Eliminating Adversarial Noise via Information Discard and Robust Representation Restoration

Dawei Zhou [*1] **Yukun Chen** [*1] **Nannan Wang** [1] **Decheng Liu** [1] **Xinbo Gao** [2] **Tongliang Liu** [3 4]

Abstract

Deep neural networks (DNNs) are vulnerable to adversarial noise. Denoising model-based defense is a major protection strategy. However, denoising models may fail and induce negative effects in fully white-box scenarios. In this work, we start from the latent inherent properties of adversarial samples to break the limitations. Unlike solely learning a mapping from adversarial samples to natural samples, we aim to achieve denoising by *destroying the spatial characteristics of adversarial noise and preserving the robust features of natural information*. Motivated by this, we propose a defense based on information discard and robust representation restoration. Our method utilize complementary masks to disrupt adversarial noise and guided denoising models to restore robust-predictive representations from masked samples. Experimental results show that our method has competitive performance against white-box attacks and effectively reverses the negative effect of denoising models.

1. Introduction

Deep neural networks (DNNs) have achieved great success in many fields, such as computer vision (He et al., 2016; Dosovitskiy et al., 2020) and speech recognition (Wang et al., 2017). However, DNNs were found to be vulnerable to adversarial samples which were crafted by adding imperceptible but adversarial noise on natural samples (Szegedy et al., 2014; Goodfellow et al., 2015). This vulnerability raised security concerns about the reliability of DNNs in decision-critical deep learning applications.

A major class of adversarial defense pre-processes input samples to alleviate the interference of adversarial noise In particular, the denoising model-based defense strategy exploited DNNs to remove adversarial noise and restore natural information of original natural samples (Liao et al., 2018; Jin et al., 2019; Naseer et al., 2020). In addition, this strategy was typically scalable, *i.e.*, the defense can be deployed to different tasks without retraining the target model (Naseer et al., 2020). Thus, defensive denoising model have shown great potential to safeguard target models from adversarial attacks.

However, experienced attackers usually considered the possibility that defenses being used (Athalye & Carlini, 2018) and deployed powerful attacks to disrupt defenses. Recent researches (Carlini & Wagner, 2017a; Tramer et al., 2020) showed that the protection for general target models by defensive denoising models was significantly weaker in a fully white-box scenario (*i.e.*, the attacker had access to all information about the defense and target models). In addition, the studies in Zhou et al. (2021) revealed a *robustness degradation effect*. That is, for adversarially trained target models, applying a defensive denoising model reduced rather than improved the robust accuracy against white-box attacks. The white-box attacks here aimed at disrupting the *overall model* (*i.e.*, the ensemble of the denoising model and target model). Moreover, simply incorporating defensive denoising models into adversarial training cannot effectively handle this negative effect (Zhou et al., 2021).

To solve above problems, we sought breakthroughs from the latent inherent properties of adversarial samples. On the one hand, considering the perturbations carefully crafted by adversarial attacks had spatial characteristics (Qian & Tolkova; Tsipras et al., 2018; Aydin et al., 2021; Wang et al., 2022), adversarial noise might lose its destructiveness when it was spatially corrupted. We conducted a proof-of-concept experiment in Section. 3.2 and its result showed that *the interference of adversarial noise was effectively suppressed via spatial information discard*. On the other hand, adversarial samples retained a lot of natural information from original natural samples (called retained natural informa-

[*]Equal contribution [1]School of Telecommunications Engineering, State Key Laboratory of Integrated Services Networks, Xidian University, Xian, Shaanxi, China [2]Chongqing Key Laboratory of Image Cognition, Chongqing University of Posts and Telecommunications, Chongqing, China [3]Mohamed bin Zayed University of Artificial Intelligence, Masdar City, Abu Dhabi, United Arab Emirates [4]University of Sydney, Darlington, NSW, Australia. Correspondence to: Nannan Wang <nnwang@xidian.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

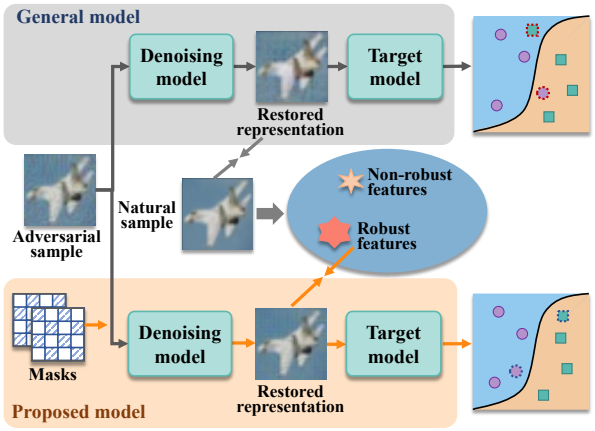


Figure 1. Difference between the proposed denoising model and general denoising model. Instead of directly bring the restored representations close to the natural samples, the proposed approach was dedicated to disrupting adversarial noise via complementary masks, and guiding the restored representations to sufficiently capture robust and predictive features in natural information.

