PMF-PURIFY: UNIFYING ACCURACY AND EFFICIENCY IN SINGLE-STEP PIXEL-SPACE ADVERSARIAL PURIFICATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep neural networks remain vulnerable to imperceptible adversarial perturbations. Adversarial purification provides a robust, plug-and-play defense by restoring perturbed inputs prior to classification without modifying the target model. While diffusion models offer powerful natural-image priors for this task, their iterative reverse process severely bottlenecks inference speed. Current acceleration methods largely attempt to compress these diffusion trajectories via distillation or guided few-step denoising. We argue for a paradigm shift: because adversarial examples inherently retain the vast majority of their original visual semantics, progressively regenerating them through a multi-step trajectory is fundamentally unnecessary. In this work, we introduce PMF-PURIFY, a single-step, latent-free adversarial defense operating in pixel space. Building on Pixel MeanFlow (pMF), PMF-PURIFY bypasses iterative sampling by formulating purification as a direct endpoint prediction. By utilizing a parameter-efficient LoRA adaptation—keeping the base generative prior and downstream classifier entirely frozen—it maps a perturbed intermediate state directly to the natural-image manifold in exactly one network evaluation ($\mathrm{NFE} = 1$). Under the rigorous ImageNet-1K/ResNet-152 evaluation protocol, PMF-PURIFY achieves a state-of-the-art average robust accuracy of 78.6% across 11 diverse attacks. Strikingly, it accomplishes this with an inference latency of merely 3.72 ms per image, successfully unifying top-tier adversarial robustness with single-digit millisecond efficiency.

1 INTRODUCTION

Deep neural networks achieve strong performance on large-scale visual recognition, yet they remain vulnerable to imperceptible adversarial perturbations. By optimizing subtle deviations against a model's decision boundary, attackers can induce incorrect predictions while leaving the visual content nearly unchanged (Goodfellow et al., 2015; Madry et al., 2018). While adversarial training improves robustness, it typically requires costly retraining of the target model (Madry et al., 2018). Adversarial purification instead operates as a standalone input-level defense, restoring perturbed inputs before classification without modifying the downstream classifier.

Existing purification methods broadly follow two directions. Lightweight approaches use filtering, compression, or super-resolution to suppress attack-sensitive components at a low computational cost (Jia et al., 2019; Liao et al., 2018; Mustafa et al., 2020). Generative approaches, conversely, restore inputs using strong natural-image priors. For instance, DiffPure adds noise to an adversarial input and recovers a natural image via a reverse stochastic differential equation (Nie et al., 2022). Subsequent works improve this paradigm through regional heterogeneity, frequency structure, or generative trajectories (Sun et al., 2025; Pei et al., 2025a;b; Song et al., 2024). While providing robust restoration, tracking the reverse diffusion process typically requires repeated evaluations of a heavy generative network.

Recent methods attempt to shorten the diffusion path through consistency distillation or guided few-step denoising (Lei et al., 2025; Li et al., 2025). However, their computation remains anchored in compressing or guiding a multi-step trajectory. This raises a fundamental discrepancy: general image generation must construct semantic content from random noise, whereas an adversarial example

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

already retains most of the original texture, structure, and semantics. Therefore, purification does not inherently require regenerating an entire image. It can instead recover the natural image directly from the perturbed state. This observation motivates our central question: *can the direct image-prediction capability of a generative prior purify an adversarial input in a single network evaluation?*

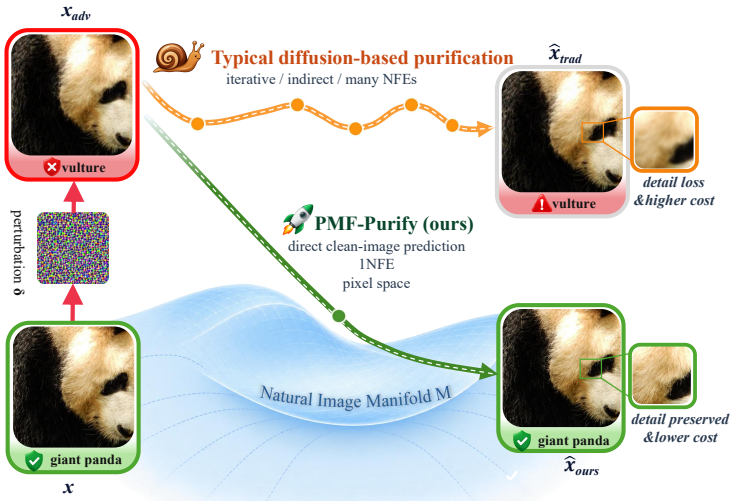

Figure 1: **Motivation for PMF-PURIFY.** A small perturbation δ moves a clean image x to a perturbed state x_{adv} near the natural-image manifold $\mathcal{M}$. Conventional purification iteratively recovers the image along a generative trajectory ($\hat{x}_{\mathrm{trad}}$). In contrast, PMF-PURIFY utilizes the retained semantics and direct pMF prediction to obtain $\hat{x}_{\mathrm{ours}}$ in exactly one network evaluation.

Recent pixel-space generative models make this approach feasible. Pixel MeanFlow (pMF) combines direct x-prediction with a one-step generation mechanism to achieve latent-free image synthesis in the original pixel space (Li & He, 2026; Geng et al., 2025; Lu et al., 2026). As illustrated in Figure 1, instead of distilling a multi-step teacher trajectory, we cast adversarial recovery as an image-endpoint prediction native to pMF. Because the perturbed input preserves semantic integrity, a LoRA-adapted pMF can directly predict the purified natural-image endpoint from an input-dependent intermediate state.

