The Thirty-Eighth AAAI Conference on Artificial Intelligence (AAAI-24)

Adversarial Purification with the Manifold Hypothesis

Zhaoyuan Yang[1], Zhiwei Xu[2], Jing Zhang[2], Richard Hartley[2], Peter Tu[1]

[1]GE Research, Niskayuna, NY
[2]Australian National University, Canberra, Australia
{zhaoyuan.yang,tu}@ge.com, {zhiwei.xu,jing.zhang,richard.hartley}@anu.edu.au

Abstract

In this work, we formulate a novel framework for adversarial robustness using the manifold hypothesis. This framework provides sufficient conditions for defending against adversarial examples. We develop an adversarial purification method with this framework. Our method combines manifold learning with variational inference to provide adversarial robustness without the need for expensive adversarial training. Experimentally, our approach can provide adversarial robustness even if attackers are aware of the existence of the defense. In addition, our method can also serve as a test-time defense mechanism for variational autoencoders. Code is available at: https://github.com/GoL2022/AdvPFY.

Introduction

State-of-the-art neural network models are known of being vulnerable to adversarial examples. With small perturbations, adversarial examples can completely change predictions of neural networks (Szegedy et al. 2014). Defense methods are then designed to produce robust models towards adversarial attacks. Common defense methods for adversarial attacks include adversarial training (Madry et al. 2018), certified robustness (Wong and Kolter 2018), *etc.*. Recently, adversarial purification has drawn increasing attention (Croce et al. 2022), which purifies adversarial examples during test time and thus requires fewer training resources.

Existing adversarial purification methods achieve superior performance when attackers are not aware of the existence of the defense; however, their performance drops significantly when attackers create defense-aware or adaptive attacks (Croce et al. 2022). Besides, most of them are empirical with limited theoretical justifications. Differently, we adapt ideas from the certified robustness and build an adversarial purification method with a theoretical foundation.

Specifically, our adversarial purification method is based on the assumption that high-dimensional images lie on low-dimensional manifolds (the manifold hypothesis). Compared with low-dimensional data, high-dimensional data are more vulnerable to adversarial examples (Goodfellow, Shlens, and Szegedy 2015). Thus, we transform the adversarial robustness problem from a high-dimensional image domain to a low-dimensional image manifold domain and present a novel adversarial purification method for non-adversarially trained models via manifold learning and variational inference (see Figure 1 for the pipeline). With our method, non-adversarially trained models can achieve performance on par with the performance of adversarially trained models. Even if attackers are aware of the defenses, our approach still provides robustness against attacks.

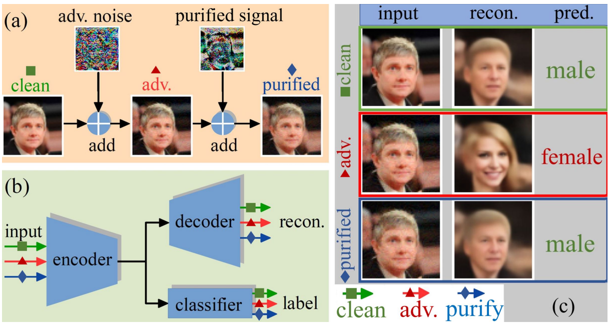


Figure 1: Adversarial purification against adversarial attacks. (a) Clean, adversarial (adv.), and purified images. (b) Jointly learning of the variational autoencoder and the classifier to achieve semantic consistency. (c) Applying semantic consistency between predictions and reconstructions to defend against attacks.

Our method is significant in introducing the manifold hypothesis to the adversarial defense framework. We improve a model's adversarial robustness from low-dimensional image manifolds than the complex high-dimensional image space. In the meantime, we provide conditions (in theory) to quantify the robustness of the predictions. Also, we present an effective adversarial purification approach combining manifold learning and variational inference, which achieves reliable performance on adaptive attacks without adversarial training. We also demonstrate the feasibility of our method to improve the robustness of adversarially trained models.

Related Work

Adversarial Training. Adversarial training is one of the most effective adversarial defense methods which incorporates adversarial examples into the training set (Goodfellow, Shlens, and Szegedy 2015; Madry et al. 2018). Such a method could degrade classification accuracy on clean data (Tsipras et al. 2019; Pang et al. 2022). To reduce the degradation in clean classification accuracy, TRADES (Zhang et al. 2019) is proposed to balance the trade-off between clean and robust accuracy. Recent works also study the effects of different hyperparameters (Pang et al. 2021; Huang et al. 2022a; Rice, Wong, and Kolter 2020) and data augmentation (Rebuffi et al. 2021; Sehwag et al. 2022; Zhao et al. 2020) to reduce robust overfitting and avoid the decrease of model's robust accuracy. Besides the standard adversarial training, others also study the adversarial training on manifolds (Stutz, Hein, and Schiele 2019; Lin et al. 2020; Zhou, Liang, and Chen 2020; Patel et al. 2020). In this work, we introduce a novel defense without adversarial training.

Adversarial Purification and Test-time Defense. As an alternative to adversarial training, adversarial purification aims to shift adversarial examples back to the representations of clean examples. Some efforts perform adversarial purification using GAN-based models (Samangouei, Kabkab, and Chellappa 2018), energy-based models (Grathwohl et al. 2020; Yoon, Hwang, and Lee 2021; Hill, Mitchell, and Zhu 2021), autoencoders (Hwang et al. 2019; Yin, Zhang, and Zuo 2022; Willetts et al. 2021; Gong et al. 2022; Meng and Chen 2017), augmentations, self-supervised learning (Pérez et al. 2021; Shi, Holtz, and Mishne 2021; Mao et al. 2021), PixelCNN (Song et al. 2018) *etc.*. Prior efforts (Athalye, Carlini, and Wagner 2018; Croce et al. 2022) have shown that methods such as Defense-GAN, PixelDefend (PixelCNN), SOAP (self-supervised), autoencoder-based purification are vulnerable to the Backward Pass Differentiable Approximation (BPDA) attacks (Athalye, Carlini, and Wagner 2018). Recently, diffusion-based adversarial purification methods have been studied (Nie et al. 2022; Xiao et al. 2022) and show adversarial robustness against adaptive attacks such as BPDA. Lee and Kim (2023), however, observe that the robustness of diffusion-based purification drops significantly when evaluated with the surrogate gradient designed for diffusion models. Similar to adversarial purification, existing test-time defense techniques (Nayak, Rawal, and Chakraborty 2022; Huang et al. 2022b) are also vulnerable to adaptive white-box attacks. In this work, we present a novel defense combining manifold learning and variational inference which achieves better performance compared with prior works and greater robustness on adaptive white-box attacks.

Methodology

In this section, we introduce an adversarial purification method with the manifold hypothesis. We first define sufficient conditions (in theory) to quantify the robustness of predictions. Then, we use variational inference to approximate such conditions in implementation and achieve adversarial robustness without adversarial training.

Let $\mathcal{D}_{XY}$ be a set of clean images and their labels where each image-label pair is defined as $(\mathbf{x}, \mathbf{y})$. The manifold hypothesis states that many real-world high-dimensional data $\mathbf{x} \in \mathbb{R}^n$ lies on a low-dimensional manifold $\mathcal{M}$ diffeomorphic to $\mathbb{R}^m$ with $m \ll n$. We define an encoder function $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a decoder function $\mathbf{f}^\dagger : \mathbb{R}^m \to \mathbb{R}^n$ to form an autoencoder, where $\mathbf{f}$ maps data point $\mathbf{x} \in \mathbb{R}^n$ to point $\mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$. For $\mathbf{x} \in \mathcal{M}$, $\mathbf{f}^\dagger$ and $\mathbf{f}$ are approximate inverses. See Yang et al. (2023) for notation details.