tion). This retained information naturally had non-robust and robust features from the natural samples (Ilyas et al., 2019). Non-robust features were susceptible to perturbations and led to wrong predictions while robust features were the opposite (Ilyas et al., 2019; Kim et al., 2021). The representations restored by denoising models were usually guided to be consistent with natural samples. They thus had non-robust features and were vulnerable to adversarial perturbations (like natural samples). Fortunately, the retained natural information contained non-robust features. If we can *emphatically restore robust representations from latent robust features of retained natural information* in the denoising process, the target model will hardly be misled by the perturbations on restored representations, *i.e.*, achieving reliable representation restoration. This will be beneficial to enhance the effectiveness of denoising models.

Motivated by above inspirations, we proposed a *defensive denoising model based on information discarding and robust representation restoration (DIR)*. As shown in Figure. 1, instead of directly learning the mapping from adversarial samples to natural samples, our method aimed to actively suppress the destruction of adversarial noise and restore robust and predictive representations.

Specifically, we **first** performed *information discard on adversarial samples* by constructing complementary masks with fixed spatial ratios. Based on the spatial redundancy of images (Bastani et al., 2010; Hu et al., 2020; He et al., 2022) and the retained natural information in adversarial samples, we leveraged a masking-related generative network (He et al., 2022) as the denoising model to restore missing natural representation from preserved neighboring pixels. However, the restored representation might contain many non-robust features. **For this reason**, a *game-based mechanism was designed to promote the robustness of the representations*. We crafted specific perturbations against the representations by mining and exploiting their residual non-robust features. The distance between the original and perturbed representations in the high-level decision space was measured. By minimizing this distance, the denoising model was guided to restore robust representations from latent robust features in masked samples. **Moreover**, adversarial training on the denoising model was performed as a basis for achieving robustness in white-box scenarios. The adversarial samples were crafted against the overall model.

The main contributions in this paper are as follows:

- Starting from inherent properties of adversarial samples, we aimed to protect target models by destroying spatial characteristics of adversarial noise and emphatically preserving latent robust features of retained natural information during the denoising process.
- We proposed a defense method based on information discard and robust representation restoring. Our method utilized complementary masks to disrupt adversarial noise, and guided denoising models to restore robust-predictive representations from masked samples by designing specific noise against denoised samples.
- Quantitative experiments on white-box and adaptive attacks were performed. The results showed that the proposed method effectively defended against adversarial noise and handled the robustness degradation effect. Multiple ablation studies were conducted to comprehensively demonstrate the effectiveness of information discard and robust representation restoration.

2. Related works

Adversarial attacks. Adversarial samples were pioneeringly proposed by Szegedy et al. (2014), they were crafted by adding imperceptible but adversarial noise on natural samples. Many works generated adversarial noise along the directions of gradients which maximized loss functions. For example, the one-step fast gradient sign method (FGSM) (Goodfellow et al., 2015) method, the multiple-step projected gradient descent (PGD) method (Madry et al., 2018) (the strongest first-order attack) and the autoattack (AA) method (Croce & Hein, 2020b) which formed a parameter-free, computationally affordable and user-independent ensemble of attacks. In addition some optimization-based methods (Carlini & Wagner, 2017b; Rony et al., 2019; Croce & Hein, 2020a) minimized the size of adversarial noise while ensuring that target models were misleading. In this work, The adversarial attacks in white-box scenarios aimed

to disrupt the defensive denoising model while perturbing the target model.

Adversarial defenses. To mitigate the threat posed by adversarial noise, the researches on adversarial defenses have drawn increasing attention. Adversarial training (AT) was proposed as a typical adversarial defense strategy, which focuses on exploiting adversarial samples to help train the target model and has shown remarkable effectiveness (Madry et al., 2018; Zhang et al., 2019; Wang et al., 2019; Wu et al., 2020a;b). Another major class of adversarial defense is pre-processing strategy, which typically has higher scalability than AT (Naseer et al., 2020) Among pre-processing defenses, compared with feature squeezing (Guo et al., 2018; Xu et al., 2017) and adversarial detection (Ma et al., 2018; Qin et al., 2019), denoising models (Liao et al., 2018; Hill et al., 2021) can remove adversarial noise and effectively retain the original information of natural samples.

Denoising model-based methods usually exploit DNNs to learn a mapping from adversarial samples to natural samples. However, experienced attackers may have access to defenses through model stealing or other malicious behaviors. The studies in Zhou et al. (2021) have shown that some defensive denoising models were broken by white-box attacks and even provided negative defense for adversarially trained target models, such as APE-GAN (Jin et al., 2019), high-level representation guided denoiser (HGD) (Liao et al., 2018) and neural representation purifier (NRP) (Naseer et al., 2020). In this work, we are committed to improving the effects of defensive denoising model in the white-box scenario. We achieve effective denoising by discarding spatial information of adversarial noise and restoring robust representation from retained natural information. Moreover, some works (Xia et al., 2019; Wu et al., 2021; Xia et al., 2022; Li et al., 2022) inferred natural data by learning a transition relationship between adversarial data and natural data. In future work, our method was expected to combine with this mechanism to explore more effective denoising-based defenses.