To this end, we introduce PMF-PURIFY, unifying high accuracy and extreme efficiency in single-step pixel-space adversarial purification. Our main contributions are threefold:

- **Single-Step Generative Purification:** We formulate adversarial defense as a direct endpoint prediction using a pretrained pMF model. It restores images in one forward pass with no latent-space encoding/decoding or iterative sampling, achieving a generative-network cost of $\mathrm{NFE} = 1$.
- **Parameter-Efficient Adaptation:** To accommodate attack-induced shifts without compromising the base generative prior, we introduce a parameter-efficient adversarial adaptation strategy. We train a small set of LoRA parameters while keeping the base pMF and the downstream classifier completely frozen.
- **State-of-the-Art Efficiency and Accuracy:** Under the ImageNet-1K/ResNet-152 evaluation protocol, PMF-PURIFY achieves an average purification accuracy of 78.6% across 11 fixed attack settings—outperforming all compared baselines—at a fraction of the computational cost (3.72 ms latency per image). Extensive adaptive white-box attacks and cross-classifier transfer experiments further validate its robustness.

2 RELATED WORK

2.1 GENERATIVE AND DIFFUSION-BASED ADVERSARIAL PURIFICATION

Generative models were used for adversarial purification before diffusion models. Early methods restored perturbed inputs with reconstruction networks, generative adversarial networks, or autoregressive models (Meng & Chen, 2017; Samangouei et al., 2018; Song et al., 2018). These studies established the central idea of generative purification: learning the natural data distribution to suppress attack-induced shifts in input space without modifying the downstream classifier.

Score-based generative models and diffusion models provide stronger models of the natural-image distribution (Yoon et al., 2021; Nie et al., 2022). DiffPure diffuses an adversarial example to a

selected time step and then recovers the image through a reverse stochastic differential equation. Its analysis attributes the suppression of local adversarial structure to the contraction of distributional distance under forward noising (Nie et al., 2022). Later methods improve purification through noise-strength selection, regional heterogeneity, frequency structure, trajectory guidance, and low-rank representations (Sun et al., 2025; Pei et al., 2025a;b; Song et al., 2024; Zollicoffer et al., 2025). Related work also studies adaptive attacks and gradient-based evaluation for diffusion purification (Lee & Kim, 2023). Although these methods use different noising and recovery strategies, they generally require multiple evaluations of the generative network.

2.2 Efficient Adversarial Purification

Progressive distillation and Consistency Models compress long generative trajectories into a small number of network evaluations through distillation and consistency constraints, respectively (Salimans & Ho, 2022; Song et al., 2023). For adversarial purification, InstantPure compresses a multi-step diffusion teacher into a single neural function evaluation through adversarial consistency distillation (Lei et al., 2025). OSGD instead uses an initial denoising result to guide a second single-step purification pass (Li et al., 2025). Other work reduces latency on mobile devices with lightweight generative components (Khalili et al., 2024). These approaches rely on teacher-trajectory distillation, additional guidance, or model compression. The single-step recovery of PMF-Purify instead follows directly from the pMF image-endpoint parameterization and requires neither a multi-step teacher nor a second guided denoising pass.

Non-diffusion defenses shorten the input-processing path through compression, reconstruction, semantic denoising, or super-resolution (Jia et al., 2019; Liao et al., 2018; Mustafa et al., 2020). Recent methods also use local implicit functions, differences in model responses, or combinations of noise and filtering for low-latency input restoration (Ho & Vasconcelos, 2022; Guesmi & Shafique, 2026; Stalder et al., 2026). Their absolute computational cost is low, but their input transformations differ from generative purification. They therefore provide complementary references for evaluating accuracy and efficiency.

2.3 One-Step Generation and Direct Image Prediction

The data manifold hypothesis assumes that natural images in a high-dimensional pixel space concentrate near a structure of much lower intrinsic dimension (Chapelle et al., 2006). Work on non-robust features further suggests that classifiers can rely on features that are predictive of the target but insensitive to human perception. Small perturbations can therefore alter a model prediction without visibly damaging the main semantics of an image (Ilyas et al., 2019). These observations support viewing an adversarial input as a perturbed state that retains most of the original content.

Pixel MeanFlow connects direct image prediction to average-velocity learning through $x \to u \to v$ (Lu et al., 2026). PMF-Purify adapts this image-endpoint parameterization to adversarially perturbed inputs for single-step purification.

3 Method

3.1 Problem Formulation and Overview

Let $x \in [0,1]^{H\times W\times 3}$ be a clean image, $y \in \{1,\dots,C\}$ its ground-truth label, and $f : \mathbb{R}^{H\times W\times 3} \to \mathbb{R}^C$ a fixed downstream classifier. We define $\Delta_p(\epsilon_{\text{atk}}) = \{\delta \in \mathbb{R}^{H\times W\times 3} : \|\delta\|_p \le \epsilon_{\text{atk}}\}$. The attacker constructs $x_{\text{adv}} = \text{clip}_{[0,1]}(x+\delta)$ for $\delta \in \Delta_p(\epsilon_{\text{atk}})$, where $\text{clip}_{[0,1]}$ restricts pixel values to $[0,1]$.

We place a pixel-space purifier P_ϕ before the classifier, where ϕ denotes the trainable adaptation parameters. Given a purification strength s and a noise sample ε, the purifier produces $\hat{x} = P_\phi(x_{\text{adv}}; s, \varepsilon)$, and the downstream classifier predicts $\hat{y} = \arg\max_{k\in\{1,\dots,C\}} f_k(\hat{x})$.