Problem Formulation

Let $\mathcal{L} = \{1, ..., c\}$ be a discrete label set of c classes and $\mathbf{h} : \mathbb{R}^m \to \mathcal{L}$ be a classifier of the latent space. Given an image-label pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{XY}$, the encoder maps the image $\mathbf{x}$ to a lower-dimensional vector $\mathbf{z} = \mathbf{f}(\mathbf{x}) \in \mathbb{R}^m$ and the functions $\mathbf{f}$ and $\mathbf{h}$ form a classifier of the image space $\mathbf{y}_{\mathrm{pred}} = \mathbf{h}(\mathbf{z}) = (\mathbf{h} \circ \mathbf{f})(\mathbf{x})$. Generally, the classifier predicts labels consistent with the ground truth labels such that $\mathbf{y}_{\mathrm{pred}} = \mathbf{y}$. However, during adversarial attacks, the adversary can generate a small adversarial perturbation $\boldsymbol{\delta}_{\mathrm{adv}}$ such that $(\mathbf{h} \circ \mathbf{f})(\mathbf{x}) \neq (\mathbf{h} \circ \mathbf{f})(\mathbf{x} + \boldsymbol{\delta}_{\mathrm{adv}})$. Thus, our purification framework aims to find a purified signal $\boldsymbol{\epsilon}_{\mathrm{pfy}} \in \mathbb{R}^n$ such that $(\mathbf{h} \circ \mathbf{f})(\mathbf{x}) = (\mathbf{h} \circ \mathbf{f})(\mathbf{x} + \boldsymbol{\delta}_{\mathrm{adv}} + \boldsymbol{\epsilon}_{\mathrm{pfy}}) = \mathbf{y}$. However, it is challenging to achieve $\boldsymbol{\epsilon}_{\mathrm{pfy}} = -\boldsymbol{\delta}_{\mathrm{adv}}$ because $\boldsymbol{\delta}_{\mathrm{adv}}$ is unknown. Thus, we aim to seek an alternative approach to estimate the purified signal $\boldsymbol{\epsilon}_{\mathrm{pfy}}$ and defend against attacks.

Theoretical Foundation for Adversarial Robustness

The adversarial perturbation is usually ℓ_p-bounded where $p \in \{0, 2, \infty\}$. We define the ℓ_p-norm of a vector $\mathbf{a} = [a_1, ..., a_n]^\intercal$ as $\|\mathbf{a}\|_p$ and a classifier of the image space as $\mathbf{G} : \mathbb{R}^n \to \mathcal{L}$. We follow Bastani et al. (2016); Leino, Wang, and Fredrikson (2021) to define the local robustness.

Definition 1 *(Locally robust image classifier) Given an image-label pair $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}_{XY}$, a classifier $\mathbf{G}$ is $(\mathbf{x}, \mathbf{y}, \tau)$-robust with respect to ℓ_p-norm if for every $\boldsymbol{\eta} \in \mathbb{R}^n$ with $\|\boldsymbol{\eta}\|_p \le \tau$, $\mathbf{y} = \mathbf{G}(\mathbf{x}) = \mathbf{G}(\mathbf{x} + \boldsymbol{\eta})$.*

Human vision is robust up to a certain perturbation. For example, given a clean MNIST image $\mathbf{x}$ with pixel values in [0, 1], if $\|\boldsymbol{\eta}\|_\infty \le 85/255$, we assign $(\mathbf{x} + \boldsymbol{\eta})$ and $\mathbf{x}$ to the same class (Madry et al. 2018). We use $\rho_H(\mathbf{x}, \mathbf{y})$ to represent the maximum perturbation budget for static human vision interpretations given an image-label pair $(\mathbf{x}, \mathbf{y})$. Exact $\rho_H(\mathbf{x}, \mathbf{y})$ is often a large value but difficult to estimate. We use it to represent the upper bound of achievable robustness.

Definition 2 *(Human-level image classifier) For every image-label pair $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{D}_{XY}$, if a classifier $\mathbf{G}_R$ is $(\mathbf{x}_i, \mathbf{y}_i, \tau_i)$-robust and $\tau_i \triangleq \rho_H(\mathbf{x}_i, \mathbf{y}_i)$ where $\rho_H(\cdot, \cdot)$ represents the maximum perturbation budget for static human vision interpretations, we define such a classifier as a human-level image classifier.*

The human-level image classifier $\mathbf{G}_R$ is an ideal classifier that is comparable to human. To construct such $\mathbf{G}_R$, we need a robust feature extractor $\mathbf{f}_R$ of the image space and a robust classifier $\mathbf{h}_R$ of the feature space to form $\mathbf{G}_R = \mathbf{h}_R \circ \mathbf{f}_R$. However, it is a challenge to construct a robust $\mathbf{f}_R$ due to the high-dimensional image space. Therefore, we aim to find an alternative solution to enhance the robustness of $\mathbf{h} \circ \mathbf{f}$ against adversarial attacks by enforcing the semantic consistency between the decoder $\mathbf{f}^\dagger$ and the classifier $\mathbf{h}$. Both functions ($\mathbf{f}^\dagger$ and $\mathbf{h}$) take inputs from a lower dimensional space (compared with the encoder); thus, they are more reliable (Goodfellow, Shlens, and Szegedy 2015).

We define a semantically consistent classifier on the manifold as $\mathbf{h}_S : \mathbb{R}^m \to \mathcal{L}$, which yields a class prediction $\mathbf{h}_S(\mathbf{z})$ given a latent vector $\mathbf{z} \in \mathbb{R}^m$.

Definition 3 *(Semantically consistent classifier on the manifold $\mathcal{M}$) A semantically consistent classifier $\mathbf{h}_S$ on the manifold $\mathcal{M}$ satisfies the following condition: for all $\mathbf{z} \in \mathbb{R}^m$, $\mathbf{h}_S(\mathbf{z}) = (\mathbf{G}_R \circ \mathbf{f}^\dagger)(\mathbf{z})$.*

A classifier (on the manifold) is a semantically consistent classifier if its predictions are consistent with the semantic interpretations of the images reconstructed by the decoder. While this definition uses the human-level image classifier $\mathbf{G}_R$, we can use the Bayesian method to approximate $\mathbf{h}_S$ without using $\mathbf{G}_R$ in experiment. Below, we provide the sufficient conditions of adversarial robustness for $\mathbf{h}_S \circ \mathbf{f}$ given an input $\mathbf{x}$, where the encoder $\mathbf{f}$ is not adversarially robust.

Proposition 1 *Let $(\mathbf{x}, \mathbf{y})$ be an image-label pair from $\mathcal{D}_{XY}$ and the human-level image classifier $\mathbf{G}_R$ be $(\mathbf{x}, \mathbf{y}, \tau)$-robust. If the encoder $\mathbf{f}$ and the decoder $\mathbf{f}^\dagger$ are approximately invertible for the given $\mathbf{x}$ such that the reconstruction error $\|\mathbf{x} - (\mathbf{f}^\dagger \circ \mathbf{f})(\mathbf{x})\|_p \triangleq \kappa \le \tau$ (**sufficient condition**), then there exists a function $\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n$ such that $(\mathbf{h}_S \circ \mathbf{f} \circ \mathbf{F})$ is $(\mathbf{x}, \mathbf{y}, \frac{\tau - \kappa}{2})$-robust.* (See Yang et al. (2023) for proof.)

The function $\mathbf{F}$ is considered to be the purifier for adversarial attacks. We construct such a function based on reconstruction errors. We assume the sufficient condition holds (bounded reconstruction errors κ for clean inputs). Lemma 1 states that adversarial attacks on a semantically consistent classifier lead to reconstruction errors larger than κ (abnormal reconstructions on adversarial examples).

Lemma 1 *If an adversarial example $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}_{\text{adv}}$ with $\|\boldsymbol{\delta}_{\text{adv}}\|_p \le \frac{\tau - \kappa}{2}$ causes $(\mathbf{h}_S \circ \mathbf{f})(\mathbf{x}_{\text{adv}}) \ne \mathbf{G}_R(\mathbf{x}_{\text{adv}})$, then $\|\mathbf{x}_{\text{adv}} - (\mathbf{f}^\dagger \circ \mathbf{f})(\mathbf{x}_{\text{adv}})\|_p > \frac{\tau + \kappa}{2} \ge \kappa$.* (Yang et al. 2023).

