Instant Adversarial Purification with Adversarial Consistency Distillation

Chun Tong Lei[1], Hon Ming Yam[1], Zhongliang Guo[2], Yifei Qian[3], Chun Pong Lau[1*]
[1]City University of Hong Kong [2]University of St Andrews [3]University of Nottingham
{ctlei2, hmyam4, cplau27}@cityu.edu.hk, zg34@st-andrews.ac.uk, yifei.qian@nottingham.ac.uk

Abstract

Neural networks have revolutionized numerous fields with their exceptional performance, yet they remain susceptible to adversarial attacks through subtle perturbations. While diffusion-based purification methods like DiffPure offer promising defense mechanisms, their computational overhead presents a significant practical limitation. In this paper, we introduce One Step Control Purification (OSCP), a novel defense framework that achieves robust adversarial purification in a single Neural Function Evaluation (NFE) within diffusion models. We propose Gaussian Adversarial Noise Distillation (GAND) as the distillation objective and Controlled Adversarial Purification (CAP) as the inference pipeline, which makes OSCP demonstrate remarkable efficiency while maintaining defense efficacy. Our proposed GAND addresses a fundamental tension between consistency distillation and adversarial perturbation, bridging the gap between natural and adversarial manifolds in the latent space, while remaining computationally efficient through Parameter-Efficient Fine-Tuning (PEFT) methods such as LoRA, eliminating the high computational budget request from full parameter fine-tuning. The CAP guides the purification process through the unlearnable edge detection operator calculated by the input image as an extra prompt, effectively preventing the purified images from deviating from their original appearance when large purification steps are used. Our experimental results on ImageNet showcase OSCP's superior performance, achieving a 74.19% defense success rate with merely 0.1s per purification — a 100-fold speedup compared to conventional approaches.

1. Introduction

Deep Neural Networks (DNNs) have fundamentally transformed the landscape of computer vision, achieving unprecedented performance across diverse applications. Despite these extraordinary achievements, a critical vulnerability lurks beneath their impressive performance, one that has emerged as a paramount concern in the research community [14, 23, 31].

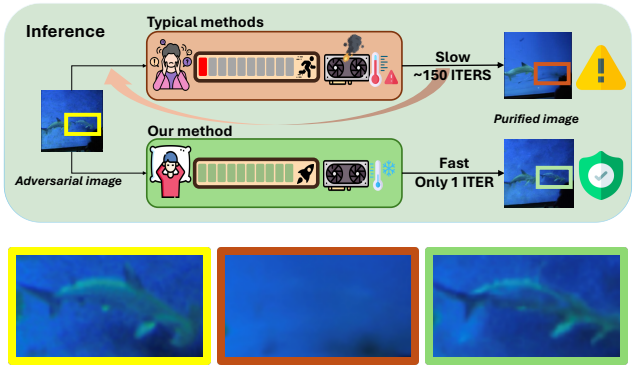


Figure 1. Comparison between existing methods and our proposed approach. Our method achieves superior performance with just a single inference step, significantly reducing computational cost. Through adversarial noise-adapted U-Net fine-tuning, we demonstrate better detail preservation after denoising, as evident in the zoomed-in regions for circles (bottom). This makes OSCP an efficient and practical solution for adversarial purification.

The Achilles' heel of these sophisticated neural architectures lies in their remarkable vulnerability to adversarial attacks [24, 43]. Through the injection of meticulously engineered perturbations that are imperceptible to human observers yet devastating to model performance, adversaries can systematically manipulate DNNs into catastrophic misclassifications [15]. While this vulnerability has found constructive applications in safeguarding against AI misuse, such as protecting intellectual property and preventing unauthorized model exploitation [13], its implications extend far beyond these beneficial uses. The susceptibility to adversarial manipulation poses a critical threat to the deployment of DNNs in high-stakes domains, where system reliability and robustness are not merely desirable but essential for public safety and security [42].

In response to these security challenges, two principal defense paradigms have emerged: adversarial training [22, 46] and adversarial purification [32, 36, 49]. Adversarial training, despite its widespread adoption, exhibits a critical limitation — it requires prior knowledge of attack methods, inherently constraining its effectiveness against

*Corresponding Author.

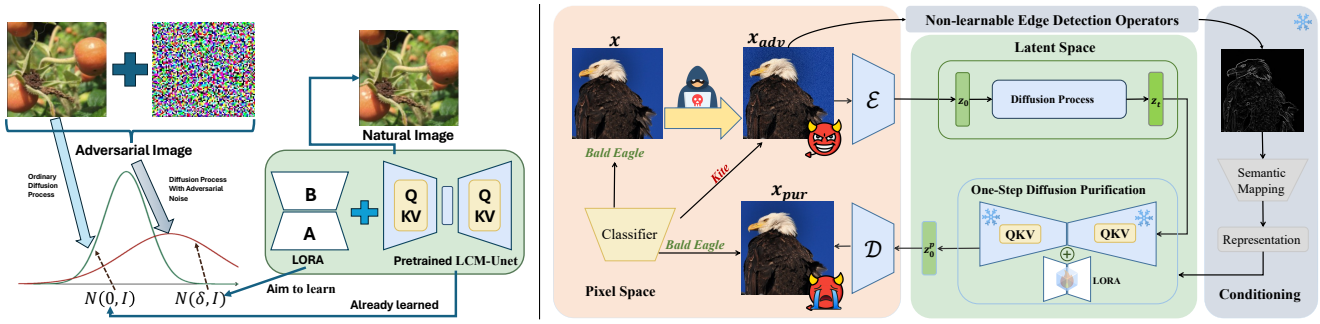


Figure 2. The pipeline of Our proposed OSCP. **(a)** the left figure shows that the adversarial images, which are crafted through intentional attacks, exhibit a shifted distribution after the diffusion process that deviates from the Standard Normal distribution. In response, our proposed GAND can learn to recover the attacked images by modeling this additional adversarial noise with LoRA. **(b)** the right figure illustrates the pipeline that our proposed CAP leverages non-learnable edge detection operators to guide the purification of adversarial samples, avoiding potential inductive bias introduced by neural networks. It is worth noting that our method achieves remarkable performance by just running a single U-Net inference step.

unknown threats. In contrast, adversarial purification offers a more versatile defense framework by focusing on perturbation removal rather than attack-specific training, making it inherently more adaptable to emerging adversarial threats.

Adversarial purification methods, particularly those leveraging diffusion models [17, 39], have demonstrated remarkable effectiveness in neutralizing adversarial attacks by mapping perturbed samples back to the natural distribution [32, 44]. Diffusion models, with their unique denoising capabilities and superior training stability, offer significant advantages over traditional generative approaches such as GANs. Their ability to systematically restore corrupted images through iterative denoising has established them as a compelling framework for adversarial defense.