PMF-Purify formulates adversarial purification as a single pMF image-endpoint prediction (Figure 2). It constructs an intermediate state from the perturbed input and Gaussian noise, then uses a LoRA-adapted pMF to predict the purified image directly in pixel space. Training updates only

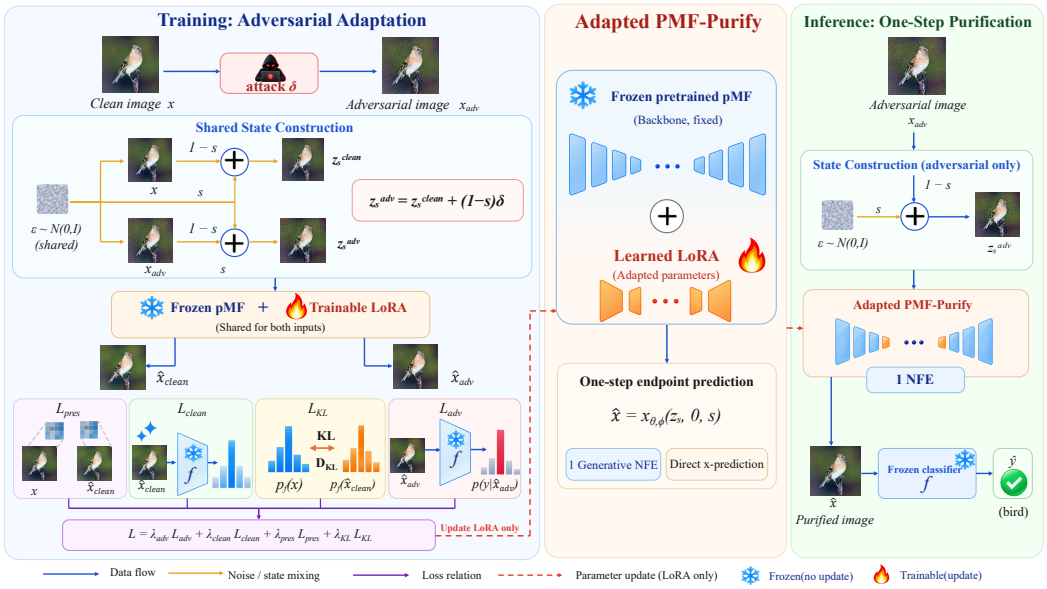

Figure 2: Overview of the training and inference pipeline of PMF-PURIFY. During training, clean and adversarial inputs are mapped to intermediate states and restored by a shared pMF image predictor. The pMF backbone and downstream classifier remain frozen, while only the LoRA parameters are updated. At inference, one pMF endpoint prediction produces the purified image for classification.

the LoRA parameters to account for attack-induced shifts, while the pretrained pMF backbone and downstream classifier remain frozen. Inference requires one generative-network evaluation.

3.2 SINGLE-STEP PURIFICATION VIA DIRECT PMF IMAGE PREDICTION

Following pMF (Lu et al., 2026), we use the linear probability path $z_t = (1-t)x + t\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$ and $t \in [0, 1]$. The endpoints $t = 0$ and $t = 1$ correspond to data and noise, respectively. At purification strength s, we construct an input-dependent intermediate state for any $a \in [0, 1]^{H\times W\times 3}$ as

$$\mathcal{S}_s(a; \varepsilon) = (1-s)a + s\varepsilon, \qquad s \in (0, 1). \tag{1}$$

Conditioned on a, this state follows $\mathcal{N}((1-s)a, s^2 I)$, so s controls the relative amount of input information and noise.

Clean and adversarial inputs share the same ε during training. Let $z_s^{\text{clean}} = \mathcal{S}_s(x; \varepsilon)$, $z_s^{\text{adv}} = \mathcal{S}_s(x_{\text{adv}}; \varepsilon)$, and $\delta_{\text{eff}} = x_{\text{adv}} - x$. Since clipping does not increase the perturbation norm,

$$z_s^{\text{adv}} - z_s^{\text{clean}} = (1-s)\delta_{\text{eff}}, \qquad \|z_s^{\text{adv}} - z_s^{\text{clean}}\|_p \le (1-s)\epsilon_{\text{atk}}. \tag{2}$$

Increasing s therefore contracts the attack-induced state shift but retains less input information, defining the main purification trade-off.

pMF parameterizes image prediction through the average-velocity field u as $x(z_t, r, t) = z_t - t\,u(z_t, r, t)$. At the data endpoint $r = 0$, the average velocity is $u(z_t, 0, t) = (z_t - z_0)/t$, so $x(z_t, 0, t) = z_0$. This endpoint identity motivates direct image prediction from the intermediate state. We define

$$\boxed{P_\phi(a; s, \varepsilon) \triangleq x_{\theta,\phi}(\mathcal{S}_s(a; \varepsilon), 0, s)}, \tag{3}$$

where θ is the frozen pretrained pMF parameter set and ϕ contains the trainable LoRA parameters. For $a = x_{\text{adv}}$, equation (3) predicts the purified image directly in pixel space with one network evaluation. Appendix A.1 gives the image-field derivation.

3.3 Parameter-Efficient Adversarial Adaptation

The intermediate states in equation (1) match the pMF pretraining interpolation but contain the attack-induced shift in equation (2). We adapt the endpoint predictor by inserting LoRA parameters ϕ into selected attention projections (Hu et al., 2022), while keeping the pretrained pMF parameters θ and downstream classifier f frozen. For each adapted projection, the effective weight is

$$W_\ell^{\text{eff}} = W_\ell + \Delta W_\ell, \qquad \Delta W_\ell = \frac{\alpha}{r_{\text{L}}} B_\ell A_\ell, \tag{4}$$

where r_{L} is the LoRA rank and α is its scaling factor. Only A_ℓ and B_ℓ are trained; Appendix A.2 specifies their dimensions and trainable parameter count.