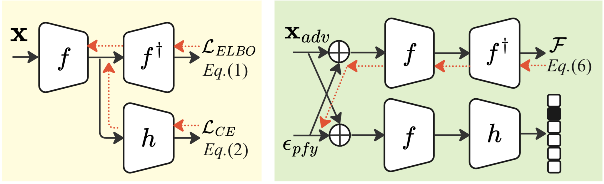
To defend against the attacks, we need to reduce the reconstruction error. Theorem 1 states that if a purified sample $\mathbf{x}_{\text{pfy}} = \mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}$ has a reconstruction error no larger than κ, the prediction from $(\mathbf{h}_S \circ \mathbf{f})(\mathbf{x}_{\text{pfy}})$ will be the same as the prediction from $\mathbf{G}_R(\mathbf{x})$.

Theorem 1 *If a purified signal $\boldsymbol{\epsilon}_{\text{pfy}} \in \mathbb{R}^n$ with $\|\boldsymbol{\epsilon}_{\text{pfy}}\|_p \le \frac{\tau - \kappa}{2}$ ensures that $\|(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}) - (\mathbf{f}^\dagger \circ \mathbf{f})(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}})\|_p \le \kappa$, then $(\mathbf{h}_S \circ \mathbf{f})(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}) = \mathbf{G}_R(\mathbf{x})$.*(Yang et al. 2023).

If $\boldsymbol{\epsilon}_{\text{pfy}} = -\boldsymbol{\delta}_{\text{adv}}$, then $\|\mathbf{x}_{\text{pfy}} - (\mathbf{f}^\dagger \circ \mathbf{f})(\mathbf{x}_{\text{pfy}})\|_p = \kappa$. Thus, feasible regions for $\boldsymbol{\epsilon}_{\text{pfy}}$ are non-empty. Let $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^n$ be a function that takes an input $\mathbf{x}$ and outputs a purified signal $\boldsymbol{\epsilon}_{\text{pfy}} = \mathbf{S}(\mathbf{x})$ by minimizing the reconstruction error, then $\mathbf{F}(\mathbf{x}) \triangleq \mathbf{x} + \mathbf{S}(\mathbf{x})$ and $\mathbf{h}_S \circ \mathbf{f} \circ \mathbf{F}$ is $(\mathbf{x}, \mathbf{y}, \frac{\tau - \kappa}{2})$-robust.

Remark 1 *For every perturbation $\boldsymbol{\delta} \in \mathbb{R}^n$ with $\|\boldsymbol{\delta}\|_p \le \nu$, if $\mathbf{S}(\mathbf{x} + \boldsymbol{\delta}) = -\boldsymbol{\delta}$, then the function $\mathbf{S} : \mathbb{R}^n \to \mathbb{R}^n$ is locally Lipschitz continuous on $\mathcal{B}_\nu \triangleq \{\hat{\mathbf{x}} \in \mathbb{R}^n \mid \|\hat{\mathbf{x}} - \mathbf{x}\|_p < \nu\}$ with a Lipschitz constant of 1.*(Yang et al. 2023).

Insights of the Theory. We transform a high-dimensional adversarial robustness problem into a low-dimensional semantic consistency problem. Since we only provide the sufficient conditions for robustness, dissatisfaction with the conditions is not necessary to be adversarially vulnerable. Our conditions indicate that higher reconstruction quality could lead to stronger robustness. Meanwhile, our method can certify robustness up to $\frac{\tau}{2}$ (reconstruction error $\kappa = 0$) when a human-level image classifier can certify robustness up to τ. The insight is that adding a purified signal on top of an adversarial example could change the image semantic, see Figure 5(c). Our framework is based on the triangle inequality and can be extended to other distance metrics.



(a) Train networks (b) Test-time purification

Figure 2: Two-stage pipeline: (a) jointly training for semantic consistency between the decoder and the classifier and (b) iterative updates of ϵ_{pfy} to purify $\mathbf{x}_{\text{adv}}$ in inference.

Relaxation. Our framework requires semantic consistency between the classifier on the manifold and the decoder on the manifold. Despite that the classifiers and decoders (on the manifold) have a low input dimension, it is still difficult to achieve high semantic consistency between them. Meanwhile, the human-level image classifier $\mathbf{G}_R$ is not available. Thus, we assume that predictions and reconstructions from high data density regions of $p(\mathbf{z}|\mathbf{x})$ are more likely to be semantically consistent (Zhou 2022). Next, we introduce a practical implementation of adversarial purification based on our framework. The implementation includes two stages: (1) enforce consistency during training and (2) test-time purification of adversarial examples, see Figure 2.

Semantic Consistency with the ELBO

Exact inference of $p(\mathbf{z}|\mathbf{x})$ is often intractable, we, therefore, use variational inference to approximate the posterior $p(\mathbf{z}|\mathbf{x})$ with a different distribution $q(\mathbf{z}|\mathbf{x})$. We define two parameters θ and ϕ which parameterize the distributions $p_\theta(\mathbf{x}|\mathbf{z})$ and $q_\phi(\mathbf{z}|\mathbf{x})$. When the evidence lower bound (ELBO) is maximized, $q_\phi(\mathbf{z}|\mathbf{x})$ is considered to be a reasonable approximation of $p(\mathbf{z}|\mathbf{x})$. To enforce the semantic consistency between the classifier and the decoder, we force the latent vector $\mathbf{z}$ inferred from the $q_\phi(\mathbf{z}|\mathbf{x})$ to contain the class label information of the input $\mathbf{x}$. We define a one-hot label vector as $\mathbf{y} = [y_1, y_2 ..., y_c]^\intercal$, where c is the number of classes and $y_i = 1$ if the image label is i otherwise it is zero. A classification head parametrized by $\boldsymbol{\psi}$ is represented as $\mathbf{h}_{\boldsymbol{\psi}}(\mathbf{z}) = [h_1(\mathbf{z}), h_2(\mathbf{z}) ..., h_c(\mathbf{z})]^\intercal$ and the cross-entropy classification loss is $-\mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{y}^\intercal \log \mathbf{h}_{\boldsymbol{\psi}}(\mathbf{z})]$ where $\log(\cdot)$ is an element-wise function for a vector. We assume the classification loss is no greater than a threshold ω and the training objective can then be written as

$$\max_{\boldsymbol{\theta}, \boldsymbol{\phi}} \underbrace{\mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{KL}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}) \| p(\mathbf{z})]}_{\text{ELBO (lower bound of } \log p_{\boldsymbol{\theta}}(\mathbf{x}))} \quad (1)$$

$$\text{s.t.} \quad -\mathbb{E}_{\mathbf{z} \sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{y}^\intercal \log \mathbf{h}_{\boldsymbol{\psi}}(\mathbf{z})] \le \omega. \quad (2)$$

We use the Lagrange multiplier with KKT conditions to optimize this objective as

$$\max_{\boldsymbol{\theta},\boldsymbol{\phi},\boldsymbol{\psi}} \text{ELBO} + \lambda \mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\mathbf{y}^{\intercal} \log \mathbf{h}_{\boldsymbol{\psi}}(\mathbf{z})], \tag{3}$$

where λ is a trade-off term to balance the two loss terms.