Despite their effectiveness, diffusion models face a critical limitation: **their multi-step denoising process incurs substantial computational overhead**, making them impractical for real-time defense scenarios. This inherent inefficiency poses a significant challenge in applications demanding rapid response times, such as real-time security systems. Consequently, there is an urgent need for efficient purification methods that can maintain robust defense while minimizing computational costs.

To address these challenges, we introduce One Step Control Purification (OSCP), a novel diffusion-based framework that achieves robust adversarial defense in a single inference step. At its core, OSCP first leverages consistency distillation for fast inference while explicitly addressing the fundamental discrepancy between adversarial and clean sample distributions. This insight leads to our first key innovation, the distillation objective — Gaussian Adversarial Noise Distillation (GAND). As shown in Fig. 1 and Fig. 2 (left), GAND enables efficient purification while maintaining strong defense performance.

Our work also addresses another fundamental challenge in diffusion-based purification — **the loss of semantic information in large-step inference**. While Latent Consistency Models offer acceleration, they often sacrifice image quality, resulting in semantic degradation and blurry outputs. To preserve both efficiency and fidelity, as shown in Fig. 2 (right), we propose Controlled Adversarial Purification (CAP) as the inference pipeline to empower GAND, which integrates ControlNet with non-learnable edge detection operators, enabling high-quality purification while maintaining semantic integrity.

Our proposed OSCP framework, composed of GAND as the distillation objective and CAP as the inference pipeline, achieves state-of-the-art performance with 74.19% robust accuracy on ImageNet while requiring merely 0.1s per purification. This breakthrough in efficiency and effectiveness not only advances the field of adversarial defense but also enables the practical deployment of robust neural networks in real-world, time-sensitive applications.

In summary, our key contributions are as follows:

- We propose GAND, a novel consistency distillation objective for adversarial training on Latent Consistency Models. Our empirical results demonstrate its exceptional robustness and transferability against unknown adversarial attacks.
- We introduce CAP, an innovative inference pipeline that leverages non-learnable edge detection operators to enhance the controllability and semantic preservation of adversarial purification.
- We develop OSCP, an integrated framework combining GAND and CAP, that achieves real-time adversarial purification while maintaining robust defense performance, making diffusion-based defenses practical for time-critical applications.

2. Related Work

2.1. Adversarial training

Adversarial training [53] has established itself as a cornerstone defense strategy against adversarial attacks by incorporating perturbed examples into the training process [21]. Numerous studies have shown its effectiveness [2, 11], showing significant improvements in model robustness against known attack patterns. However, this approach exhibits an inherent limitation: models tend to overfit known attacks, compromising their resilience against novel attack vectors [22]. Recent advances attempt to address this limitation by leveraging diffusion models to generate diverse adversarial samples [12, 46], aiming to enhance generalization and prevent overfitting.

2.2. Adversarial Purification

Adversarial purification has evolved significantly in its approach to defending against attacks through input restoration. Initially pioneered with GAN-based methods [36], the field underwent a paradigm shift with the advent of diffusion models [17, 39, 40]. These diffusion-based approaches have demonstrated superior purification capabilities [32, 44, 49]. Notably, DiffPure [32] simultaneously removes adversarial perturbations and Gaussian noise from the forward process, theoretically justified by the reduced KL divergence between clean and adversarial image distributions. Despite these advances, the computational overhead of multiple denoising steps remains a critical bottleneck [44], limiting practical deployment.

2.3. Diffusion Models

Diffusion models have demonstrated remarkable capabilities across diverse domains, including text-to-image synthesis [34, 35], video generation [3, 18], and 3D content creation [28, 33]. The foundational Denoising Diffusion Probabilistic Models (DDPM) [17] established a two-phase framework: a forward process that gradually applies Gaussian noise following Markov properties, and a reverse process that learns to reconstruct the original image through a reverse Markov chain. However, the sequential nature of this process results in prohibitively long inference times, as computational cost scales linearly with the number of denoising steps.

2.4. Efficient Diffusion Models

To address the computational bottleneck, Consistency Models [41] introduced a novel consistency training paradigm, enabling image generation in significantly fewer steps. This advancement led to the development of Latent Consistency Models (LCM) [29], which further accelerated the generation process. LCM-LoRA [30] enhanced efficiency through Low-Rank Adaptation, a Parameter-Efficient Fine-Tuning (PEFT) [19] approach that substantially reduces computational requirements. Additionally, ControlNet [51] introduced flexible conditioning mechanisms for Stable Diffusion, enabling precise control over generation through various visual guides including edge maps, pose estimation, sketches, and depth information.

3. Preliminaries

3.1. Diffusion model

Denoising Diffusion Probabilistic Models (DDPM) [17] generate images by learning from the reverse Markov chain with Gaussian noise added to the original image. The forward process can be formulated as linear combination of original image $\boldsymbol{x}_0$ and standard Gaussian noise $\boldsymbol{\epsilon}$, $\bar{\alpha}_t$ denoted cumulative product from α_1 to α_t, $\alpha_t = 1 - \beta_t$ for any t, β_t is predefined variance schedule of diffusion process:

$$\boldsymbol{x}_t = \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{1}$$

and the denoising step can be expressed as:

$$\hat{\boldsymbol{x}}_{t-1} = \frac{1}{\sqrt{\alpha_t}}\left(\boldsymbol{x}_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\boldsymbol{\epsilon}_{\boldsymbol{\theta}^*}(\boldsymbol{x}_t, t)\right) + \sqrt{\beta_t}\boldsymbol{\epsilon}, \tag{2}$$

where model parameter $\boldsymbol{\theta}^*$ minimize the loss between actual noise and predict noise:

$$\boldsymbol{\theta}^* = \arg\min_{\theta} \mathbb{E}_{\boldsymbol{x}_0,t,\boldsymbol{\epsilon}}\left[\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t)\|_2^2\right]. \tag{3}$$

3.2. Diffusion-Base Purification (DBP)

DiffPure [32] proposes that diffusion models can remove adversarial noise by performing a sub-process of the normal reverse process ($t = 0$ to $t = T$). By adding predefined t^* ($t^* < T$) of noise to the adversarial image $\boldsymbol{x}_{adv}$, which is formed by the sum of original image $\boldsymbol{x}$ and adversarial noise $\boldsymbol{\delta}$, $\boldsymbol{\delta}$ can be generated by L_p attack [31] or AutoAttack [6]:

$$\boldsymbol{\delta} = \arg\max_{\delta} \mathcal{L}(C(\boldsymbol{x} + \boldsymbol{\delta}), y), \tag{4}$$

where C is the classifier, y is the true label. The forward process of diffusion purification method uses:

$$\boldsymbol{x}(t^*) = \sqrt{\bar{\alpha}(t^*)}\boldsymbol{x} + \sqrt{1-\bar{\alpha}(t^*)}\boldsymbol{\epsilon}, \tag{5}$$

and solve the reverse process of DDPM from time step t^* to 0, to get the purified $\hat{\boldsymbol{x}}_{adv}^0$ that is close to the original image $\boldsymbol{x}$, allowing the classifier to classify the image with the correct label.