For a clean sample (x, y) and its adversarial counterpart x_{adv}, both purification paths use the same noise ε and training strength s_{tr}. We denote their outputs by $\hat{x}_{\text{clean}} = P_\phi(x; s_{\text{tr}}, \varepsilon)$ and $\hat{x}_{\text{adv}} = P_\phi(x_{\text{adv}}; s_{\text{tr}}, \varepsilon)$. We optimize four complementary terms:

$$\begin{aligned} \mathcal{L}_{\text{adv}} &= \text{CE}(f(\hat{x}_{\text{adv}}), y)\,, & \mathcal{L}_{\text{clean}} &= \text{CE}(f(\hat{x}_{\text{clean}}), y)\,, \\ \mathcal{L}_{\text{pres}} &= \rho_{\text{char}}(\hat{x}_{\text{clean}} - x)\,, & \mathcal{L}_{\text{KL}} &= D_{\text{KL}}(\text{sg}[p_f(x)] \,\|\, p_f(\hat{x}_{\text{clean}}))\,. \end{aligned} \tag{5}$$

Here, $p_f(a) = \text{softmax}(f(a))$, $\text{sg}[\cdot]$ denotes stop-gradient, and ρ_{char} is the Charbonnier penalty.

The final training objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{adv}}\mathcal{L}_{\text{adv}} + \lambda_{\text{clean}}\mathcal{L}_{\text{clean}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}}, \tag{6}$$

where all λ values are nonnegative. $\mathcal{L}_{\text{adv}}$ optimizes post-purification classification of adversarial inputs, $\mathcal{L}_{\text{clean}}$ preserves the correct class on clean inputs, $\mathcal{L}_{\text{pres}}$ preserves pixel content, and $\mathcal{L}_{\text{KL}}$ retains class relationships beyond the hard label. Gradients pass through the frozen classifier to the purifier and update only ϕ. Appendix A.2 further specifies the LoRA parameterization and Charbonnier penalty.

3.4 Single-Step Inference and Computational Cost

For the i-th test input x_i^{test}, we generate $\varepsilon_i \sim \mathcal{N}(0, I)$ from a global random seed and the sample index i, and keep this noise fixed throughout evaluation. Let s_{te} be the test-time purification strength. We first construct $z_i = \mathcal{S}_{s_{\text{te}}}(x_i^{\text{test}}; \varepsilon_i)$, obtain the purified image $\hat{x}_i = P_\phi(x_i^{\text{test}}; s_{\text{te}}, \varepsilon_i) = x_{\theta,\phi}(z_i, 0, s_{\text{te}})$, and predict $\hat{y}_i = \arg\max_{k \in \{1,\dots,C\}} f_k(\hat{x}_i)$.

This inference procedure consists of one state construction, one pMF image prediction, and one downstream classification. Let $D = 3HW$ denote the input dimension. State construction uses only elementwise linear operations and has complexity $\mathcal{O}(D)$. Purification then calls the image predictor $x_{\theta,\phi}$ once, so the generative-network cost is $\text{NFE} = 1$. Here, NFE counts only calls to the generative backbone and excludes state construction and the downstream classifier.

Because pMF directly outputs a pixel-space image, the purification path contains no separate VAE encoder or decoder and requires no numerical solution of a reverse SDE or ODE. NFE measures the number of generative-network calls but does not fully capture differences in the cost of a single call across backbones. We therefore report both NFE and measured inference latency using the timing protocol in Section 4.1.

4 Experiments

4.1 Experimental Setup

Datasets and implementation. We first evaluate PMF-Purify on ImageNet-1K (Deng et al., 2009) with a torchvision-pretrained ResNet-152 (He et al., 2016) as the downstream classifier. The evaluation set contains 1,000 ImageNet validation images correctly classified by the undefended classifier, giving the undefended model an initial accuracy of 100% on this subset. All methods use the same evaluation images, adversarial examples, and downstream classifier.

PMF-Purify is initialized from the official pretrained pMF-B/16 weights (Lu et al., 2026). We freeze the base pMF model and downstream classifier and perform adversarial adaptation only through

LoRA. Training uses 5,000 ImageNet images, with another 256 images reserved for validation. Training examples are generated by three-step PGD with random initialization, and the model is trained for one epoch. We set the training and test purification strengths to $s_{\mathrm{tr}} = 0.22$ and $s_{\mathrm{te}} = 0.28$, respectively. All main experiments use unconditional pMF and a fixed Gaussian noise sample determined by the sample index, with one network evaluation per purification.

Attacks and metrics. The fixed-attack evaluation includes FGSM, BIM, and PGD under an ℓ_∞ constraint with perturbation budgets $\epsilon \in \{2, 4, 8\}/255$. We also evaluate C&W and AA-4 (Croce & Hein, 2020), using $\epsilon = 4/255$ for AA-4. Each attack is first generated against the undefended classifier, after which the same adversarial examples are passed to every defense. Section 4.5 separately reports adaptive white-box evaluation with BPDA-PGD.

We refer to classification accuracy on purified clean images as clean purification accuracy. Average purification accuracy is the equally weighted accuracy over the 11 fixed attack settings and excludes results on clean images.

Baselines and efficiency evaluation. The generative adversarial-purification baselines are DiffPure (Nie et al., 2022), SSNI (Sun et al., 2025), D&C (Pei et al., 2025a), and InstantPure (Lei et al., 2025). The efficient non-diffusion defenses are DISCO (Ho & Vasconcelos, 2022), DRIFT (Guesmi & Shafique, 2026), and Noise+Bilateral (Stalder et al., 2026). Efficiency is measured in BF16 precision as CUDA-synchronized forward latency per image for the purification module. The reported time excludes attack generation and downstream-classifier inference.