We follow Kingma and Welling (2014) to define the prior $p(\mathbf{z}) = \mathcal{N}(0, I)$ and the posterior (encoder) $q_{\phi}(\mathbf{z}|\mathbf{x})$ by using a normal distribution with diagonal covariance. Given an input vector $\mathbf{x}$, an encoder model parameterized by ϕ is used to model the posterior distribution $q_{\phi}(\mathbf{z}|\mathbf{x}) = \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}), \text{diag}(\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x})))$. The model predicts the mean vector $\boldsymbol{\mu}_{\boldsymbol{\phi}}(\mathbf{x}) = [\mu_1(\mathbf{x}), \mu_2(\mathbf{x})..., \mu_m(\mathbf{x})]^{\intercal}$ and the diagonal covariance $\boldsymbol{\sigma}^2_{\boldsymbol{\phi}}(\mathbf{x}) = [\sigma^2_1(\mathbf{x}), \sigma^2_2(\mathbf{x})..., \sigma^2_m(\mathbf{x})]^{\intercal}$. We define $p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z}) = (1/\beta)\text{exp}(-(1/\gamma)\|\mathbf{x} - f^{\dagger}_{\boldsymbol{\theta}}(\mathbf{z})\|^2_2)$, where γ controls the variance, β is a normalization constant, and $f^{\dagger}_{\boldsymbol{\theta}}$ is a decoder parametrized by $\boldsymbol{\theta}$, which maps data from the latent space to the image space. The illustrated network optimizing over Eq. (3) is provided in Figure 2(a).

Adversarial Attack and Purification

We mainly focus on the white-box attacks here. During inference time, the attackers can create adversarial perturbations on the classification head. Given a clean image $\mathbf{x}$, an adversarial example can be crafted as $\mathbf{x}_{\text{adv}} = \mathbf{x} + \boldsymbol{\delta}_{\text{adv}}$ with

$$\boldsymbol{\delta}_{\text{adv}} = \arg\max_{\boldsymbol{\delta}\in\mathcal{C}_{\text{adv}}} -\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}+\boldsymbol{\delta})}[\mathbf{y}^{\intercal} \log \mathbf{h}_{\boldsymbol{\psi}}(\mathbf{z})], \tag{4}$$

where $\mathcal{C}_{\text{adv}} \triangleq \{\boldsymbol{\delta} \in \mathbb{R}^n \mid \mathbf{x} + \boldsymbol{\delta} \in [0, 1]^n \text{ and } \|\boldsymbol{\delta}\|_p \leq \delta_{\text{th}}\}$ is the feasible set for δ_{th}-bounded perturbations.

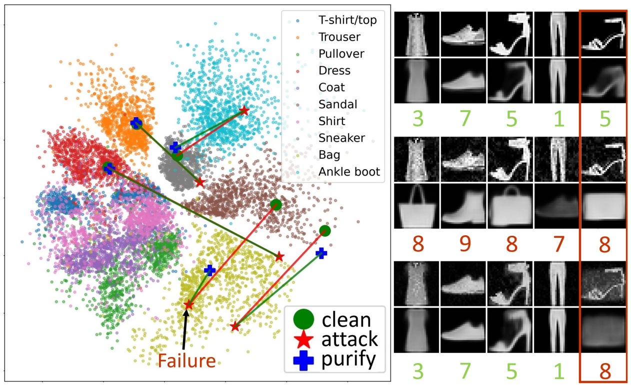
To avoid changing the semantic interpretation of the image, we need to estimate the purified signal with a ℓ_p-norm no greater than a threshold value ϵ_{th}. The purified signal also needs to project the sample to a high-density region of the manifold with a small reconstruction loss. The ELBO contains ℓ_2 reconstruction loss (relaxation from ℓ_p-metric), and it can be applied to maximize the posterior $q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})$. Therefore, we present a defense method that optimizes the ELBO during the test time to degrade the effects of the attacks. Given an adversarial example $\mathbf{x}_{\text{adv}}$, a purified sample can be obtained by $\mathbf{x}_{\text{pfy}} = \mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}$ with

$$\boldsymbol{\epsilon}_{\text{pfy}} = \arg\max_{\boldsymbol{\epsilon}\in\mathcal{C}_{\text{pfy}}} \mathbb{E}_{\mathbf{z}\sim \hat{q}_{\boldsymbol{\phi}}(\mathbf{z})}[\log \hat{p}_{\boldsymbol{\theta}}(\boldsymbol{\epsilon})] - D_{\text{KL}}[\hat{q}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon})\|p(\mathbf{z})], \tag{5}$$

where $\hat{p}_{\boldsymbol{\theta}}(\boldsymbol{\epsilon}) \triangleq p_{\boldsymbol{\theta}}(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}|\mathbf{z})$, $\hat{q}_{\boldsymbol{\phi}}(\boldsymbol{\epsilon}) \triangleq q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon})$ and $\mathcal{C}_{\text{pfy}} \triangleq \{\boldsymbol{\epsilon} \in \mathbb{R}^n \mid \mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon} \in [0, 1]^n \text{ and } \|\boldsymbol{\epsilon}\|_p \leq \epsilon_{\text{th}}\}$ which is a feasible set for purification. Compared with training a model to produce the purified signal $\boldsymbol{\epsilon}_{\text{pfy}} = \mathbf{S}(\mathbf{x})$, the test-time optimization of the ELBO is more efficient.

Algorithm 1: Test-time Purification

```
Input: $\mathbf{x}_{\text{adv}}$: input (adv) data; $\alpha$: learning rate; $T$: number of purification iterations; $\epsilon_{\text{th}}$: purification budget.
Output: $\mathbf{x}_{\text{pfy}}$: purified data; $s$: purification score.
procedure PURIFY($\mathbf{x}, \alpha, T, \epsilon_{\text{th}}$)
    $\boldsymbol{\epsilon}_{\text{pfy}} \sim \mathcal{U}_{[-\epsilon_{\text{th}}, \epsilon_{\text{th}}]}$    ▷ random initialization
    for t = 1, 2, ..., T do
        $\boldsymbol{\epsilon}_{\text{pfy}} \leftarrow \boldsymbol{\epsilon}_{\text{pfy}} + \alpha \cdot \text{sign}(\nabla_{\boldsymbol{\epsilon}_{\text{pfy}}} \mathcal{F}(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}))$
        $\boldsymbol{\epsilon}_{\text{pfy}} \leftarrow \min(\max(\boldsymbol{\epsilon}_{\text{pfy}}, -\epsilon_{\text{th}}), \epsilon_{\text{th}})$
        $\boldsymbol{\epsilon}_{\text{pfy}} \leftarrow \min(\max(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}, 0), 1) - \mathbf{x}_{\text{adv}}$
    $\mathbf{x}_{\text{pfy}} \leftarrow \mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}_{\text{pfy}}$    ▷ purified data
    $s \leftarrow \mathcal{F}(\mathbf{x}_{\text{pfy}})$    ▷ purification score
    return $\mathbf{x}_{\text{pfy}}, s$
```



(a) Trajectory: clean-attack-purify

(b) image pairs

Figure 3: (a) 2D trajectories of Fashion-MNIST. (b) Input and reconstructed image pairs on clean (top), adversarial (middle), and purified (bottom) examples with model predictions (numerical index) and failure cases (in box).

We focus on ℓ_∞-bounded purified vectors while our method is effective for both ℓ_2 and ℓ_∞ attacks in our experiments. We define α as the learning rate and $\text{Proj}_{\mathcal{S}}$ as the projection operator which projects a data point back to its feasible region when it is out of the region. We use a clipping function as the projection operator to ensure $\|\mathbf{x}_{\text{pfy}} - \mathbf{x}\|_\infty = \|\boldsymbol{\epsilon}_{\text{pfy}}\|_\infty \leq \epsilon_{\text{th}}$ and $\mathbf{x}_{\text{pfy}} = \mathbf{x} + \boldsymbol{\epsilon}_{\text{pfy}} \in [0, 1]^n$, where $\mathbf{x}$ is an (adversarial) image and ϵ_{th} is the budget for purification. We then define $\mathcal{F}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ as

$$\mathbb{E}_{\mathbf{z}\sim q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})}[\log p_{\boldsymbol{\theta}}(\mathbf{x}|\mathbf{z})] - D_{\text{KL}}[q_{\boldsymbol{\phi}}(\mathbf{z}|\mathbf{x})\|p(\mathbf{z})] \tag{6}$$

and iterative purification given adversarial example $\mathbf{x}_{\text{adv}}$ as

$$\boldsymbol{\epsilon}^{t+1} = \text{Proj}_{\mathcal{S}}\left(\boldsymbol{\epsilon}^t + \alpha\, \text{sgn}(\nabla_{\boldsymbol{\epsilon}^t} \mathcal{F}(\mathbf{x}_{\text{adv}} + \boldsymbol{\epsilon}^t; \boldsymbol{\theta}, \boldsymbol{\phi}))\right), \tag{7}$$

where the element-wise sign function $\text{sgn}(x) = x/|x|$ if x is non-zero otherwise it is zero. A detailed procedure is provided in Figure 2(b) and Algorithm 1.