3.3. Consistency Model

Diffusion models are known to have long inference time, which limits their usage in the real world. To tackle this situation, Consistency model [41] has been proposed, which makes it possible to generate images within 2 to 4 U-Net inference steps, by distilling a Consistency Model from a pretrianed diffusion model.

Due to the efficiency of latent space models compared

to pixel-based models, Latent Consistency Model [29] has been proposed, making use of pretrained encoder and decoder to transform images from pixel space to latent space. Latent Consistency Model's definition is

$$f_\theta(\boldsymbol{z}_t, t) = f_\theta(\boldsymbol{z}_{t'}, t') \quad \forall t, t' \in [0, T] \tag{6}$$

and f_θ can be parameterized as:

$$f_\theta(\boldsymbol{z}, \boldsymbol{c}, t) = c_{\text{skip}}(t)\boldsymbol{z} + c_{\text{out}}(t)\left(\frac{\boldsymbol{z} - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}, \boldsymbol{c}, t)}{\sqrt{\bar{\alpha}_t}}\right), \tag{7}$$

where $c_{\text{skip}}(t)$ and $c_{\text{out}}(t)$ are differentiable functions such that $c_{\text{skip}}(\epsilon) = 1$ and $c_{\text{out}}(\epsilon) = 0$. Distilled with LoRA is introduced in LCM-LoRA [30], which dramatically reduces training time and computation cost in distillation by hugely reducing trainable parameters.

4. Method

In this paper, we aim to solve the slow purification problem in DBP by leveraging the Latent Consistency Distillation to speed up the purification backbone, enabling single-step adversarial purification.

Recognizing that adversarial noise differs from the Gaussian noise that diffusion models are typically designed to remove, we propose Gaussian Adversarial Noise Distillation (GAND). This novel distillation objective specifically targets adversarial noise, enhancing the purification performance. Our approach builds upon the insight that combining adversarial purification with adversarial training can yield superior results [25], effectively addressing the distinct distributions of Gaussian and adversarial noise.

However, diffusion models tend to produce images deviating from the originals when purification steps t^* are large [44]. To address this, we introduce Controlled Adversarial Purification (CAP), shown in Fig. 2 (right), which utilizes the unlearnable edge detection operator computed on the input to guide the purification process.

4.1. One Step Control Purification

Problem Definition Our goal can be formulated as:

$$\begin{aligned} \boldsymbol{x}_{gt} \simeq \hat{\boldsymbol{x}}^0_{adv} &= \mathcal{D}(f(\mathcal{E}(\boldsymbol{x}_{adv}), t^*)) \\ \text{s.t.} \quad C(\boldsymbol{x}_{gt}) &= C(\hat{\boldsymbol{x}}^0_{adv}), \end{aligned} \tag{8}$$

where $\boldsymbol{x}_{gt}$ is groundtruth image, $\boldsymbol{x}_{adv}$ is the adversarial image, f is a model with the purification function in latent space, t^* is predefined purification step, $\mathcal{E}$ and $\mathcal{D}$ are pretrained image encoder and decoder respectively. For simplifying symbolic, we denote $\mathcal{E}(\boldsymbol{x}_{adv})$ as $\boldsymbol{z}_{adv}$.

An overview of our method OSCP is illustrated in Fig. 2; OSCP can be separated into two components: a) training a sped-up backbone model with noise and denoise function; b) pipeline for purification process with non-text guidance. We propose Gaussian Adversarial Noise Distillation (GAND) and Controlled Adversarial Purification (CAP) to respond to those two unique requirements.

Algorithm 1 GAND

Require: dataset $\mathcal{X}$, class label of the image $\boldsymbol{y}$, classifier C, LCM f_θ and it's parameter $\boldsymbol{\theta}$, ODE solver Ψ and distance metric d. $\mathcal{L}_C$ and $\mathcal{L}_G$ indicate $\mathcal{L}_{CIG}$ (Eq. (14)) and $\mathcal{L}_{GAND}$ (Eq. (15)) respectively. t' denote time step t_{n+k} in diffusion process.

while not convergence **do**
  Sample $\boldsymbol{x} \sim \mathcal{X}$, $n \sim \mathcal{U}[1, (N-k)/2]$
  $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$
  $\boldsymbol{\delta}_{adv} = \arg\max_\delta L(C(\mathcal{D}(\boldsymbol{z}+\boldsymbol{\delta})), \boldsymbol{y})$ Eq. (12)
  $\boldsymbol{z}^*_{t'} = \sqrt{\bar{\alpha}_{t'}}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_{t'}}(\boldsymbol{\epsilon} + \boldsymbol{\delta}_{adv})$ Eq. (11)
  $\hat{\boldsymbol{z}}^\Psi_{t_n} \leftarrow \boldsymbol{z}^*_{t'} + \Psi(\boldsymbol{z}^*_{t'}, t', t_n, \varnothing)$ Eq. (13)
  $\mathcal{L}_G(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \leftarrow d(f_\theta(\boldsymbol{z}^*_{t'}, \varnothing, t'), f_{\theta^-}(\hat{\boldsymbol{z}}^\Psi_{t_n}, \varnothing, t_n))$
  $\mathcal{L}_C(\boldsymbol{\theta}) \leftarrow d(f_\theta(\boldsymbol{z}^*_{t'}, \varnothing, t'), \boldsymbol{z})$
  $\mathcal{L}_T(\boldsymbol{\theta}, \boldsymbol{\theta}^-) \leftarrow \mathcal{L}_G(\boldsymbol{\theta}, \boldsymbol{\theta}^-) + \lambda_{CIG}\mathcal{L}_C(\boldsymbol{\theta})$ Eq. (16)
  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \eta\nabla_\theta\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\theta}^-)$
end while

4.2. Gaussian Adversarial Noise Distillation

To respond to the first component we defined, we propose GAND as illustrated in Fig. 2 (left) and detailed in Alg. 1, a novel adversarial distillation objective, which combines adversarial training, Latent Consistency Distillation, and learning how to remove the adversarial and Gaussian noise from the shifted standard Gaussian distribution.