Randomized smoothing. Randomized smoothing samples K Gaussian noises for one input sample, and our method samples K masks for one input sample. Both our method and randomized smoothing use the sampling operation. However, our method also has some differences from randomized smoothing. On the one hand, our approach aims to improve the robustness of the target model by removing the adversarial noise through denoising the model. Randomized smoothing aims to use samples with Gaussian noise to train a smoothed classifier. The optimization object of our method is not the same as that of randomized smoothing. On the other hand, the proposed method samples the mask but random smoothing samples the Gaussian noise. We use the mask to discard part of the adversarial noise in adversarial samples to destroy its structure while preserving sufficient useful information for sample restoration. This is different from random smoothing that adds Gaussian noise to the original natural samples.

3. Methodology

3.1. Notation

The random variables and their specific realization were denoted by *capital* letters and *lower-case* letters, respectively. Let X, Y represent the variables for natural samples and corresponding labels, and X', Y' represent the variables for adversarial samples and corresponding adversarial labels. Variables (X, Y) constituted a data distribution, where $(X, Y) \in \mathcal{X} \times \{1, 2, \ldots, C\}$, $\mathcal{X}$ was the feature space of X and C was the number of label categories. A set of natural examples $\{(x_i, y_i)\}_{i=1}^{N}$ was sampled from (X, Y), where n was the number of examples. We defined a classification function as $f : \mathcal{X} \to \{1, 2, \ldots, C\}$ and a denoising function as $g : \mathcal{X}' \to \mathcal{X}$, where $\mathcal{X}'$ denoted the feature space of the variable for adversarial samples. The classification and denoising functions can be parameterized via deep neural networks (DNNs). We used h_θ and g_ω as the classification model (*i.e.*, the target model) and the denoising model respectively, where θ and ω are the model parameters. Given the classification model h_θ, denoising model h_ω and a natural example (x, y), the adversarial sample x' in the white-box scenario was formulated as follows:

$$softmax(h_\theta(g_\omega(x'))) \neq \boldsymbol{y} \quad \text{s.t.} \quad \|x - x'\| \leq \epsilon, \quad (1)$$

where $\boldsymbol{y}$ denoted the label y in the one-hot vector form and $\|\cdot\|$ denoted the norm (*e.g.*, L_∞-norm: $\|\cdot\|_\infty$).

3.2. Motivation

DNNs are facing threats from widespread adversarial attacks. As protectors, adversarial defenses need to take into account worst-case scenarios, such as white-box scenarios where both target models and defense models are leaked to the attackers). The protections provided by the defensive denoising models for the naturally trained target models can be significantly broken by the white-box attacks (Carlini & Wagner, 2017a; Tramer et al., 2020). Even, for robust target models (trained by standard AT (Madry et al., 2018), TRADES (Ding et al., 2019) or MART (Wang et al., 2019)), deploying a denoising model can be counterproductive, leading to a decrease in robust accuracy (Zhou et al., 2021). This may be due to the fact that the denoising model suffers from the corruption of white-box attacks and generates apparently anomalous textures (Kos et al., 2018; Ru et al., 2019; Sun et al., 2020).

A straightforward solution was to construct adversarial samples against the overall model and used them to train the

denoising model. Unfortunately, although the adversarially trained denoising model can alleviate the apparently anomalous textures, the generated representations still contained residual noise. This approach improved the adversarial accuracy, but it still lagged behind the performance of a single adversarially trained target model (see the experiment in Section. 4.3). To enhance the abilities of denoising models, we explored the potential inherent properties of adversarial samples from the perspectives of adversarial noise and retained natural information to handle above issues.

(I) Spatial characteristics of adversarial noise. Previous studies showed that carefully crafted adversarial noise often exhibited potential spatial characteristics (even if adversarial attacks do not involve spatially specialized designs) (Tsipras et al., 2018; Aydin et al., 2021). For the spatial semantics of the objective, the adversarial noise usually had corresponding perturbations in the global structure (Yang et al., 2019; Wang et al., 2022). Given that denoising models had the ability to learn spatial information (*e.g.*, preserving the spatial semantics of objectives from input samples), *the spatial characteristics of adversarial noise might be carried over into the residual noise*.

If the adversarial noise in input samples was corrupted in pixel space, the spatial characteristics of residual noise in restored representations will be affected, and the interference to target models will be reduced as a result. We conducted a proof-of-concept experiment to validate this hypothesis. More details were presented in Appendix. A. As shown in Figure. 2, we removed the adversarial noise at different spatial rates and calculated attack success rates against naturally and adversarially trained models. The evaluations on several different attack algorithms showed that the interference of adversarial noise was significantly reduced when it was spatially removed by more than 50%.