The test-time optimization of the ELBO projects adversarial examples back to their feasible regions with a high posterior $q(\mathbf{z}|\mathbf{x})$ (regions where decoders and classifiers have strong semantic consistency) and a small reconstruction error (defend against adversarial attacks). To better demonstrate the process, we build a classification model in a 2-dimensional latent space (Figure 3 (a)) on Fashion-MNIST and show examples of clean, attack, and purified trajectories in Figure 3. Correspondingly, adversarial attacks are likely to push latent vectors to abnormal regions which cause abnormal reconstructions (Lemma 1). Through the test-time optimization over the ELBO, the latent vectors can be brought back to their original regions (Theorem 1).

If attackers are aware of the existence of purification, they could take advantage of this knowledge to create adaptive attacks. A straightforward formulation is to perform the multi-objective attacks with a trade-off term λ_a to balance the classification loss $\mathcal{H}(\mathbf{x}, \mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\psi})$ and the purification objective $\mathcal{F}(\mathbf{x}; \boldsymbol{\theta}, \boldsymbol{\phi})$ (Mao et al. 2021; Shi, Holtz, and Mishne 2021). The adversarial perturbation of the multi-objective attacks is

$$\boldsymbol{\delta}_{\mathrm{adv}} = \arg\max_{\boldsymbol{\delta} \in \mathcal{C}_{\mathrm{adv}}} \mathcal{H}(\mathbf{x}+\boldsymbol{\delta}, \mathbf{y}; \boldsymbol{\theta}, \boldsymbol{\psi}) + \lambda_a \mathcal{F}(\mathbf{x}+\boldsymbol{\delta}; \boldsymbol{\theta}, \boldsymbol{\phi}). \quad (8)$$

Another popular adaptive attack is the BPDA attack (Athalye, Carlini, and Wagner 2018). Consider a purification process as $\mathbf{x}_{\mathrm{pfy}} = \mathbf{F}(\mathbf{x})$ and a classifier as $\mathbf{G}(\mathbf{x}_{\mathrm{pfy}}) = (\mathbf{G} \circ \mathbf{F})(\mathbf{x})$. The BPDA attack uses an approximation $\nabla_{\hat{\mathbf{x}}} \mathbf{F}(\hat{\mathbf{x}}) \approx \mathbf{I}$ (identity) to estimate the gradient as $\nabla_{\hat{\mathbf{x}}} \mathbf{G}(\mathbf{F}(\hat{\mathbf{x}}))|_{\hat{\mathbf{x}}=\mathbf{x}} \approx \nabla_{\hat{\mathbf{x}}} \mathbf{G}(\hat{\mathbf{x}})|_{\hat{\mathbf{x}}=\mathbf{F}(\mathbf{x})}$. Many adversarial purification methods are vulnerable to the BPDA attack (Athalye, Carlini, and Wagner 2018; Croce et al. 2022). We show that even if attackers are aware of the defense, our method can still achieve effective robustness to this white-box adaptive attacks.

Experiments

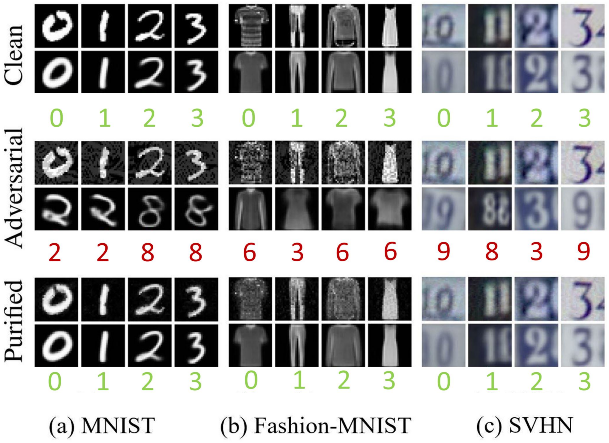
We first evaluate our method on MNIST (LeCun, Cortes, and Burges 2010), Fashion-MNIST (Xiao, Rasul, and Vollgraf 2017), SVHN (Netzer et al. 2011), and CIFAR-10 (Krizhevsky and Hinton 2009), followed by CIFAR-100 (Krizhevsky and Hinton 2009) and CelebA (64×64 and 128×128) for gender classification (Liu et al. 2015). See Yang et al. (2023) for the dataset details.

Model Architectures and Hyperparameters. We use three types of models (encoders) in our experiments: (1) tiny ResNet with standard training (for ablation study), (2) ResNet-50 with standard training (He et al. 2016) (for comparison with benchmark), and (3) PreActResNet-18 with adversarial training (Rebuffi et al. 2021) (to study the impact on adversarially trained models). We use several residual blocks to construct decoders and use linear layers for classification. We empirically set the weight of the classification loss (λ in Eq. (3)) to 8. See Yang et al. (2023) for details.

Adversarial Attacks. We evaluate our method on standard adversarial attacks and adaptive attacks (multi-objective and BPDA). All attacks are untargeted. For standard adversarial attacks (Eq. (4)), we use Foolbox (Rauber, Brendel, and Bethge 2017) to generate the PGD (ℓ_∞) attacks (Madry et al. 2018). We use Croce and Hein (2020) for the AutoAttack (ℓ_∞, ℓ_2). For the adaptive attacks, we use Torchattacks (Kim 2020) for the BPDA-PGD/APGD (ℓ_∞) attacks (Athalye, Carlini, and Wagner 2018), and standard PGD (ℓ_∞) for the multi-objective attacks. For MNIST and Fashion-MNIST, we report the attack hyperparameters and numerical results in Yang et al. (2023). For SVHN, CIFAR-10/100, and CelebA, we set the ℓ_∞ attack budget δ_{th} to $8/255$ and the ℓ_2 attack budget to 0.5. We run 100 iterations with step size $2/255$ for PGD (ℓ_∞) and 50 iterations with step size $2/255$ for the BPDA attack.

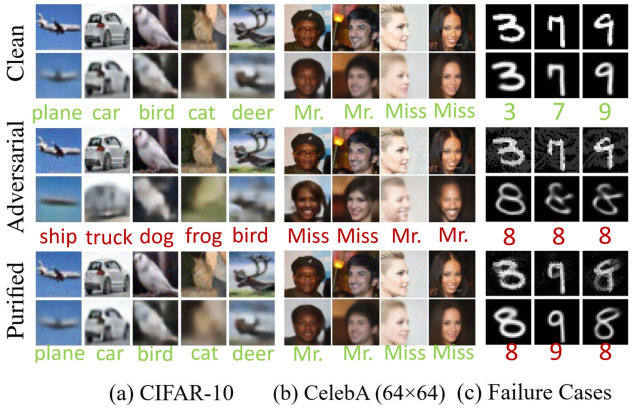
We also evaluate our ResNet-50 (CIFAR-10) model on the RayS (blackbox) attack (Chen and Gu 2020), the FGSM (ℓ_∞) attack (Goodfellow, Shlens, and Szegedy 2015) and the C&W (ℓ_2) attack (Carlini and Wagner 2017) in Yang et al. (2023) and our defense is effective for these attacks.