The original latent consistency distillation first encodes the image to latent code by encoder $\mathcal{E}$, such that $\boldsymbol{z} = \mathcal{E}(\boldsymbol{x})$ and trains the model base on loss function:

$$\begin{aligned} &\mathcal{L}(\theta, \theta^-; \Psi) = \\ &\quad \mathbb{E}_{\boldsymbol{z}, \boldsymbol{c}, n}[d(f_\theta(\boldsymbol{z}_{t_{n+k}}, \boldsymbol{c}, t_{n+k}), f_{\theta^-}(\hat{\boldsymbol{z}}^{\Psi,\omega}_{t_n}, \boldsymbol{c}, t_n))], \end{aligned} \tag{9}$$

where $d(\cdot,\cdot)$ is distance metric, $\Psi(\cdot,\cdot,\cdot,\cdot)$ is DDIM PF-ODE solver Ψ_{DDIM} [29] , $f(\cdot,\cdot,\cdot)$ is latent consistency function in E.q. 7 and $\hat{\boldsymbol{z}}^{\Psi,\omega}_{t_n}$ is an estimation of the solution of the PF-ODE from $t_{n+k} \to t_n$ using the DDIM PF-ODE solver Ψ:

$$\begin{aligned} \hat{\boldsymbol{z}}^{\Psi,\omega}_{t_n} \leftarrow \boldsymbol{z}_{t_{n+k}} &+ (1+\omega)\Psi(\boldsymbol{z}_{t_{n+k}}, t_{n+k}, t_n, c) \\ &- \omega\Psi(\boldsymbol{z}_{t_{n+k}}, t_{n+k}, t_n, \varnothing), \end{aligned} \tag{10}$$

ω is guidance scale and it is sampled from $[\omega_{\min}, \omega_{\max}]$.

We aim to learn the denoise trajectory from a combination of Gaussian noise and adversarial noise to natural image; we need a model satisfying that $f_\theta(\boldsymbol{z}_{adv}(t), \varnothing, t) = f_\theta(\boldsymbol{z}(t), \varnothing, t) = \boldsymbol{z}_\epsilon$ for all t. $\boldsymbol{z}_\epsilon$ is the limit of $\boldsymbol{z}_t$ when $t \to 0$ where $\boldsymbol{z}_\epsilon \simeq \boldsymbol{z}$. However, if we look closely at the formulation of f_θ in Eq. (7), $f_\theta(\boldsymbol{z}_{adv}(t), \varnothing, t)$ is converging to $\boldsymbol{z}_{adv}$ when $t \to 0$. $f_\theta(\boldsymbol{z}_{adv}(t), \varnothing, t) - f_\theta(\boldsymbol{z}(t), \varnothing, t) \to \boldsymbol{\delta}_{adv}$ when $t \to 0$, where $\boldsymbol{\delta}_{adv} = \boldsymbol{z}_{adv} - \boldsymbol{z}$. This violates the definition of a Latent Consistency Model mentioned in Eq. (6). To address this disharmony, we introduce $\boldsymbol{z}^*_t$

$$\boldsymbol{z}^*_t = \sqrt{\bar{\alpha}_t}\boldsymbol{z} + \sqrt{1-\bar{\alpha}_t}(\boldsymbol{\epsilon} + \boldsymbol{\delta}_{adv}), \tag{11}$$

by borrowing the idea of forward diffusion process, making $\boldsymbol{z}_t^*$ be an linear combination of $\boldsymbol{z}$ and $\boldsymbol{z}_{adv}$, worth to mention that these $\boldsymbol{\delta}_{adv}$ used in GAND are generated from latent space attack.

$$\boldsymbol{\delta}_{adv} = \arg\max_{\delta} \mathcal{L}(C(\mathcal{D}(\mathcal{E}(\boldsymbol{x}) + \boldsymbol{\delta})), \boldsymbol{y}). \tag{12}$$

The $\boldsymbol{z}_t^* \to \boldsymbol{z}$ when $t \to 0$, and $\boldsymbol{z}_t^* \to \boldsymbol{\epsilon} + \boldsymbol{\delta}_{adv}$ when $t \to T$, which converges to the natural image when t is small and matches with our goal, removing the adversarial noise and Gaussian noise concurrently from the shifted normal distribution. Hence, we eliminate disharmony between adversarial training and the original latent consistency distillation objective.

Then, as we do not give condition embeddings, we remove the ω in the formulation of estimation $\hat{\boldsymbol{z}}_{t_n}^{\Psi,\omega}$ to be:

$$\hat{\boldsymbol{z}}_{t_n}^{\Psi} \leftarrow \boldsymbol{z}_{t_{n+k}} + \Psi(\boldsymbol{z}_{t_{n+k}}, t_{n+k}, t_n, \varnothing), \tag{13}$$

where $\varnothing$ means null conditional embedding.

Furthermore, our goal is purifying images instead of image generation, we can train our LCM in a weaker constraint, where we only need the consistency function satisfying on $[0, t]$, $t < T$. Hence, we only simulate the n of time step t_n in $\mathcal{U}[1, (N-k)/2]$. This means that our LCM only satisfies the consistency function in the first half of the time steps.

In addition, inspired by [20], we add a Clear Image Guide (CIG) loss to further ensure our distillation process is training towards the proposed purification goal. CIG loss is given by:

$$\mathcal{L}_{CIG}(\boldsymbol{\theta}) = \mathbb{E}_{z,n}[d(f_\theta(\boldsymbol{z}_{t_{n+k}}^*, \varnothing, t_{n+k}), \boldsymbol{z})]. \tag{14}$$

Therefore, our total distillation loss is the combination of GAND loss

$$\begin{aligned}\mathcal{L}_{GAND}(\boldsymbol{\theta},\boldsymbol{\theta}^-) = \\ \mathbb{E}_{z,n}[d(f_\theta(\boldsymbol{z}_{t_{n+k}}^*, \varnothing, t_{n+k}), f_{\theta^-}(\hat{\boldsymbol{z}}_{t_n}^{\Psi}, \varnothing, t_n))],\end{aligned} \tag{15}$$

and CIG loss as:

$$\mathcal{L}_{Total}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) = \mathcal{L}_{GAND}(\boldsymbol{\theta}, \boldsymbol{\theta}^-) + \lambda_{CIG}\mathcal{L}_{CIG}(\boldsymbol{\theta}), \tag{16}$$

λ_{CIG} is used to ensure two losses are on the same scale.

The following theorem shows if the loss of GAND converges to an arbitrarily small number, the difference between the purified image and the clean image will be arbitrarily small. We can use this as guidance on how to choose our purification t in the inference step since our method is reaching the theoretical optimal performance at a specific t.

4.3. Controlled Adversarial Purification

To address the first component we defined, we propose CAP (as shown in Fig. 3) - a purification inference pipeline that utilizes visual prompts instead of text guidance. We chose visual prompts because text-based guidance can be vulnerable to caption semantic attacks [47], which may compromise the accuracy of generated text prompts. Therefore, we opt for a more traditional and robust approach by using an unlearnable edge detection operator to generate edge guidance of the input adversarial sample.