(II) Latent features of retained natural information. For adversarial samples, they were similar to natural samples in terms of human vision and can be correctly classified by humans. This indicated that adversarial samples retained a lot of information from natural samples (called natural information). The studies in Ilyas et al. (2019) had indicated that natural samples had latent non-robust features, which were easily disturbed by small perturbations to affect the prediction of models (Ilyas et al., 2019; Kim et al., 2021). Denoising models usually guided the restored representations to be consistent with natural samples for obtaining correct predictions. However, the reliability of restored representations had not been well considered. In the absence of specific constraints, the restored representations were prone to preserve non-robust features from retained natural information due to their close resemblance to natural samples. *The residual noise can exploit the latent non-robust features in the restored representations to mislead target models*.

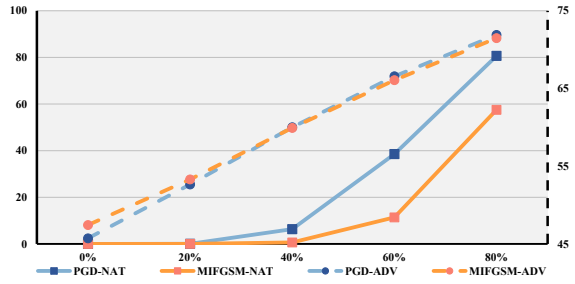


Figure 2. The impact of spatial disruption on adversarial noise and natural information. We randomly discarded pixels with different spatial ratios (horizontal coordinate) and then calculated the robust accuracy of the naturally trained target model (solid line) and the target model trained by standard AT (Madry et al., 2018) (dotted line). The results reflected that the destructiveness of adversarial noise (crafted by PGD and MIFGSM (Dong et al., 2018)) was significantly weakened when it was discarded by more than 50%.

Fortunately, in addition to non-robust features, natural samples also contained latent robust features and the retained natural information naturally had sufficient robust features. These features were predictive and can guide target models to correctly classify objectives in noisy environments (Ilyas et al., 2019). In the denoising process, if robust features can be maintained while non-robust features were marginalized, the restored representations were expected to be robust to target models, *i.e.*, perturbations on restored representations will not break target models. This was beneficial to improve the protection ability of denoising models.

3.3. The proposed defense

Motivated by above inspirations, we proposed a denoising model-based defense based on two main modules: *information discard* and *robust representation restoration*. The schematic diagram of our method was shown in Figure. 3.

Information discard. Our method destroyed the spatial characteristics of adversarial noise by discarding the spatial information of adversarial samples. We randomly constructed complementary masks $\{m\}_1^K$ (K is the number of masks) at a fixed ratio, and respectively overlaid them on the input adversarial sample x' to achieve the information discard. Of course, part of the natural information retained in the adversarial sample was also lost, which may lead to the failure of classification. Fortunately, in view of the spatial redundancy of images, the missing spatial semantics can be recovered and denoised by the global position of each patch in the complete image using a deep neural network with the pixel information of the patch itself., *e.g.*, the masked autoencoder (He et al., 2022).

Based on this, we leveraged a masking-related generative network as the backbone of the denoising model g_ω. To