Table 1: **Purification accuracy (%) under fixed attacks on ImageNet-1K with ResNet-152.** We evaluate 1,000 ImageNet validation images correctly classified by the undefended classifier. All methods share the same adversarial examples. Average denotes the equally weighted mean over 11 attack settings and excludes clean images. The best result in each column is bold.

		FGSM			BIM			PGD					
Method	Clean	2/255	4/255	8/255	2/255	4/255	8/255	2/255	4/255	8/255	C&W	AA-4	Avg.
Diffusion-based adversarial purification													
DiffPure (2022)	90.6	**85.5**	72.3	49.4	**85.0**	79.6	67.0	82.4	**81.2**	73.2	86.0	82.3	76.7
SSNI (2025)	92.1	79.5	63.6	44.0	84.1	79.2	62.0	82.5	80.4	69.3	87.7	82.0	74.0
D&C-50 (2025a)	90.3	74.4	52.0	34.5	83.0	66.7	29.9	**82.6**	76.0	41.1	**89.7**	82.5	64.8
InstantPure (2025)	89.2	77.6	68.8	59.0	82.4	74.3	67.0	80.5	81.1	69.0	87.2	81.0	75.3
Efficient non-diffusion defenses													
DRIFT (2026)	85.0	72.0	62.5	53.0	75.8	67.6	58.2	75.0	72.5	61.0	82.8	75.0	68.7
N&B+FT (2026)	79.5	75.3	72.4	62.6	77.0	75.3	69.2	77.5	76.3	70.8	78.0	75.6	73.6
DISCO (2022)	**93.0**	66.8	61.6	14.7	69.0	54.5	15.3	67.7	63.2	32.7	85.0	72.4	54.8
PMF-PURIFY (Ours)	88.9	80.4	**75.6**	**65.8**	81.0	**79.7**	**75.0**	81.8	80.0	**75.4**	86.7	**83.0**	**78.6**

4.2 MAIN RESULTS ON IMAGENET-1K

Purification accuracy. Table 1 shows that PMF-PURIFY achieves the highest average purification accuracy of 78.6% across 11 attack settings, 1.9 percentage points above DiffPure. At $\epsilon = 8/255$, it exceeds the strongest baseline by 3.2, 5.8, and 2.2 percentage points under FGSM, BIM, and PGD, respectively. From $\epsilon = 2/255$ to $8/255$, its accuracy drops by 14.6, 6.0, and 6.4 points, compared with 36.1, 18.0, and 9.2 points for DiffPure. The smaller drops across single-step and iterative attacks are consistent with the attack-shift contraction in equation (2) and adversarial LoRA adaptation.

Inference efficiency. Figure 3 compares average purification accuracy and purification-module latency under the unified timing protocol. With one generative-network evaluation, PMF-PURIFY reaches 78.6% average purification accuracy at 3.72 ms per image. Relative to the sub-millisecond N&B+FT and DRIFT defenses, it gains 5.0 and 9.9 percentage points in accuracy, respectively. InstantPure takes 247.92 ms per image, while the remaining diffusion-based methods take seconds. PMF-PURIFY thus combines the highest average accuracy in this comparison with single-digit-millisecond generative purification.

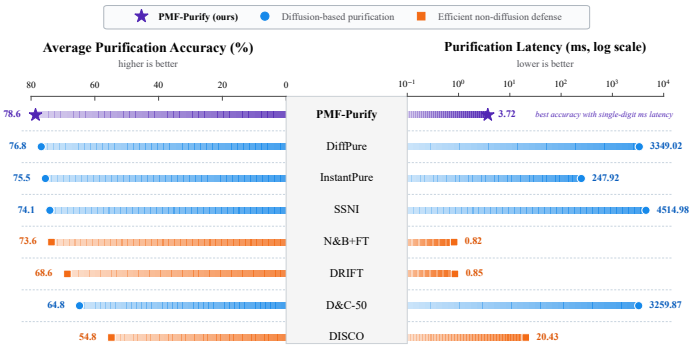

Figure 3: **Accuracy–latency comparison under fixed attacks.** The left panel reports average purification accuracy over 11 attack settings (higher is better), and the right panel reports purification-module latency per image on a logarithmic scale (lower is better). The star denotes our PMF-PURIFY.

4.3 CROSS-CLASSIFIER TRANSFER

We test whether the LoRA parameters adapted with ResNet-152 can be reused with WideResNet-50-2 (Zagoruyko & Komodakis, 2016) without retraining the purifier. For WideResNet-50-2, we select ImageNet validation images correctly classified by the undefended model and generate PGD-4, PGD-8, and AA-4 adversarial examples against it. All defenses receive the same examples and retain their original purification parameters. Table 2 reports results for both downstream classifiers. PMF-PURIFY leads under PGD-8 and AA-4 with ResNet-152 and under all three attack settings with WideResNet-50-2. The latter results support reuse of the adapted purifier with a different downstream classifier.

Table 2: **Cross-classifier transfer on ImageNet (%).** Defenses evaluated with WideResNet-50-2 are not retrained for the new classifier. The classifiers use different evaluation subsets, so results are compared only within each group. The best result in each group and column is bold.