Test-time Purification. Key hyperparameters and experimental details are provided below, and only the ℓ_∞-bounded purification is considered in this work. We initialize the purified signal $\boldsymbol{\epsilon}_{\mathrm{pfy}}$ by sampling from an uniform distribution $\mathcal{U}_{[-\epsilon_{\mathrm{th}}, \epsilon_{\mathrm{th}}]}$ where ϵ_{th} is the purification budget. We run purification 16 times in parallel with different initializations and select the signal with the best purification score measured by the reconstruction loss or the ELBO. Step size α is alternated between $\{1/255, 2/255\}$ for each run. For SVHN, CIFAR-10/100, and CelebA, we set the ℓ_∞-purification budget ϵ_{th} to $8/255$ with 32 iterations. Despite the aforementioned hyperparameters, our method also works on other hyperparameter settings as shown in Figure 7.



(a) MNIST (b) Fashion-MNIST (c) SVHN

Figure 4: Examples using the VAE-Classifier models on clean, adversarial, and purified images with model predictions (text), similar to Figure 3(b). We report predictions from Fashion-MNIST in numerical index.

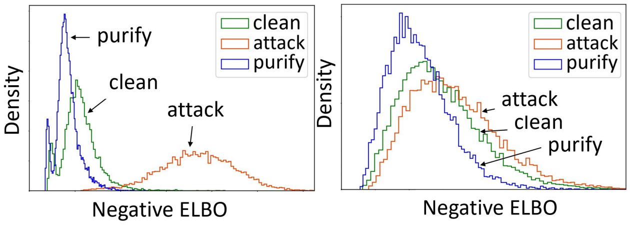

Clean
plane car bird cat deer Mr. Mr. Miss Miss 3 7 9
Adversarial
ship truck dog frog bird Miss Miss Mr. Mr. 8 8 8
Purified
plane car bird cat deer Mr. Mr. Miss Miss 8 9 8

(a) CIFAR-10 (b) CelebA (64×64) (c) Failure Cases

Figure 5: Examples of more datasets with failed cases on MNIST in (c) in addition to Figure 4.

Baselines. We compare our VAE-Classifier (V-CLF) with the standard autoencoders, denoted by Standard-AE-Classifier (S-CLF), by replacing the ELBO with reconstruction loss. One should note that the classifiers of the Standard-AE-Classifier may not have consistency with their decoders. See Yang et al. (2023) for details.

Objectives of Test-time Purification. The autoencoders are



(a) MNIST (b) CIFAR-10

Figure 6: Illustration of negative ELBOs. Adversarial attacks yield higher negative ELBOs while our purification reverses the ELBO shifts. Compared with MNIST, diverse backgrounds of CIFAR-10 lead to less obvious ELBO shifts.

trained with the reconstruction loss and the ELBO respectively. The Standard-AE-Classifier can only minimize the reconstruction loss during inference while the VAE-Classifier can optimize both the reconstruction loss and the ELBO during inference. We use "TP (R)" to represent the test-time optimization of the reconstruction loss and "TP (E)" for the test-time optimization of the ELBO.

Inference Time. We evaluate our method on an NVIDIA Tesla P100 GPU in PyTorch. For ResNet-50 with a batch size of 256 on CIFAR-10, the average run time per batch for 32 purification steps is 17.65s. Following the standard in (Croce et al. 2022), our method is roughly 102×. To reduce the inference time, one can adapt APGD (adaptative learning rate and momentum) during purification.

Experiment Results

Standard Adversarial Attacks. For the standard adversarial attacks, only the classification heads are attacked. We observe that, for MNIST, Fashion-MNIST, and SVHN, the standard adversarial attacks of the VAE-Classifier create abnormal reconstructed images while this is not applied to the Standard-AE-Classifier. It indicates that the classifiers and the decoders of the VAE-Classifier are strongly consistent. Figure 4 shows various sample predictions and reconstructions on clean, attack, and purified images from the VAE-Classifier. For clean images, the VAE-Classifier models achieve qualified reconstruction and predictions. For adversarial examples, the VAE-Classifier models generate abnormal reconstructions which are correlated with abnormal predictions from the classifiers (implied by Lemma 1). In other words, if the prediction of an adversarial example is 2, the digit on the reconstructed image may look like 2 as well. If we can estimate the purified vectors by minimizing the errors between inputs and reconstructions, the attacks could be defended (implied by Theorem 1). In our experiments, the predictions and reconstructions of adversarial examples become normal after using the test-time optimization (ELBO).

Figure 5 shows various sample predictions and reconstructions from the VAE-Classifier on CIFAR-10 and CelebA. The results are slightly different from those on MNIST, Fashion-MNIST, and SVHN: in Figures 5(a)-(b), although the reconstructed images are more blurry compared with MNIST, Fashion-MNIST, and SVHN, the VAE-Classifier is still robust under adversarial attacks. Distribution of the negative ELBO for clean, attack, and purified examples are shown in Figure 6.

	Method	Clean	PGD	AA-ℓ_∞	AA-ℓ_2	BPDA
SVHN	S-CLF	94.00	0.00	0.00	1.67	-
	TP (R)	93.03	40.83	44.82	59.93	7.65
	V-CLF	95.27	16.01	0.33	6.61	-
	TP (R)	90.66	72.44	**73.92**	79.15	**66.68**
	TP (E)	86.29	**72.72**	73.47	76.21	64.70
CIFAR-10	S-CLF	90.96	0.00	0.00	0.98	-
	TP (R)	87.80	11.65	13.74	44.57	0.70
	V-CLF	91.82	17.82	0.05	2.36	-
	TP (R)	78.51	51.20	51.63	59.35	43.02
	TP (E)	77.97	**57.21**	**58.78**	63.38	**47.43**

Table 1: Classification accuracy on SVHN and CIFAR-10 with $\ell_\infty = 8/255$ and $\ell_2 = 0.5$. "S-CLF": Standard-AE-Classifier, "V-CLF": VAE-Classifier, "TP (R)": Test-time minimization of the reconstruction loss, "TP (E)": Test-time minimization of the negative ELBO, and "AA": AutoAttack. We evaluate the model with both BPDA-(PGD/APGD) and report the minimum accuracy.

Arch.	Method	Clean	PGD	AA-ℓ_∞	BPDA
R50	V-CLF	94.82	23.84	0.04	-
	TP (E)	85.12	**63.09**	**63.16**	**57.15**
PR18	V-CLF	87.35	61.08	58.65	-
	TP (E)	85.14	**61.98**	**63.73**	**60.52**

Table 2: VAE-Classifier with ResNet-50 (R50) and adversarially trained PreActResNet-18 (PR18) on CIFAR-10. We set $\ell_\infty = 8/255$. See Yang et al. (2023) for more results.

We use tiny ResNet backbones for ablation study between the Standard-AE-Classifier and the VAE-Classifier. Classification accuracy of CIFAR-10 and SVHN is provided in Table 1 (see Yang et al. (2023) for results on MNIST and Fashion-MNIST). According to our results, optimizing the ELBO during the test time is more effective than only optimizing the reconstruction loss. Table 2 shows test-time purification (CIFAR-10) with larger backbones such as ResNet-50 (standard training) and PreActResNet-18 (adversarial training). With our defense, the robust accuracy of ResNet-50 on CIFAR-10 increases by more than 50%. Our method can also be applied to adversarially trained models (PreActResNet-18) to further increase their robust accuracy.

Multi-objective Attacks. We evaluate our method with the multi-objective attacks on CIFAR-10 and provide accuracy with respect to trade-off term λ_a of Eq. (8) in Figure 7(a). We observe that the classification accuracy of adversarial examples increases as the trade-off term increases while impacts on our defense are not significant. Thus, our defense is robust to the multi-objective attacks. We show some multi-objective adversarial examples in Yang et al. (2023).