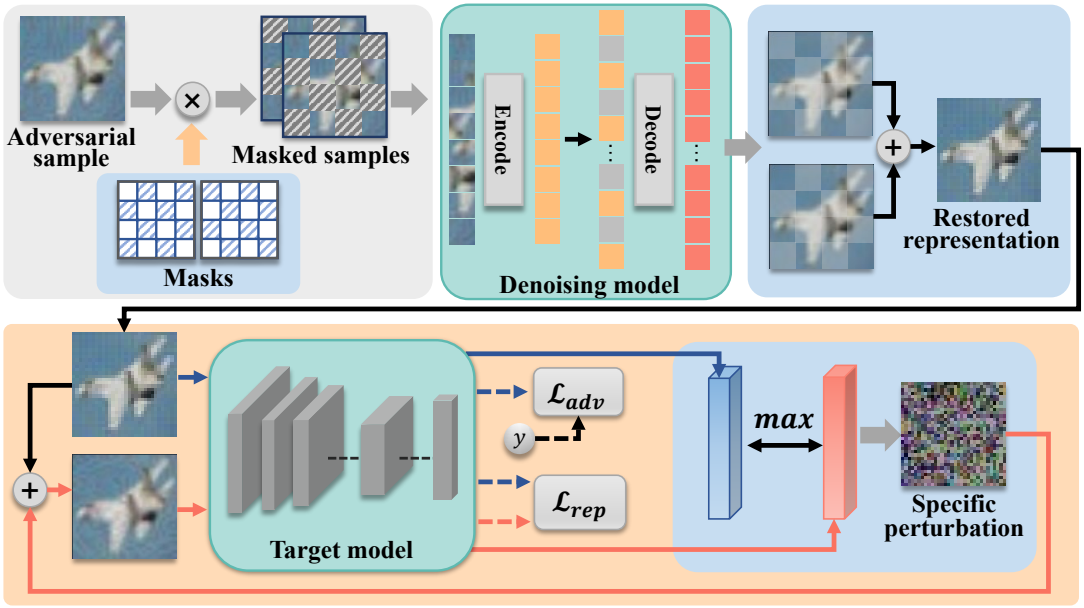


Figure 3. The schematic diagram of the proposed method. Our method constructed a defensive denoising model based on information discarding and robust representation restoration (DIR). The information discard was performed to destroy the spatial characteristics of adversarial noise by introducing random and complementary masks on adversarial samples. To restore robust representations from the latent robust features of natural information, a game-based mechanism was designed. We crafted specific perturbations on restored representations to maximize the distance (*e.g.*, KL divergence) between original representations and perturbed representations in the decision space (see blue box at the bottom). Adversarially, the denoising model was optimized to minimize the distance by minimizing $\mathcal{L}_{rep}$. In addition, we performed adversarial training on the denoising model by minimizing $\mathcal{L}_{adv}$. The adversarial samples used are crafted against the ensemble of the denoising model and target model.

drive the denoising model to achieve the basic representation restoration and denoising, we pre-trained the denoising model according to the following optimization objective:

$$\min_{\omega} \sum_{i=1}^{N} \sum_{k=1}^{K} \left\| g_\omega \left(x'_i \times m_i^k \right) - x_i \right\|_2^2, \tag{2}$$

where m_i^k denoted the $k-th$ constructed mask for x'_i. Note that the denoising model was not adversarially trained here, and the used adversarial samples x' are only crafted against the target model h_θ. The purpose of pre-training was only to facilitate the ability of the denoising model to extract the effective objective semantics from the remaining neighboring pixels (as was done in (He et al., 2022)).

Robust representation restoring. For a denoising model, robust restored representations was defined to satisfy the following condition:

$$h_\theta(\tilde{x} + \delta) = h_\theta(\tilde{x}), \tag{3}$$

where $\tilde{x} = g_\omega(x' \times m)$ and δ denoted the perturbation crafted against the target model. That is, perturbations deployed on the robust restored representations did not manipulate the predictions of the target model. However, when the restored representations were closely aligned with the natural samples, the non-robust features from the retained natural information were easily be exploited to induce adversarial perturbations. We aimed to guide the denoising model to reduce the involvement of non-robust features and focus on restoring representations from potential robust features. Unfortunately, the distance between the restored representations and the latent robust features of natural information cannot be measured directly. We therefore designed a game-based mechanism to indirectly guide the restoration of robust representations.

Based on the definition of robust restored representations, we naturally constructed specific perturbations on the restored representations according to the following objective:

$$\max_{\delta} KL(h_\theta(\tilde{x} + \delta), h_\theta(\tilde{x})), \tag{4}$$

where $KL(\cdot, \cdot)$ denoted Kullback-Leibler Divergence. The perturbations mined the residual non-robust features in the restored representations and exploited them to maximize the anomalies in the high-level feature space of the target model. Conversely, the denoising needed to block the aggressive behavior of specific perturbations. It thus worked to minimize the distance between the original representations

Algorithm 1 Defensive denoising model based on information discarding and robust representation restoration (DIR).

Input: Target model h with adversarially trained parameters θ, denoising model g parameterized by ω, batch size n, epoch number E, training dataset $\mathcal{D}$ and perturbation budget ϵ.

Pre-train g_ω by using natural samples and adversarial samples crafted against h_θ according to Eq. 2;
for $e = 1$ to E **do**
  Read mini-batch $\mathcal{B} = \{x_i\}_{i=1}^n$ from training set $\mathcal{D}$;
  Randomly construct a group of complementary masks $\{m_i^k\}_{k=1}^K$ for each sample x_i in $\mathcal{B}$;
  Craft white-box adversarial samples $\{x_i'\}_{i=1}^n$ against the ensemble of g_ω and h_θ at the given perturbation budget ϵ for each sample x_i in $\mathcal{B}$;
  for $i = 1$ to n (in parallel) **do**
    Forward-pass $\{x_i' \times m_i^k\}_{k=1}^K$ through g_ω and obtain restored representation $\tilde{x}_i$;
    Craft specific perturbation δ_i on $\tilde{x}_i$ via Eq. 4;
  end for
  Calculate $\mathcal{L}_{rep}$, $\mathcal{L}_{adv}$ and $\mathcal{L}_{den}$ using Eq. 5, Eq. 6 and Eq. 7, respectively;
  Back-pass and update ω;
end for.

and perturbed representations in the feature space. The loss function for the denoising model is shown as follows:

$$\begin{aligned} \mathcal{L}_{rep}(x', m; \omega) &= \sum_{i=1}^{N} KL\left(h_\theta\left(\tilde{x}_i + \delta_i\right), h_\theta\left(\tilde{x}_i\right)\right), \\ \tilde{x}_i &= \frac{1}{K}\sum_{k=1}^{K} g_\omega\left(x_i' \times m_i^k\right), \end{aligned} \tag{5}$$

where δ_i denoted the specific perturbation on $\tilde{x}_i$. The restored representation was obtained by fusing the representations generated for the different masks and averaging them. By alternatively gaming, the denoising model was expected to focus on preserving potential robust features of retained natural information and have the ability to restoring robust representations.