Classifier	Method	Clean	PGD-4	PGD-8	AA-4
ResNet-152	InstantPure	**89.2**	**81.0**	69.0	81.0
	DRIFT	85.0	72.3	61.0	74.7
	PMF-PURIFY	88.9	80.0	**75.3**	**83.0**
WideResNet-50-2	InstantPure	85.1	80.6	73.4	81.2
	DRIFT	82.2	70.2	57.0	73.4
	PMF-PURIFY	**86.6**	**81.8**	**76.3**	**81.5**

4.4 CROSS-DATASET VALIDATION ON CIFAR-10

We evaluate a separate PMF-PURIFY model matched to CIFAR-10 (Krizhevsky, 2009), with adversarial adaptation performed independently of ImageNet. The pMF weights, LoRA parameters, and purification strength are not transferred between datasets. Using WideResNet-28-10 (Zagoruyko & Komodakis, 2016) as the frozen classifier, we evaluate all 10,000 official test images with $s = 0.06$. PGD-2, PGD-4, and PGD-8 use $\epsilon = 2/255$, $4/255$, and $8/255$, respectively; AA-4 uses $\epsilon = 4/255$. Attacks target the undefended classifier, and every purifier receives the same adversarial examples.

Table 3: **Cross-dataset results on CIFAR-10** with WideResNet-28-10 (%). Average is the equally weighted mean over four fixed attacks and excludes clean images. The best result in each column is bold.

Method	Clean	PGD-2	PGD-4	PGD-8	AA-4	Avg.
DiffPure	88.6	85.2	84.5	81.4	86.2	84.3
SSNI	**90.8**	85.0	84.2	78.5	86.5	83.6
DISCO	88.5	85.2	84.0	**82.0**	83.0	83.6
PMF-PURIFY	89.6	**85.8**	**84.7**	81.8	**86.8**	**84.8**

Table 3 shows that PMF-PURIFY reaches 84.8% average purification accuracy on CIFAR-10, compared with 84.3% for DiffPure, and leads under PGD-2, PGD-4, and AA-4. Thus, the single-step design remains effective when trained and adapted for lower-resolution images.

4.5 EFFECT OF ADVERSARIAL ADAPTATION & ADAPTIVE WHITE-BOX EVALUATION

To separate the pretrained image prior from task-specific adaptation, we compare the unadapted pMF with PMF-PURIFY. As shown in Table 4, the base model already reaches 77.1% average purification accuracy, while LoRA adaptation raises it to 78.6%. Under AA-4, accuracy rises significantly from 71.5% to 83.0%, while clean accuracy remains close to that of the base model.

In the adaptive white-box setting, the attacker has access to the complete purification pipeline and optimizes the post-defense classification loss. We use BPDA-PGD to approximate gradients through the purifier during backpropagation. Under the 50-step BPDA-PGD attacks in Table 5, PMF-PURIFY achieves 41.7% accuracy at $\epsilon = 4/255$ and 14.0% at $\epsilon = 8/255$, the highest results among the compared methods. The leading ranking persists when the attack targets the defended pipeline through BPDA.

Table 4: **Effect of adversarial adaptation** on purification accuracy (%). Average is the equally weighted mean over all 11 settings.

Method	Clean	PGD-8	AA-4	Avg.
Base pMF	88.6	73.9	71.5	77.1
PMF-PURIFY	**88.9**	**75.3**	**83.0**	**78.6**

Table 5: **Adaptive white-box BPDA-PGD** results on ImageNet-1K (%). BPDA-4 and BPDA-8 denote 50-step attacks, respectively.

Method	Clean	BPDA-4	BPDA-8
DiffPure	90.6	38.8	13.6
SSNI	**92.1**	40.5	13.8
InstantPure	89.2	6.8	2.7
PMF-PURIFY	86.5	**41.7**	**14.0**

4.6 PURIFICATION STRENGTH AND NOISE STABILITY

The purification strength s trades off suppression of adversarial shifts against retention of the input: by equation (2), increasing s contracts the state-shift bound by $(1 - s)$ while reducing the input component. We sweep $s \in [0.10, 0.36]$ on 1,000 ImageNet validation images correctly classified by ResNet-152. In Table 6, mean accuracy rises from 63.5% at $s = 0.10$ to 79.0% at $s = 0.30$, then declines, while clean accuracy falls from 91.2% as s increases. We choose $s = 0.28$: relative to $s = 0.30$, it sacrifices 0.4 percentage points in mean accuracy but gains 1.0 point in clean accuracy.

Table 6: **Effect of purification strength** s on clean and average purification accuracy (%) over 11 fixed attacks. The default setting is highlighted in bold.

Strength s	0.10	0.12	0.14	0.16	0.18	0.20	0.22	0.24	0.26	**0.28**	0.30	0.32	0.34	0.36
Clean Acc.	91.2	90.8	90.6	90.4	90.2	90.0	89.8	89.5	89.2	**88.9**	87.9	85.9	84.3	83.0
Avg. Robust	63.5	64.6	65.9	67.3	69.4	70.8	72.5	74.4	76.8	**78.6**	79.0	76.9	74.2	72.0

Fixed noise seed. The main evaluation fixes Gaussian noise by image index. Holding the model, purification strength, and adversarial examples constant, we vary only the noise seed. As shown in Table 7, the standard deviation across five distinct seeds is at most 0.23 percentage points, demonstrating high stability.

Input	Seed 0	Seed 1	Seed 2	Seed 3	Seed 4	Mean $\pm$ Std.
Clean	88.9	88.8	89.0	88.8	89.1	88.92 ± 0.13
PGD-8	75.4	75.5	75.1	75.5	75.0	75.30 ± 0.23
AA-4	83.0	83.1	82.9	83.0	82.8	82.96 ± 0.11

Table 7: **Purification accuracy across five fixed noise seeds (%).** Variation in the initialization noise seed has a minimal effect on the final performance.

4.7 Classifier Feature Distributions and Grad-CAM Responses

We project frozen-classifier features $h(\cdot)$ from clean, adversarial, noisy, and purified images with a jointly fitted, shared PCA matrix. Figure 4 shows adversarial features shifting away from the clean distribution and purified features moving back toward it after one step.