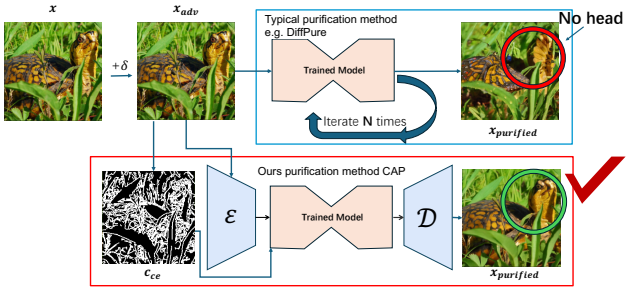


Figure 3. CAP used edge image of the adversarial image to control our purification process, maximizing the remaining semantic information of the purified image.

In purification process, we first encode the adversarial image $\boldsymbol{z}_{adv} = \mathcal{E}(\boldsymbol{x}_{adv})$ to latent space using pre-trained image encoder $\mathcal{E}$ and sample a random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ in the dimension of the latent space. Then, we diffuse the $\boldsymbol{z}_{adv}$ with predefined strength t^*, using forward latent diffusion process:

$$\boldsymbol{z}_{adv}(t^*) = \sqrt{\bar{\alpha}(t^*)}\boldsymbol{z}_{adv} + \sqrt{1 - \bar{\alpha}(t^*)}\boldsymbol{\epsilon}. \tag{17}$$

Then, we purify $\boldsymbol{z}_{adv}(t^*)$ using our LCM trained by GAND (in Alg. 1), $f_\theta(\boldsymbol{z}, \boldsymbol{c}, t)$, $\boldsymbol{z}$ is image latent, $\boldsymbol{c}$ is the condition embedding (e.g., text, canny edge image) and t is time step. The Latent Consistency Model has been introduced in Eq. (7). The purified image latent comes from the latent consistency function, $\hat{\boldsymbol{z}}_{adv}^0 = f_\theta(\boldsymbol{z}_{adv}(t^*), \boldsymbol{c}_{ce}, t^*)$, where $\boldsymbol{c}_{ce}$ means canny edge images which are provided by edge detection operators [4]. Although ControlNet is an extra plug-in tool for Stable Diffusion, $\boldsymbol{c}_{ce}$ is not exactly an input of f_θ, we treat it as condition embedding here for simplifying the equation. To further reduce the effect of the adversarial image, we remove the $c_{skip}(t)\boldsymbol{z}_{adv}(t)$ in LCM f_θ and denote this LCM as f_θ^-, since this term will maintain most of the adversarial noise. Therefore, our purified image latent is actually:

$$\begin{aligned}\hat{\boldsymbol{z}}_{adv}^0 &= f_\theta^-(\boldsymbol{z}_{adv}(t^*), \boldsymbol{c}_{ce}, t^*) \\ &= c_{\text{out}}(t)\left(\frac{\boldsymbol{z}_{adv} - \sqrt{1-\bar{\alpha}_t}\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{z}_{adv}, \boldsymbol{c}_{ce}, t)}{\sqrt{\bar{\alpha}_t}}\right).\end{aligned} \tag{18}$$

Same as setting $c_{skip}(t) \equiv 0$. Finally, We reconstruct the purified image by the image decoder $\mathcal{D}$, $\hat{\boldsymbol{x}}_{adv}^0 = \mathcal{D}(\hat{\boldsymbol{z}}_{adv}^0)$.

5. Experiments and Results

Training setting Our adversarial latent consistency model is distilled from Stable Diffusion v1.5 [34], which utilizes ϵ-prediction [17]. We split the first 10,000 images and the remaining 40,000 images in the ImageNet [8] valset as our testing set and training set, respectively. The GAND framework is trained on our training set, with all images resized to 512×512 resolution. Training proceeds for 20,000 iterations with a batch size of 4, employing an initial learning

Table 1. Accuracy (%) results for ImageNet. The best models against different attack methods are in **bold**. The methods marked with * mean the data are borrowed from the original paper.

Defense	Attack Method	Standard Accuracy	Robust Accuracy	Architecture
Without defense	untargeted PGD-100	80.55%	0.01%	ResNet-50
Without defense	AutoAttack	80.55%	0.00%	ResNet-50
Without defense	random-targeted PGD-40	82.33%	0.04%	ResNet-152
Diffusion Based Purification				
Wang et al. [44]	untargeted PGD-100	73.53%	72.97%	ResNet-50
Wang et al. [44]	random-targeted PGD-40	78.10%	77.86%	ResNet-152
DiffPure [32]	AutoAttack	75.77%	73.02%	ResNet-50
Adversarial Training				
Amini et al. [1]*	AutoAttack	**77.96%**	59.64%	ConvNeXt-L
Singh et al. [38]*	AutoAttack	77.00%	57.70%	ConvNeXt-L
Hybrid				
OSCP (ours)	untargeted PGD-100	**77.63%**	**73.89%**	ResNet-50
OSCP (ours)	AutoAttack	77.63%	**74.19%**	ResNet-50
OSCP (ours)	random-targeted PGD-40	**79.81%**	**78.78%**	ResNet-152

rate of 8e-6 and a 500-step warm-up period. For the diffusion process, we employ DDIM-Solver [39] as the PF-ODE solver Ψ with a skipping step $k = 20$ in Equation 13. Adversarial latents are generated using PGD-10 with $\epsilon = 0.03$ against a ResNet50 victim model, and the consistency improvement gradient weight λ_{CIG} is set to 0.001. Notably, our experiments demonstrate that PEFT via LoRA is sufficient for achieving optimal performance.

Attack Settings We conduct our evaluations on our testing set. Our evaluation encompasses diverse architectures, spanning traditional CNNs (ResNet 50 and 152 [16], WideResNet [50]) and modern Vision Transformers (ViT-B [10], Swin-B [27]). For adversarial attacks, we employ standard L_p norm-based methods including PGD [31] and AutoAttack [6]. The PGD attacks are configured with L_∞ bounds $\gamma \in \{4/255, 16/255\}$ and corresponding step sizes $\eta \in \{1/255, 0.025 \cdot 16/255\}$, while AutoAttack uses L_∞ bound $\gamma = 4/255$. Here, PGD-n denotes PGD attack with n iterations. In our evaluation metrics, we define standard accuracy as performance on clean data with our defense framework, robust accuracy as performance on adversarial data with our defense, and clean accuracy as performance on clean data without any defense. We fix the random seed of every experiment to avoid the randomness.