Adversarial training on denoising model. In addition to the information discard and robust representation restoration, as a basis, we performed adversarial training on the denoising model. The adversarial training data x' was generated via PGD attack against the ensemble of the denoising model g_ω and the target model h_θ. The adversarial loss for the denoising model was formulated as:

$$\mathcal{L}_{adv}(x', m; \omega) = \sum_{i=1}^{N}\sum_{c=1}^{C} y_i^c \log\left(p^c\left(\tilde{x}_i\right)\right), \tag{6}$$

where $p^c(\cdot)$ denoted the operation to obtain the predicted probability corresponding class c. The overall loss function used to optimize the denoising model was formulated as:

$$\mathcal{L}_{den} = \mathcal{L}_{adv} + \alpha \cdot \mathcal{L}_{rep}, \tag{7}$$

where α was the trade-off hyperparameters. Detailed settings were presented in Section. 4.1 and the algorithm of our method was shown in Algorithm. 1. The code can be found in https://github.com/chenyyyykun/DIR.

4. Experiments

4.1. Experiment settings

Target models. In this work, we conducted experiments on the *SVHN* (Netzer et al., 2011) and *CIFAR-10* (Krizhevsky et al., 2009) datasets. For both datasets, a ResNet-18 network (He et al., 2016) was utilized as the backbone of the target model. We used standard AT strategy to train the target model and utilized this target model to participate in the training of the denoising model. The perturbation budget is 8 / 255, the step number is 5 and the step size is 2 / 255. In addition, we used TRADES (Ding et al., 2019) and MART (Wang et al., 2019) strategies to train target models respectively and employed them to evaluate the defenses.

Defense models. Considering that our method contained an information discard module, we needed a mask-related generative network for denoising and representation restoration. Referring to the works in Dosovitskiy et al. (2020), we adopted a vision transformer block-based network (Vaswani et al., 2017) as the backbone of the denoising model. Meanwhile, according to the conclusions in simMIM (Xie et al., 2022), we choose a spatial ratio of 50% to construct masks, *i.e.*, $K = 2$. We used four denoising model-based methods as the baselines. They were APE-GAN (Jin et al., 2019), high-level representation guided denoiser (HGD) (Liao et al., 2018), neural representation purifier (NRP) (Naseer et al., 2020) and joint adversarial training based pre-processing (JATP) (Zhou et al., 2021). The denoising models of these methods were trained by using adversarial samples against the adversarially trained target model (trained by standard AT). The other settings of baselines were consistent with those on CIFAR-10 in their original papers. For our method, the denoising model was trained using SGD (Andrew & Gao, 2007) with momentum 0.9, weight decay 2×10^{-4}. The learning rate was initially set to 10^{-2}, which is divided by 10 at the $75-th$ and $90-th$ epoch. The epoch for the pre-processing process was 20. Referring to the early stopping strategy (Rice et al., 2020), the epoch number was set to 90 and the hyperparameter α is set to 2.0. The attack used in all methods was PGD attack (Madry et al., 2018) with a perturbation budget of 8 / 255, a step number of 5 (for faster training), and a step size of 2 / 255.

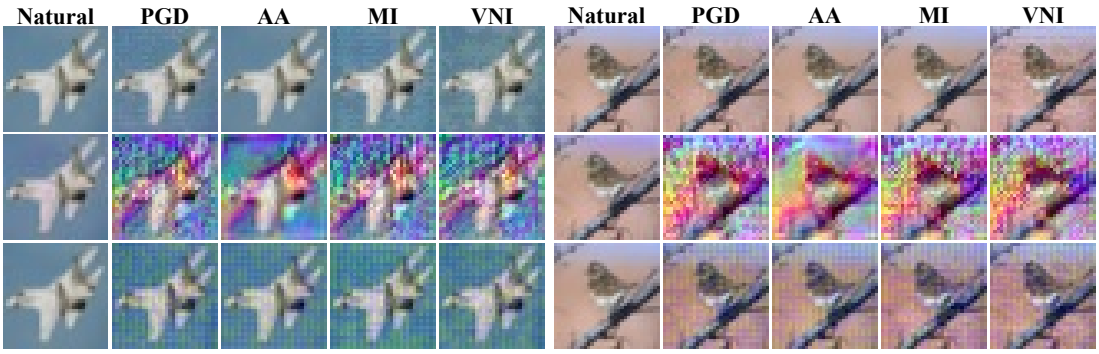


Figure 4. The examples of restored representations. The images from top to bottom are natural samples or white-box adversarial samples against our model, representations restored by a baseline (*e.g.*, HGD) and representations restored by our method. More details are presented in Appendix. B (*e.g.*, white-box adversarial samples against the model of HGD and restored representations of natural samples). It can be seen that our method effectively reduces the apparent anomalies in the restored representations.