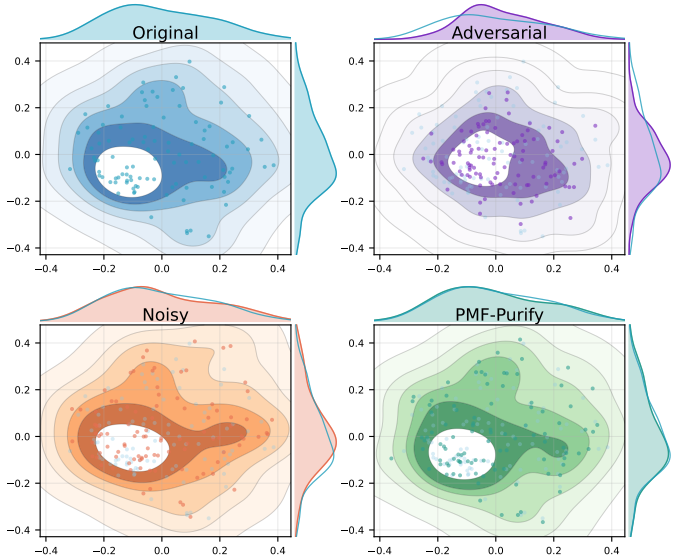

Figure 4: **Classifier feature distributions in a shared PCA space.** All four input states share the same PCA projection matrix. Adversarial features shift away from clean-image features, while purified features effectively move back toward the clean distribution.

Spatial responses. We use Grad-CAM (Selvaraju et al., 2017) to compare classifier responses on clean, adversarial, and purified images. In the examples in Figure 5, adversarial perturbations move the classifier's main response regions. After purification, the responses return to object locations consistent with those for the clean images.

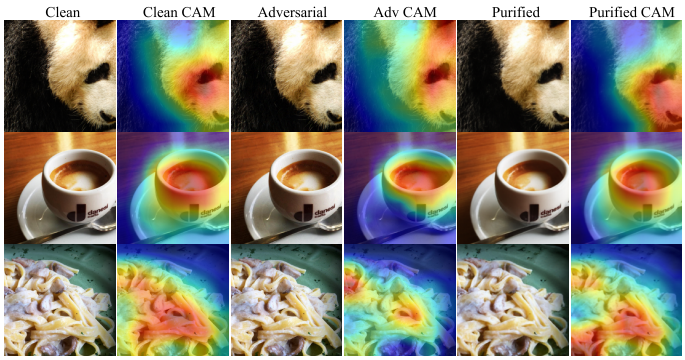

Figure 5: **Grad-CAM visualizations of the downstream classifier.** While adversarial perturbations mislead the classifier's attention, our purification successfully restores the response regions to locations consistent with the clean images.

5 Conclusion

In this work, we introduced PMF-Purify, a single-step, latent-free adversarial defense that unifies state-of-the-art robustness with extreme computational efficiency. By recognizing that adversarial examples inherently retain the vast majority of their original texture and semantics, we challenge the conventional reliance on multi-step generative trajectories. Instead of iteratively tracking a reverse diffusion process or distilling a complex teacher model, we formulate adversarial recovery as a direct image-endpoint prediction. Leveraging a parameter-efficient, LoRA-adapted Pixel MeanFlow (pMF) prior, PMF-Purify restores natural images directly from perturbed states in a single forward pass (NFE $= 1$). On ImageNet-1K with ResNet-152, it achieves the highest average accuracy among the compared methods across 11 fixed attacks (78.6%), with a purification-module latency of 3.72 ms per image. Ablation, transfer, CIFAR-10, and adaptive white-box results further support this single-step alternative to iterative diffusion-based purification.

AI USE STATEMENT

Generative AI assisted with Chinese-to-English translation, language editing, manuscript organization, and identifying related literature. The authors verified all equations, results, citations, and technical claims against the source materials and take responsibility for the paper.

REFERENCES

Olivier Chapelle, Bernhard Schölkopf, and Alexander Zien. *Semi-Supervised Learning*. MIT Press, 2006.

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.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2009.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, J. Zico Kolter, and Kaiming He. Mean Flows for one-step generative modeling. In *Advances in Neural Information Processing Systems*, 2025.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Amira Guesmi and Muhammad Shafique. DRIFT: Divergent response in filtered transformations for robust adversarial defense. In *International Conference on Learning Representations*, 2026.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2016.

Chih-Hui Ho and Nuno Vasconcelos. Adversarial defense with local implicit functions. In *Advances in Neural Information Processing Systems*, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. In *Advances in Neural Information Processing Systems*, 2019.

Xiaojun Jia, Xingxing Wei, Xiaochun Cao, and Hassan Foroosh. ComDefend: An efficient image compression model to defend adversarial examples. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2019.

H. Khalili, S. Park, V. Li, et al. LightPure: Realtime adversarial image purification for mobile devices using diffusion models. In *ACM International Conference on Mobile Computing and Networking*, 2024.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Minjong Lee and Dongwoo Kim. Robust evaluation of diffusion-based adversarial purification. In *IEEE International Conference on Computer Vision*, 2023.

Chun Tong Lei, Hon Ming Yam, Zhongliang Guo, Yifei Qian, and Chun Pong Lau. Instant adversarial purification with adversarial consistency distillation. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025.

Tianhong Li and Kaiming He. Back to basics: Let denoising generative models denoise. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2026.

Y. Li, Z. Li, L. Huang, et al. Adversarial purification with one-step guided diffusion model. *Neural Networks*, 192:107877, 2025.

Fangzhou Liao, Ming Liang, Yinpeng Dong, Tianyu Pang, Xiaolin Hu, and Jun Zhu. Defense against adversarial attacks using high-level representation guided denoiser. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with Pixel Mean Flows. *arXiv preprint arXiv:2601.22158*, 2026.