5.1. Main Result

Our method is regarded as a hybrid method, consisting of adversarial training and purification. Hence we compare our method with both categories of the defenses. Based on the provided attack setting, our method demonstrates significant computational efficiency, completing evaluations in 3 hours on an NVIDIA F40 GPU compared to GDMP's [44] 48 hours. As shown in Tab. 1, diffusion based purification achieves satisfactory robust accuracy while sacrificing standard accuracy. On the other hand, adversarial training has little standard accuracy loss but the robust accuracy is unsatisfactory. Our method achieves promising results on both standard and robust accuracy with three different attack methods. We achieve robust accuracy improvements of 0.92% and 1.17% against PGD-100 ($\gamma = 4/255$) and AutoAttack respectively, while improving from 77.86% to 78.78% against random targeted PGD-40 ($\gamma = 16/255$). We empirically determine $t^* = 200$ as the optimal purification strength.

Table 2. Accuracy (%) and attack success rate (ASR %) results in various architectures, where the attack method is PGD-100, with budget $4/255$, attack step size 1/255.

Architecture	Clean	ASR	Standard	Robust
WRN-50-2	82.6%	100%	77.6%	75.2%
Vit-b-16	81.6%	100%	78.2%	71.6%
Swin-b	83.6%	100%	79.2%	77.8%

5.2. Analysis With Different Architectures

While our GAND-trained LCM is initially optimized using adversarial examples generated for ResNet50, we evaluate its generalization capability across different architectures to address potential overfitting concerns. As shown in Tab. 2, our method maintains strong robust accuracy across architectures, with even the lowest performance of 71.6% on ViT_{b-16} demonstrating effective transfer. These experiments, along with subsequent ImageNet evaluations, are conducted on a representative subset of 500 images from

Table 3. Robust Accuracy (%) on our method under Diff-PGD-10 attack $\gamma = 8/255$ ($\eta = 2/255$) on ImageNet [8].

Defence method	ResNet-50	ResNet-152	WideResNet-50-2	Vit-b-16	Swin-b	ConvNeXt-b
DiffPure ($t^* = 100$)	53.8%	49.4%	52.2%	16.6%	45.1%	42.9%
Ours ($t^* = 250$)	**59.0%**	**56.5%**	**57.9%**	**34.1%**	**53.9%**	**49.1%**

Table 4. Inference time of purification models to purify one image on an NVIDIA F40 GPU. Notably, GDMP refers to Wang et al. [44], DiffPure refers to Nie et al. [32].

Method	Dataset	$t^* \in$ [0,1000]	runtime (s)
GDMP	ImageNet	250	~ 9
DiffPure	ImageNet	150	~ 11
OSCP(Ours)	ImageNet	20/200/500/1000	$\sim$ **0.1**
OSCP(Ours)	CelabA-HQ	20/200/500/1000	~ 0.5

our previously mentioned 10,000 images test set, as we observe consistent performance between these sample sizes.

5.3. Defend Adaptive Attack

In Tab. 3, we test our method under version 2 of Diff-PGD attack [48], which is a SOTA attack method against diffusion based purification. Our method boost the robust accuracy in Vit-b-16 from 16.6% to 34.1%. In ResNet-50, ResNet-152, WideResNet-50-2, the robust accuracy increase 5.2%, 7.1% and 5.7% respectively. For robust accuracy in Swin-b and ConvNeXt, our method acheive 53.9% and 49.1%, which is 8.8% and 6.2% higher than DiffPure. Worth to mention that our method uses LCM model which is deterministic, and hence, Expectation over Transformation, which is usually use in other robust evaluation for DBP, is not needed here.

5.4. Inference Time

Conventional diffusion-based purification methods are computationally intensive, making them impractical for real-time applications. As shown in Tab. 4, our method achieves purification in just **0.1s** for ImageNet and **0.5s** for CelebA-HQ (1024 $\times$ 1024) on a NVIDIA F40 GPU, **independent of** t^* where the inference time of other method scales up depending on t^*. More detailed experiments can be found in the supplementary materials. Our method significantly speeds up the purification process, and this acceleration enables deployment in time-critical scenarios such as autonomous driving and purification in video.

5.5. Face Recognition

We further test our model on a subset of 1000 images from CelebA-HQ [26], choosing purification step t^* as 200, control scale 0.8, and utilizing the GAND weights trained on ImageNet. We defend against targeted PGD attacks that attempt to manipulate Arcface [9] (AF), FaceNet [37] (FN), and MobileFaceNet [5] (MFN) to misidentify any input face as a specific target person. Our defense goal is to prevent this misidentification by ensuring the cosine similarity between the purified image embedding and the target image remains below the recognition threshold. To validate the effectiveness of ImageNet-trained GAND weights on CelebA-HQ, we compare OSCP with CAP (OSCP without GAND) as shown in Tab. 5. The GAND weights significantly improve robust accuracy from 83.4% to 86.8% on Arcface and from 82.8% to 84.9% on MobileFaceNet. Notably, both our methods (CAP and OSCP) achieve an exceptional robust accuracy of 97.8% on FaceNet.

Table 5. Robust accuracy (%) for CelebA-HQ under targeted PGD-10 $L_\infty\gamma(\gamma = 4/255)$, η = 0.5*5/255. The best model is in **bold**.

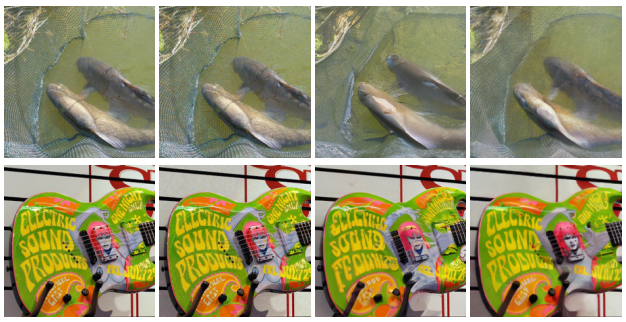
Defense Method	AF	FN	MFN
Without defense	0%	0.3%	0%
GassianBlur (σ = 7.0)	2.8%	51.4%	2.8%
Das et al. [7] (n = 60)	17.3%	84.1%	27.6%
CAP(Ours)	83.4%	**97.8%**	82.8%
OSCP(Ours)	**86.8%**	**97.8%**	**84.9%**

Table 6. IQAs for $\epsilon = 8/255$. ↑/↓ indicate higher/lower value consists the better image quality. Notably, $\mathrm{IQA}(\cdot)$ indicates one of IQA we leverage, $\boldsymbol{x}$ refers to the clean image, $\boldsymbol{x}_{\text{adv}}$ refers to the adversarial sample, $\boldsymbol{x}_{\text{ours}}$ refers to the purified image by our method and $\boldsymbol{x}_{\text{DiffPure}}$ refers to purified image by DiffPure.