4.2. Evaluations of defense effectiveness

Robust accuracy against white-box attacks. We first used white-box adversarial samples to evaluate the denoising model-based defenses. The adversarial samples were crafted against the ensemble of the denoising model and target model. Four strong adversarial attack algorithms were deployed, they were PGD, AA, MIFGSM (MI) (Dong et al., 2018) and VNIFGSM (VNI) (Wang & He, 2021). The main parameters of PGD, MI and VNI were set as follows: a L_∞-norm, a perturbation budget of 8 / 255, a step number of 20 and a step size of 2 / 255.

The restored representations are shown in Figure. 3.3. Our method effectively reduces the apparent anomalies in the restored representations, which indicates the higher reliability of our denoising model. Of course, there are local differences in our restored representations from the original natural samples. The color textures in many patches are altered. This may be because these information contains more non-robust features (Ilyas et al., 2019) and is therefore more likely to be perturbed by residual noise to mislead the target model. Our denoising model is guided to focus on latent robust features and may thus modify these information to restore robust representations. Quantitative results in Table. 1 also verify the effectiveness of our method. It can be seen that our method achieved better performances on both *CIFAR-10* and *SVHN*. As shown in the results of the third, fourth, fifth and sixth rows in the table, the robust degradation effect is induced in such white-box scenario, *i.e.*, their robust accuracy is lower than that of a single adversarially trained target model (the results in the *None* row). Our method reverses this negative effect and achieves higher robust accuracy. In addition, we observe the separation between different classes of restored representations under the white-box PGD attack via UMAP (McInnes et al., 2018). As shown in Figure. 5, our method presents a clearer separation

Table 1. The robust accuracy (percentage) against white-box attacks on *CIFAR-10* and *SVHN*. We showed the best result with **bold**. *None* denoted the target model without defensive denoising model, and it was adversarially trained by standard AT.

Dataset	Defense	Natural	PGD	AA	MI	VNI
CIFAR	None	**85.20**	45.72	42.26	47.43	47.39
	APE-GAN	84.79	28.65	12.91	30.02	29.96
	HGD	84.36	18.07	13.85	19.67	19.63
	NRP	84.53	21.90	15.32	22.43	22.37
	JATP	85.04	44.16	40.27	44.60	44.56
	DIR(Our)	85.11	**54.27**	**51.34**	**55.62**	**55.57**
SVHN	None	**91.27**	50.83	45.16	51.90	51.85
	APE-GAN	90.76	33.14	17.89	35.07	34.98
	HGD	90.64	27.49	20.37	28.83	28.76
	NRP	90.71	34.73	21.84	36.29	36.21
	JATP	90.82	47.72	43.87	49.34	49.27
	DIR(Our)	90.79	**55.30**	**51.61**	**56.70**	**56.64**

compared to the HGD defense. Moreover, we evaluated the cross-model defense effectiveness of our method by transferring the well-trained denoising model (for the target model adversarially trained by standard AT) to the target models adversarially trained by TRADES and MART, the results were shown in Table. 2.

Robust accuracy against adaptive attacks. Considering that masking, pre-processing and randomness are involved in our method, we utilized three adaptive attacks for further evaluation. We first designed a masking-related adaptive attack based on the following rule:

$$x^*_{t+1} = \Pi_{x+\mathcal{S}}\left(x^*_t + \tau \operatorname{sign}\left(\nabla_{x^*}\ell(h_\theta(g_\omega(x^*_t)), y)\right)\right), \quad (8)$$

where $x^* = x \times m$, S denoted a set of allowed perturbations and τ denoted the step size. This adaptive attack had access to the masks used in our defense and crafted targeted adversarial noise for each mask. The adversarial accuracy of our

Table 2. The robust accuracy (percentage) against white-box attacks on *CIFAR-10*. We transferred the denoising model to other adversarially trained target models.