Backward Pass Differentiable Approximation (BPDA). We apply PGD and APGD to optimize the objective of the BPDA attacks. We highlight the minimum classification ac-

	Method	St.A-ℓ_∞	Adap.A
SVHN	Semi-SL* (Mao et al. 2021)	**62.12**	-
	PR18* (Rice et al. 2020)	**61.00**	-
	WR28* (Rebuffi et al. 2021)	**61.09**	-
	WR28 (Lee and Kim 2023)	-	**49.65**
	ResNet-Tiny (**ours**)	72.72	**64.70**
CIFAR-10	WRN28 (Shi et al. 2021)	53.58	**3.70**
	ResNet (Song et al. 2018)	46.00	**9.00**
	WR28 (Hill et al. 2021)	78.91	**54.90**
	WR28 (Yoon et al. 2021)	85.45	**33.70**
	PR18 (Mao et al. 2021)	**34.40**	-
	Semi-SL* (Mao et al. 2021)	64.44	**58.40**
	WR34* (Madry et al. 2018)	**44.04**	-
	WR34* (Zhang et al. 2019)	**53.08**	-
	WR70* (Wang et al. 2023)	**70.69**	-
	WR70* (Rebuffi et al. 2021)	**66.56**	-
	PR18* (Rebuffi et al. 2021)	**58.50**	-
	WR70 (Nie et al. 2022)	71.29	**51.13**
	WR70 (Lee and Kim 2023)	70.31	**56.88**
	ResNet50 (**ours**)	63.09	**57.15**

Table 3: Benchmark on SVHN and CIFAR-10 with $\ell_\infty = 8/255$. Accuracy is directly reported from the respective paper except for the adaptive attack which is reported from Lee and Kim (2023) for diffusion-based purification and Athalye et al (2018); Croce et al. (2022) for BPDA. Numbers in **bold** are the minimum/robust accuracy. "$*$": adversarially trained models, "-": missing from the references, "St.A": strongest standard attacks and "Adap.A": strongest adaptive attacks. "WR": Wide-ResNet. "PR": PreActResNet. Other works follow different evaluation standards, and we compare with them in Yang et al. (2023). We achieves robust performance without adversarial training.

	V-CLF		+TP (E)		
Dataset	Clean	AA-ℓ_∞	Clean	AA-ℓ_∞	BPDA
CIFAR100	72.37	0.10	42.96	**26.13**	**16.87**
ClbA-64^2	97.86	0.28	95.36	**90.34**	**73.77**
ClbA-128^2	97.78	0.00	96.81	**93.91**	**74.02**

Table 4: Classification accuracy on CIFAR-100 and CelebA (size: ClbA-x^2) using the VAE-Classifier with $\ell_\infty = 8/255$.

curacy from our experiments in Tables 1-2. Although the BPDA attack is the strongest attack compared with the standard adversarial attacks and the multi-objective attacks, our defense still achieves desirable adversarial robustness. In our experiments, models with larger backbones are more robust to the BPDA attacks. Compared with others in Table 3, we achieves superior performance against the BPDA attacks without adversarial training.

Effects of Hyperparameters. We study the effects of the purification budgets ϵ_{pfy} on clean classification accuracy and adversarial robustness. We use the VAE-Classifier (tiny ResNet encoder) on CIFAR-10 and show results in Figure 7 that our defense can provide adversarial robustness with various hyperparameters. See Yang et al. (2023) for details.

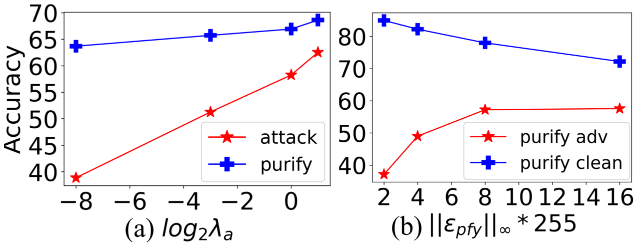


Figure 7: Accuracy affected by (a) the trade-off term λ_a of the multi-objective attacks, (b) the purification budget $\|\epsilon_{\text{pfy}}\|_\infty$ when $\|\boldsymbol{\delta}_{\text{adv}}\|_\infty = 8/255$.

Larger Datasets. We use larger datasets to study the impacts of image resolution (CelebA 64×64 and 128×128) and number of classes (CIFAR-100) on our defense. Table 4 indicates the scalability of our method for high-resolution data. However, the performance on data with a larger number of classes is limited as an accurate estimation of $p(\mathbf{z}|\mathbf{x})$ for each class is required. Compared with 5,000 training images per class in CIFAR-10, CIFAR-100 only provides 500 per class, leading to a less accurate estimation of $p(\mathbf{z}|\mathbf{x})$. See Yang et al. (2023) for more discussion of this perspective.

Theory and Experiments. The theory in our methodology section also provides insights for the experimental analysis. For example, CIFAR-10 and SVHN have the same number of classes and dimensions, but our method shows stronger robustness on SVHN. The insight is that reconstruction errors are smaller since the manifold of SVHN is easier to model. Meanwhile, the Standard-AE-Classifier is less robust compared with the VAE-Classifier since the classifier and the decoder of the Standard-AE-Classifier are not semantically consistent. However, the accuracy drops on clean data is not an implication of the theory. There could be multiple reasons for such phenomena. First, optimizing the ELBO for clean images causes distribution drifts. Second, there is a tradeoff between robustness and accuracy (Tsipras et al. 2019). Third, the function generating the purified signals should be locally Lipschitz continuous (Remark 1); however, such property is not guaranteed in Eq. (7). Consequently, the purification process may move samples to unstable regions causing the drop. One can apply our method only when attacks are detected to avoid the accuracy drop.

Conclusion

We formulate a novel adversarial purification framework via manifold learning and variational inference. Our test-time purification is evaluated with several attacks and shows its adversarial robustness for non-adversarially trained models. Even if attackers are aware of our defense method, we can still achieve competitive adversarial robustness. Our method is also capable of being combined with adversarially trained models to further increase their adversarial robustness.

Acknowledgments

Research is supported by DARPA Geometries of Learning (GoL) program under the agreement No. HR00112290075.

The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. government.

References

Athalye, A.; Carlini, N.; and Wagner, D. A. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, 274–283. PMLR.

Bastani, O.; Ioannou, Y.; Lampropoulos, L.; Vytiniotis, D.; Nori, A. V.; and Criminisi, A. 2016. Measuring Neural Net Robustness with Constraints. In *NIPS*, 2613–2621.

Carlini, N.; and Wagner, D. A. 2017. Towards Evaluating the Robustness of Neural Networks. In *IEEE Symposium on Security and Privacy*, 39–57. IEEE Computer Society.

Chen, J.; and Gu, Q. 2020. RayS: A Ray Searching Method for Hard-label Adversarial Attack. In *KDD*, 1739–1747. ACM.

Croce, F.; Gowal, S.; Brunner, T.; Shelhamer, E.; Hein, M.; and Cemgil, A. T. 2022. Evaluating the Adversarial Robustness of Adaptive Test-time Defenses. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 4421–4435. PMLR.

Croce, F.; and Hein, M. 2020. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 2206–2216. PMLR.

Gong, Y.; Yao, Y.; Li, Y.; Zhang, Y.; Liu, X.; Lin, X.; and Liu, S. 2022. Reverse Engineering of Imperceptible Adversarial Image Perturbations. In *ICLR*. OpenReview.net.

Goodfellow, I. J.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR (Poster)*.

Grathwohl, W.; Wang, K.; Jacobsen, J.; Duvenaud, D.; Norouzi, M.; and Swersky, K. 2020. Your classifier is secretly an energy based model and you should treat it like one. In *ICLR*. OpenReview.net.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *CVPR*, 770–778. IEEE Computer Society.

Hill, M.; Mitchell, J. C.; and Zhu, S. 2021. Stochastic Security: Adversarial Defense Using Long-Run Dynamics of Energy-Based Models. In *ICLR*. OpenReview.net.

Huang, S.; Lu, Z.; Deb, K.; and Boddeti, V. N. 2022a. Revisiting Residual Networks for Adversarial Robustness: An Architectural Perspective. *arXiv preprint arXiv:2212.11005*.