	LPIPS↓	PSNR↑	SSIM↑
$\mathrm{IQA}(\boldsymbol{x}, \boldsymbol{x}_{\text{adv}})$	0.1547	34.21	0.8906
$\mathrm{IQA}(\boldsymbol{x}, \boldsymbol{x}_{\text{ours}})$	**0.2370**	**24.13**	**0.7343**
$\mathrm{IQA}(\boldsymbol{x}, \boldsymbol{x}_{\text{DiffPure}})$	0.2616	24.11	0.7155

5.6. Image Quality Assessment

Regarding image quality assessments, we compare our method with DiffPure using a subset of 1000 ImageNet images from our testing set. We use PSNR to reflect the overall distortion level of images; SSIM [45] to indicate how human eyes perceive image structures; LPIPS [52] to assesses perceptual similarity. Both methods are configured with t^*=250 for fair comparison. As shown in Tab. 6, our method demonstrates superior performance across multiple metrics: LPIPS decreases from 0.2616 to 0.2370, while PSNR and SSIM improve from 24.11 to 24.13 and 0.7155 to 0.7343. These improvements consistently indicate that our purified images maintain greater fidelity to the original inputs, validating the effectiveness of our purification approach. The visualization is shown in Fig. 4.



(a) (b) (c) (d)

Figure 4. Visualization of the IQA experiment which compares with DiffPure and the proposed method. (a) Input image. (b) Adversarial image. (c) DiffPure. (d) Ours.

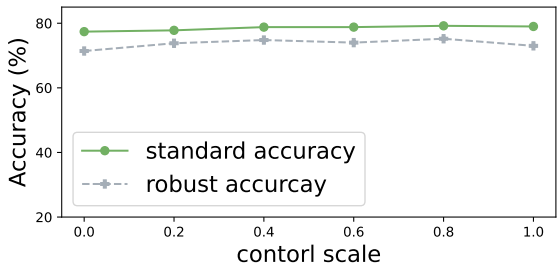


Figure 5. Performance of our method on different t^* under PGD-100 $L_\infty\gamma$ ($\gamma = 4/255$), $\eta = 0.01$*4/255, where we evaluate on ResNet50 on ImageNet.

Table 7. Accuracy (%) under PGD-100 $L_\infty\gamma$ (γ = 4/255), η = 0.01*4/255. The best result is in **bold**, the 2nd best is underlined.

CAP	GAND	Standard	Robust
✗	✗	75.0%	72.4%
✗	✓	74.0%	72.6%
✓	✗	**77.0%**	73.2%
✓	✓	76.8%	**75.0%**

5.7. Ablation Studies

We conduct comprehensive ablation studies on 500 images subset from our testing set. First, as shown in Fig. 5, we optimize CAP's performance across different ControlNet conditioning scales, identifying 0.8 as the optimal value. Tab. 7 demonstrates the individual and combined effects of our components. The baseline LCM-LoRA already shows promising performance, indicating its potential for adversarial purification. GAND-trained weights yield marginal improvements, while CAP integration increases standard and robust accuracy by 2% and 0.8% respectively. The full OSCP framework (combining both) significantly improves robust accuracy from 72.4% to 75%, validating our design choices. In Tab. 8, we analyze the impact of the $c_{skip}(t)\boldsymbol{z}_t$ term from Eq. (18). While standard accuracy remains stable, removing this term improves robust accuracy, likely because $\boldsymbol{z}_t$ retains adversarial noise from the input image. This observation leads us to exclude this term from our LCM implementation.

Table 8. Accuracy with and without $c_{skip}(t)\boldsymbol{z}_t$ term in LCM in our method under PGD-100 ($\gamma = 8/255$), $\eta = 2/255$ on ResNet50 on a subset of 500 images from our ImageNet testing set.

$c_{skip}(t)\boldsymbol{z}_t$	Acc.	t^*=20	t^*=60	t^*=100	t^*=200	t^*=250
✗	Robust	43.4%	**56.0%**	59.4%	**66.4%**	**68.4%**
✗	Standard	81.2%	81.4%	81.2%	79.2%	76.6%
✓	Robust	**44.2%**	54.8%	**60.4%**	66.0%	67.8%
✓	Standard	81.2%	81.4%	81.2%	79.2%	76.6%

Table 9. Accuracy of our method under AutoAttack ($\gamma = 8/255$) on ResNet50 on a subset of 1000 images from ImageNet on different sets of training time steps. ($t^* = 200$)

Acc.	$\frac{N-k}{4}$	$\frac{N-k}{2}$	$\frac{3(N-k)}{4}$	$N-k$
Robust	74.0%	**74.3%**	73.7%	73.6%
Standard	78.8%	**79.4%**	79.1%	79.1%

Finally, Tab. 9 explores different training time step ranges: $\mathcal{U}[1,(N-k)/4], \mathcal{U}[1,(N-k)/2], \mathcal{U}[1,3(N-k)/4]$, and $\mathcal{U}[1,N-k]$ (original). Results show that smaller time step ranges maintain better performance, suggesting that fine-tuning LCM with smaller time steps is sufficient for purification tasks.

6. Conclusion and Discussion

In this paper, we present One Step Control Purification (OSCP), a novel framework that addresses two critical challenges in diffusion-based adversarial purification: computational efficiency and semantic preservation. Our approach integrates the knowledge of pre-trained Diffusion Models with the proposed Gaussian Adversarial Noise Distillation (GAND) to achieve robust defense in a single inference step. Then as the inference framework, the Controlled Adversarial Purification (CAP), utilizing the unlearnable edge detection operator, effectively preserves semantic information of the input during purification. Extensive experiments validate OSCP's effectiveness across diverse architectures and datasets. On ImageNet, we achieve 74.19% robust accuracy against strong attacks while requiring only 0.1s per purification—a 100-fold speedup compared to existing methods. The framework demonstrates strong cross-domain generalization, from natural images to face recognition tasks, while our ablation studies confirm the synergistic benefits of combining GAND and CAP.

While our work represents significant progress in real-time adversarial defense, several promising directions remain, including the adaptive purification strength mechanisms and alternative control conditions. These advances could further enhance the applicability of diffusion-based defenses in time-critical scenarios.