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.

Dongyu Meng and Hao Chen. MagNet: A two-pronged defense against adversarial examples. In *ACM SIGSAC Conference on Computer and Communications Security*, 2017.

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.

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.

Gaozheng Pei, Shaojie Lyu, Gong Chen, Ke Ma, Qianqian Xu, Yingfei Sun, and Qingming Huang. Divide and conquer: Heterogeneous noise integration for diffusion-based adversarial purification. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2025a.

Gaozheng Pei, Ke Ma, Yingfei Sun, Qianqian Xu, and Qingming Huang. Diffusion-based adversarial purification from the perspective of the frequency domain. In *International Conference on Machine Learning*, 2025b.

Tim Salimans and Jonathan Ho. Progressive distillation for fast sampling of diffusion models. In *International Conference on Learning Representations*, 2022.

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

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 *IEEE International Conference on Computer Vision*, 2017.

K. Song, H. Lai, Y. Pan, and J. Yin. MimicDiffusion: Purifying adversarial perturbation via mimicking clean diffusion model. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2024.

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.

Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, 2023.

Nicolas Stalder, Benjamin F. Grewe, Matteo Saponati, and Pau Vilimelis Aceituno. A combination of noise and bilateral filters achieve supralinear and scalable adversarial robustness in CNNs. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2026.

Yuhao Sun, Jiacheng Zhang, Zesheng Ye, Chaowei Xiao, and Feng Liu. Sample-specific noise injection for diffusion-based adversarial purification. In *International Conference on Machine Learning*, 2025.

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

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference*, 2016.

G. Zollicoffer, M. N. Vu, B. Nebgen, et al. LoRID: Low-rank iterative diffusion for adversarial purification. In *AAAI Conference on Artificial Intelligence*, 2025.

A Pixel MeanFlow Parameterization and Adaptation Details

A.1 Image-Field Parameterization

Along an underlying ODE trajectory, MeanFlow defines the average velocity over an interval $[r, t]$ as follows (Geng et al., 2025):

$$u(z_t, r, t) = \frac{1}{t-r} \int_r^t v(z_\tau, \tau)\, d\tau, \qquad 0 \le r < t \le 1, \tag{7}$$

where $v(z_\tau, \tau)$ denotes the instantaneous velocity of the trajectory at time τ. As $r \to t$, the average velocity continuously reduces to the instantaneous velocity, giving $u(z_t, t, t) = v(z_t, t)$.

pMF combines the average-velocity representation of MeanFlow with the direct image prediction of JiT (Li & He, 2026; Lu et al., 2026). For the true average velocity field u, pMF defines a generalized denoised-image field

$$x(z_t, r, t) \triangleq z_t - t\, u(z_t, r, t). \tag{8}$$

Instead of predicting the average velocity directly, pMF uses a network to predict an image-valued output $x_\theta(z_t, r, t)$ and converts it to the corresponding average-velocity parameterization through

$$x_\theta(z_t, r, t) := \mathrm{net}_\theta(z_t, r, t), \qquad u_\theta(z_t, r, t) = \frac{z_t - x_\theta(z_t, r, t)}{t}, \quad t > 0. \tag{9}$$

When the target endpoint is the data endpoint $r = 0$, substituting it into equation (7) and using equation (8) gives

$$u(z_t, 0, t) = \frac{z_t - z_0}{t}, \qquad x(z_t, 0, t) = z_0, \tag{10}$$

where z_0 is the data endpoint of the corresponding ODE trajectory and follows p_{data}. Thus, at $r = 0$, the pMF image field represents the data-endpoint image of the trajectory. We initialize the image predictor from a pretrained pMF and adapt it to intermediate states constructed from perturbed inputs.

A.2 Low-Rank Adaptation and Loss Details

Let $\mathcal{L}$ be the set of attention projection layers equipped with LoRA. For the frozen weight $W_\ell \in \mathbb{R}^{d_{\mathrm{out}} \times d_{\mathrm{in}}}$ of layer $\ell \in \mathcal{L}$, equation (4) uses $A_\ell \in \mathbb{R}^{r_{\mathrm{L}} \times d_{\mathrm{in}}}$ and $B_\ell \in \mathbb{R}^{d_{\mathrm{out}} \times r_{\mathrm{L}}}$. Since $\mathrm{rank}(\Delta W_\ell) \le r_{\mathrm{L}}$, each projection layer introduces $r_{\mathrm{L}}(d_{\mathrm{in}} + d_{\mathrm{out}})$ trainable parameters while the base weight W_ℓ remains fixed. The complete trainable parameter set is $\phi = \{A_\ell, B_\ell\}_{\ell \in \mathcal{L}}$. The adapted $x_{\theta,\phi}$ retains the direct image-prediction form of pMF and leaves the single-step inference path in equation (3) unchanged.

In equation (5), $p_f(a) = \mathrm{softmax}(f(a))$ is the class-probability distribution produced by the frozen classifier for input a, and $\mathrm{sg}[\cdot]$ denotes stop-gradient. The Charbonnier function is

$$\rho_{\mathrm{char}}(q) = D^{-1} \sum_{i=1}^{D} \sqrt{q_i^2 + \epsilon_{\mathrm{char}}^2}, \qquad D = 3HW, \quad \epsilon_{\mathrm{char}} > 0. \tag{11}$$

Sharing the same noise sample between clean and adversarial purification removes variation from noise sampling and focuses optimization on the attack-induced state shift. The classifier remains frozen, but gradients propagate through $f(\hat{x}_{\mathrm{adv}})$ and $f(\hat{x}_{\mathrm{clean}})$ to the purifier and update only the LoRA parameters ϕ.