing competition to boost the transferability of targeted adversarial examples through clean feature mixup. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023. 1

[11] Nicholas Carlini, Florian Tramer, Krishnamurthy Dj Dvijotham, Leslie Rice, Mingjie Sun, and J Zico Kolter. (certified!!) adversarial robustness for free! In *International Conference on Learning Representations*, 2023. 1

[12] Hyungjin Chung, Byeongsu Sim, Dohoon Ryu, and Jong Chul Ye. Improving diffusion models for inverse problems using manifold constraints. *Advances in Neural Information Processing Systems*, 2022. 5

[13] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020. 1, 6

[14] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Li Chen, Michael E Kounavis, and Duen Horng Chau. Keeping the bad guys out: Protecting and vaccinating deep learning with jpeg compression. *arXiv preprint arXiv:1705.02900*, 2017. 3

[15] Yilun Du and Igor Mordatch. Implicit generation and modeling with energy based models. *Advances in Neural Information Processing Systems*, 2019. 3

[16] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853*, 2016. 3

[17] Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. 6, 7

[18] Leon A Gatys, Alexander S Ecker, and Matthias Bethge. Image style transfer using convolutional neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 4, 5

[19] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020. 6

[20] Sven Gowal, Sylvestre-Alvise Rebuffi, Olivia Wiles, Florian Stimberg, Dan Andrei Calian, and Timothy A Mann. Improving robustness using generated data. *Advances in Neural Information Processing Systems*, 2021. 6

[21] 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*, 2020. 3

[22] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *International Conference on Learning Representations*, 2018. 3

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2016. 8

[24] Mitch Hill, Jonathan Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. *International Conference on Learning Representations*, 2020. 6

[25] Mitch Hill, Jonathan Craig Mitchell, and Song-Chun Zhu. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations*, 2021. 3

[26] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 2020. 2, 3, 6

[27] Ziyu Jiang, Tianlong Chen, Ting Chen, and Zhangyang Wang. Robust pre-training by adversarial contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 1

[28] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, 2016. 2, 4

[29] Mintong Kang, Dawn Song, and Bo Li. Diffattack: Evasion attacks against diffusion-based adversarial purification. *Advances in Neural Information Processing Systems*, 2024. 6

[30] Minseon Kim, Jihoon Tack, and Sung Ju Hwang. Adversarial self-supervised contrastive learning. *Advances in Neural Information Processing Systems*, 2020. 1

[31] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2017. 4

[32] Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2023. 1, 3, 6

[33] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2017. 6, 7

[34] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1

[35] Alexander Mordvintsev, Christopher Olah, and Mike Tyka. Inceptionism: Going deeper into neural networks. *Google Research Blog*, 2015. 8

[36] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 2020. 3

[37] Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, 2021. 2

[38] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. 1, 2, 3, 4, 5, 6, 7, 8

[39] Tianyu Pang, Min Lin, Xiao Yang, Jun Zhu, and Shuicheng Yan. Robustness and accuracy could be reconcilable by (proper) definition. In *International Conference on Machine Learning*, 2022. 6

[40] Aaditya Prakash, Nick Moran, Solomon Garber, Antonella DiLillo, and James Storer. Deflecting adversarial attacks with pixel deflection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018. 3

[41] Sylvestre-Alvise Rebuffi, Sven Gowal, Dan A Calian, Florian Stimberg, Olivia Wiles, and Timothy Mann. Fixing data augmentation to improve adversarial robustness. *arXiv preprint arXiv:2103.01946*, 2021. 6

[42] Hadi Salman, Andrew Ilyas, Logan Engstrom, Ashish Kapoor, and Aleksander Madry. Do adversarially robust imagenet models transfer better? *Advances in Neural Information Processing Systems*, 2020. 6, 7

[43] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-GAN: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 3

[44] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2017. 2

[45] Taizhang Shang, Qiuju Dai, Shengchen Zhu, Tong Yang, and Yandong Guo. Perceptual extreme super-resolution network with receptive field block. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2020. 4

[46] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 3, 8

[47] Yang Song, Taesup Kim, Sebastian Nowozin, Stefano Ermon, and Nate Kushman. Pixeldefend: Leveraging generative models to understand and defend against adversarial examples. In *International Conference on Learning Representations*, 2018. 3, 6

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *International Conference on Learning Representations*, 2021. 3, 4

[49] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021. 3

[50] Soobin Um, Suhyeon Lee, and Jong Chul Ye. Don't play favorites: Minority guidance for diffusion models. In *International Conference on Learning Representations*, 2024. 5

[51] Pascal Vincent. A connection between score matching and denoising autoencoders. *Neural Computation*, 2011. 3

[52] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In

European Conference on Computer Vision Workshops, 2018. 4, 6, 7

[53] Eric Wong, Leslie Rice, and J. Zico Kolter. Fast is better than free: Revisiting adversarial training. In *International Conference on Learning Representations*, 2020. 6, 7

[54] Fei Wu, Wenxue Yang, Limin Xiao, and Jinbin Zhu. Adaptive wiener filter and natural noise to eliminate adversarial perturbation. *Electronics*, 2020. 3

[55] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L Yuille, and Kaiming He. Feature denoising for improving adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019. 3

[56] Yuzhe Yang, Guo Zhang, Dina Katabi, and Zhi Xu. Me-net: Towards effective adversarial robustness with matrix estimation. In *International Conference on Machine Learning*, 2019. 6

[57] Hongxu Yin, Pavlo Molchanov, Jose M Alvarez, Zhizhong Li, Arun Mallya, Derek Hoiem, Niraj K Jha, and Jan Kautz. Dreaming to distill: Data-free knowledge transfer via deepinversion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020. 8

[58] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, 2021. 1, 3, 6

[59] MD Zeiler. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*, 2014. 4

[60] Boya Zhang, Weijian Luo, and Zhihua Zhang. Enhancing adversarial robustness via score-based optimization. *Advances in Neural Information Processing Systems*, 2023. 2, 3, 5, 6, 8