References

[1] Sajjad Amini, Mohammadreza Teymoorianfard, Shiqing Ma, and Amir Houmansadr. Meansparse: Post-training robustness enhancement through mean-centered feature sparsification. *CoRR*, 2024. 6

[2] 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*, pages 274–283, 2018. 3

[3] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22563–22575, 2023. 3

[4] John Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, 1986. 5

[5] Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition*, pages 428–438. Springer International Publishing, 2018. 7

[6] Francesco Croce and Matthias Hein. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pages 2206–2216, 2020. 3, 6

[7] Nilaksh Das, Madhuri Shanbhogue, Shang-Tse Chen, Fred Hohman, Siwei Li, Li Chen, Michael E Kounavis, and Duen Horng Chau. Shield: Fast, practical defense and vaccination for deep learning using jpeg compression. In *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 196–204, 2018. 7

[8] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009. 5, 7

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 7

[10] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 6

[11] Sven Gowal, Chongli Qin, Jonathan Uesato, Timothy Mann, and Pushmeet Kohli. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv e-prints*, pages arXiv–2010, 2020. 3

[12] 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*, 34:4218–4233, 2021. 3

[13] Zhongliang Guo, Junhao Dong, Yifei Qian, Kaixuan Wang, Weiye Li, Ziheng Guo, Yuheng Wang, Yanli Li, Ognjen Arandjelović, and Lei Fang. Artwork protection against neural style transfer using locally adaptive adversarial color attack. In *ECAI 2024*, pages 1414–1421. IOS Press, 2024. 1

[14] Zhongliang Guo, Chun Tong Lei, Lei Fang, Shuai Zhao, Yifei Qian, Jingyu Lin, Zeyu Wang, Cunjian Chen, Ognjen Arandjelović, and Chun Pong Lau. A grey-box attack against latent diffusion model-based image editing by posterior collapse. *arXiv preprint arXiv:2408.10901*, 2024. 1

[15] Zhongliang Guo, Weiye Li, Yifei Qian, Ognjen Arandjelovic, and Lei Fang. A white-box false positive adversarial attack method on contrastive loss based offline handwritten signature verification models. In *International Conference on Artificial Intelligence and Statistics*, pages 901–909, 2024. 1

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016. 6

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3, 5

[18] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022. 3

[19] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799, 2019. 3

[20] Dongjun Kim, Chieh-Hsin Lai, Wei-Hsiang Liao, Naoki Murata, Yuhta Takida, Toshimitsu Uesaka, Yutong He, Yuki Mitsufuji, and Stefano Ermon. Consistency trajectory models: Learning probability flow ODE trajectory of diffusion. In *International Conference on Learning Representations*, 2024. 5

[21] Chun Pong Lau, Jiang Liu, Hossein Souri, Wei-An Lin, Soheil Feizi, and Rama Chellappa. Interpolated joint space adversarial training for robust and generalizable defenses. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):13054–13067, 2023. 3

[22] Wei-An Lin, Chun Pong Lau, Alexander Levine, Rama Chellappa, and Soheil Feizi. Dual manifold adversarial robustness: Defense against lp and non-lp adversarial attacks. In *Advances in Neural Information Processing Systems*, pages 3487–3498, 2020. 1, 3

[23] Jiang Liu, Chun Pong Lau, Hossein Souri, Soheil Feizi, and Rama Chellappa. Mutual adversarial training: Learning together is better than going alone. *IEEE Transactions on Information Forensics and Security*, 17:2364–2377, 2022. 1

[24] Jiang Liu, Chen Wei, Yuxiang Guo, Heng Yu, Alan Yuille, Soheil Feizi, Chun Pong Lau, and Rama Chellappa. Instruct2attack: Language-guided semantic adversarial attacks. *arXiv preprint arXiv:2311.15551*, 2023. 1

[25] Yiming Liu, Kezhao Liu, Yao Xiao, Ziyi Dong, Xiaogang Xu, Pengxu Wei, and Liang Lin. Towards better adversar-

ial purification via adversarial denoising diffusion training. *arXiv preprint arXiv:2404.14309*, 2024. 4

[26] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2015. 7

[27] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10012–10022, 2021. 6

[28] Shitong Luo and Wei Hu. Diffusion probabilistic models for 3d point cloud generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2837–2845, 2021. 3

[29] Simian Luo, Yiqin Tan, Longbo Huang, Jian Li, and Hang Zhao. Latent consistency models: Synthesizing high-resolution images with few-step inference. *arXiv e-prints*, pages arXiv–2310, 2023. 3, 4

[30] Simian Luo, Yiqin Tan, Suraj Patil, Daniel Gu, Patrick von Platen, Apolinário Passos, Longbo Huang, Jian Li, and Hang Zhao. Lcm-lora: A universal stable-diffusion acceleration module. *arXiv e-prints*, pages arXiv–2311, 2023. 3, 4

[31] 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, 3, 6

[32] Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Animashree Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, pages 16805–16827, 2022. 1, 2, 3, 6, 7

[33] Ben Poole, Ajay Jain, Jonathan T. Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. In *International Conference on Learning Representations*, 2023. 3

[34] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE, 2022. 3, 5

[35] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022. 3

[36] Pouya Samangouei, Maya Kabkab, and Rama Chellappa. Defense-gan: Protecting classifiers against adversarial attacks using generative models. In *International Conference on Learning Representations*, 2018. 1, 3

[37] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015. 7

[38] Naman Deep Singh, Francesco Croce, and Matthias Hein. Revisiting adversarial training for imagenet: Architectures, training and generalization across threat models. *Advances in Neural Information Processing Systems*, 36, 2024. 6

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. In *International Conference on Learning Representations*, 2021. 2, 3, 6

[40] Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019. 3

[41] Yang Song, Prafulla Dhariwal, Mark Chen, and Ilya Sutskever. Consistency models. In *International Conference on Machine Learning*, pages 32211–32252, 2023. 3

[42] Hossein Souri, Pirazh Khorramshahi, Chun Pong Lau, Micah Goldblum, and Rama Chellappa. Identification of attack-specific signatures in adversarial examples. *arXiv preprint arXiv:2110.06802*, 2021. 1

[43] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 1

[44] Jinyi Wang, Zhaoyang Lyu, Dahua Lin, Bo Dai, and Hongfei Fu. Guided diffusion model for adversarial purification. *arXiv preprint arXiv:2205.14969*, 2022. 2, 3, 4, 6, 7

[45] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 7

[46] Zekai Wang, Tianyu Pang, Chao Du, Min Lin, Weiwei Liu, and Shuicheng Yan. Better diffusion models further improve adversarial training. In *International Conference on Machine Learning*, pages 36246–36263, 2023. 1, 3

[47] Yan Xu, Baoyuan Wu, Fumin Shen, Yanbo Fan, Yong Zhang, Heng Tao Shen, and Wei Liu. Exact adversarial attack to image captioning via structured output learning with latent variables. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 5

[48] Haotian Xue, Alexandre Araujo, Bin Hu, and Yongxin Chen. Diffusion-based adversarial sample generation for improved stealthiness and controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 7

[49] Jongmin Yoon, Sung Ju Hwang, and Juho Lee. Adversarial purification with score-based generative models. In *International Conference on Machine Learning*, pages 12062–12072, 2021. 1, 3

[50] Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *British Machine Vision Conference 2016*. British Machine Vision Association, 2016. 6

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3836–3847, 2023. 3

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 7

[53] Shuai Zhao, Meihuizi Jia, Zhongliang Guo, Leilei Gan, Xiaoyu Xu, Xiaobao Wu, Jie Fu, Feng Yichao, Fengjun Pan, and Anh Tuan Luu. A survey of recent backdoor attacks and defenses in large language models. *Transactions on Machine Learning Research*, 2025. Survey Certification. 3