Huang, Z.; Liu, C.; Salzmann, M.; Süsstrunk, S.; and Zhang, T. 2022b. Improving Adversarial Defense with Self-supervised Test-time Fine-tuning.

Hwang, U.; Park, J.; Jang, H.; Yoon, S.; and Cho, N. I. 2019. PuVAE: A Variational Autoencoder to Purify Adversarial Examples. In *IEEE Access*, volume 7, 126582–126593.

Kim, H. 2020. Torchattacks: A pytorch repository for adversarial attacks. *arXiv preprint arXiv:2010.01950*.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In *ICLR*.

Krizhevsky, A.; and Hinton, G. 2009. Learning multiple layers of features from tiny images. Technical Report 0, University of Toronto, Toronto, Ontario.

LeCun, Y.; Cortes, C.; and Burges, C. 2010. MNIST handwritten digit database. *ATT Labs [Online]. Available: http://yann.lecun.com/exdb/mnist*, 2.

Lee, M.; and Kim, D. 2023. Robust Evaluation of Diffusion-Based Adversarial Purification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 134–144.

Leino, K.; Wang, Z.; and Fredrikson, M. 2021. Globally-Robust Neural Networks. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 6212–6222. PMLR.

Lin, W.; Lau, C. P.; Levine, A.; Chellappa, R.; and Feizi, S. 2020. Dual Manifold Adversarial Robustness: Defense against Lp and non-Lp Adversarial Attacks. In *NeurIPS*.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Madry, A.; Makelov, A.; Schmidt, L.; Tsipras, D.; and Vladu, A. 2018. Towards Deep Learning Models Resistant to Adversarial Attacks. In *ICLR (Poster)*. OpenReview.net.

Mao, C.; Chiquier, M.; Wang, H.; Yang, J.; and Vondrick, C. 2021. Adversarial Attacks are Reversible with Natural Supervision. In *ICCV*, 641–651. IEEE.

Meng, D.; and Chen, H. 2017. MagNet: A Two-Pronged Defense against Adversarial Examples. In *CCS*, 135–147. ACM.

Nayak, G. K.; Rawal, R.; and Chakraborty, A. 2022. DAD: Data-free Adversarial Defense at Test Time. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, 3788–3797. IEEE.

Netzer, Y.; Wang, T.; Coates, A.; Bissacco, A.; Wu, B.; and Ng, A. Y. 2011. Reading digits in natural images with unsupervised feature learning. In *NIPS Workshop on Deep Learning and Unsupervised Feature Learning*.

Nie, W.; Guo, B.; Huang, Y.; Xiao, C.; Vahdat, A.; and Anandkumar, A. 2022. Diffusion Models for Adversarial Purification. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 16805–16827. PMLR.

Pang, T.; Lin, M.; Yang, X.; Zhu, J.; and Yan, S. 2022. Robustness and Accuracy Could Be Reconcilable by (Proper) Definition. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, 17258–17277. PMLR.

Pang, T.; Yang, X.; Dong, Y.; Su, H.; and Zhu, J. 2021. Bag of Tricks for Adversarial Training. In *ICLR*. OpenReview.net.

Patel, K.; Beluch, W.; Zhang, D.; Pfeiffer, M.; and Yang, B. 2020. On-manifold Adversarial Data Augmentation Improves Uncertainty Calibration. In *ICPR*, 8029–8036. IEEE.

Pérez, J. C.; Alfarra, M.; Jeanneret, G.; Rueda, L.; Thabet, A. K.; Ghanem, B.; and Arbeláez, P. 2021. Enhancing Adversarial Robustness via Test-time Transformation Ensembling. In *ICCVW*, 81–91. IEEE.

Rauber, J.; Brendel, W.; and Bethge, M. 2017. Foolbox: A Python toolbox to benchmark the robustness of machine learning models. In *Reliable Machine Learning in the Wild Workshop, 34th International Conference on Machine Learning*.

Rebuffi, S.; Gowal, S.; Calian, D. A.; Stimberg, F.; Wiles, O.; and Mann, T. A. 2021. Fixing Data Augmentation to Improve Adversarial Robustness. *CoRR*, abs/2103.01946.

Rice, L.; Wong, E.; and Kolter, J. Z. 2020. Overfitting in adversarially robust deep learning. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, 8093–8104. PMLR.

Samangouei, P.; Kabkab, M.; and Chellappa, R. 2018. Defense-GAN: Protecting Classifiers Against Adversarial Attacks Using Generative Models. In *ICLR (Poster)*. OpenReview.net.

Sehwag, V.; Mahloujifar, S.; Handina, T.; Dai, S.; Xiang, C.; Chiang, M.; and Mittal, P. 2022. Robust Learning Meets Generative Models: Can Proxy Distributions Improve Adversarial Robustness? In *ICLR*. OpenReview.net.

Shi, C.; Holtz, C.; and Mishne, G. 2021. Online Adversarial Purification based on Self-supervised Learning. In *ICLR*. OpenReview.net.

Song, Y.; Kim, T.; Nowozin, S.; Ermon, S.; and Kushman, N. 2018. PixelDefend: Leveraging Generative Models to Understand and Defend against Adversarial Examples. In *ICLR (Poster)*. OpenReview.net.

Stutz, D.; Hein, M.; and Schiele, B. 2019. Disentangling Adversarial Robustness and Generalization. In *CVPR*, 6976–6987. Computer Vision Foundation / IEEE.

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I. J.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR (Poster)*.

Tsipras, D.; Santurkar, S.; Engstrom, L.; Turner, A.; and Madry, A. 2019. Robustness May Be at Odds with Accuracy. In *ICLR (Poster)*. OpenReview.net.

Wang, Z.; Pang, T.; Du, C.; Lin, M.; Liu, W.; and Yan, S. 2023. Better Diffusion Models Further Improve Adversarial Training. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, 36246–36263. PMLR.

Willetts, M.; Camuto, A.; Rainforth, T.; Roberts, S. J.; and Holmes, C. C. 2021. Improving VAEs' Robustness to Adversarial Attack. In *ICLR*. OpenReview.net.

Wong, E.; and Kolter, J. Z. 2018. Provable Defenses against Adversarial Examples via the Convex Outer Adversarial Polytope. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, 5283–5292. PMLR.

Xiao, C.; Chen, Z.; Jin, K.; Wang, J.; Nie, W.; Liu, M.; Anandkumar, A.; Li, B.; and Song, D. 2022. DensePure: Understanding Diffusion Models towards Adversarial Robustness. *CoRR*, abs/2211.00322.

Xiao, H.; Rasul, K.; and Vollgraf, R. 2017. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*.

Yang, Z.; Xu, Z.; Zhang, J.; Hartley, R.; and Tu, P. 2023. Adversarial Purification with the Manifold Hypothesis. *arXiv preprint arXiv:2210.14404*.

Yin, S.; Zhang, X.; and Zuo, L. 2022. Defending against adversarial attacks using spherical sampling-based variational auto-encoder. *Neurocomputing*, 478: 1–10.

Yoon, J.; Hwang, S. J.; and Lee, J. 2021. Adversarial Purification with Score-based Generative Models. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, 12062–12072. PMLR.

Zhang, H.; Yu, Y.; Jiao, J.; Xing, E. P.; Ghaoui, L. E.; and Jordan, M. I. 2019. Theoretically Principled Trade-off between Robustness and Accuracy. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, 7472–7482. PMLR.

Zhao, L.; Liu, T.; Peng, X.; and Metaxas, D. N. 2020. Maximum-Entropy Adversarial Data Augmentation for Improved Generalization and Robustness. In *NeurIPS*.

Zhou, J.; Liang, C.; and Chen, J. 2020. Manifold Projection for Adversarial Defense on Face Recognition. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, 288–305. Springer.

Zhou, Y. 2022. Rethinking Reconstruction Autoencoder-Based Out-of-Distribution Detection. In *CVPR*, 7369–7377. IEEE.