Target model	Defense	Natural	PGD	AA
Trained by TRADES	None	**82.67**	49.59	46.63
	APE-GAN	82.14	31.70	13.59
	HGD	82.37	23.49	14.63
	NRP	82.23	27.65	14.90
	JATP	82.61	48.13	44.46
	DIR(Our)	82.53	**55.43**	**51.72**
Trained by MART	None	**81.93**	50.65	48.12
	APE-GAN	81.69	33.67	14.24
	HGD	81.64	25.01	15.17
	NRP	81.70	29.34	16.91
	JATP	81.85	48.30	45.16
	DIR(Our)	81.76	**56.04**	**52.23**

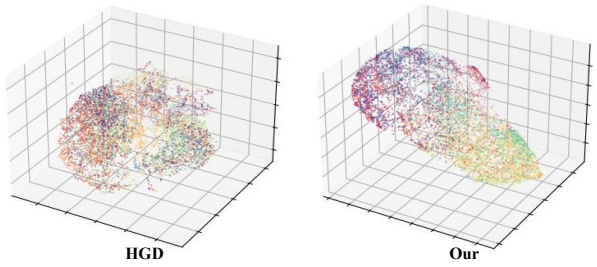


Figure 5. The separation between different classes of restored representations under the white-box PGD attack. Compared with the left panel, the points with the same color (*i.e.*, samples from the same category) were more aggregated in the right panel, and the points with different colors (*i.e.*, samples from different categories) were more clearly separated. This indicated that our method presented a better denoising effect for classification.

method against this adaptive attack (with a step number of 40 and a step size of 2 / 255) is 53.68% on *CIFAR-10*.

We then combined the backward pass differentiable approximation (BPDA) strategy (Athalye et al., 2018a) and PGD (with a step size of 40) to construct an adaptive attack. BPDA found a differentiable approximation such as $g^*(x) \approx g(x)$ and computed the gradient by replacing $g(x)$ with $g^*(x)$ on the backward pass. The BPDA strategy can be used on an arbitrary network, even if it is already differentiable, to obtain a more useful gradient (Athalye et al., 2018a). In addition, we leveraged the expectation over time (Athalye et al., 2018b) strategy to further target the randomness used in defenses. We toke two related defenses for comparison. ME-Net (Yang et al., 2019) did not have a DNN-based denoising model, but it utilized a random masking strategy. Similar to ME-Net, the work in Hill et al. (2021) constructed a energy-based model for adversarial purification and introduced random variation. The adversarial accuracy against adaptive attacks were shown in Table. 3.

Table 3. The robust accuracy (percentage) against adaptive attacks on *CIFAR-10*. Two defenses in Yang et al. (2019) were evaluated, one is naturally trained (ME-Net) and the other is adversarial trained (ME-Net-A). We called the defense in Hill et al. (2021) as EBM. The adversarial algorithm is PGD with a step number of 40.

Attack	ME-Net	ME-Net-A	EBM	Ours
BPDA+EOT	15.26	32.10	51.43	**52.09**

Table 4. Robust accuracy (percentage) of naturally trained target model and the time cost (millisecond) of the training process in one epoch on *CIFAR-10*. We compared the results with HGD and JATP on the same dataset.

Method	HGD	JATP	Ours
Accuracy	20.57	27.06	39.12
Time	291k	349k	527k

Our method maintained the defense effectiveness against masking-related adaptive attacks and presented better performances against BPDA and EOT adaptive attacks.

Moreover, we performed attacks against the target model on restored representations to verify the robustness of restored representations:

$$x_{t+1}^{\circ} = \Pi_{\tilde{x}+\mathcal{S}} \left(x_t^{\circ} + \tau \operatorname{sign} \left(\nabla_{\tilde{x}} \ell(h_\theta(x_t^{\circ}), y) \right) \right), \quad (9)$$

where x° denoted the perturbation on restored representation. The robust accuracy of JATP and our method against this attack (20 steps) was 26.79% and 42.16%. Our restored representations exhibited relatively better robustness.

Robust accuracy based natural model. For a naturally trained target model (i.e., a simple undefended target model), as shown in the result of Table. 4, our method still shows good robust accuracy.

However, our method requires more time cost which results are shown in Table. 4, this is due to the use of vision transformer and the introduction of an additional adversarial learning process in the robust representation restoration module. The additional time cost is a limitation of our method. Considering the significant gain in robust accuracy resulting from the proposed method, the cost is relatively worthwhile.

4.3. Ablation studies

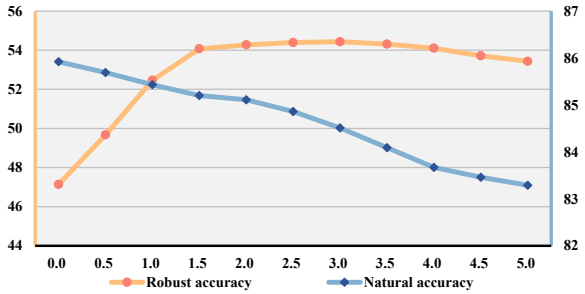
Information discard. We explored the influence of masks with three different spatial ratios (i.e., 30%, 50%, and 70%) on our denoising model. Corresponding to the three ratios, the natural and adversarial accuracies were 87.34%, 85.11%, 77.68% and 49.63%, 54.27%, 50.16%, respectively. Our model exhibited the optimal comprehensive performance when the spatial ratio was 50%, which was similar to the finding in Xie et al. (2022).

Robust representation restoration. We evaluated the influence of the robust representation restoration by varying the hyperparameter α in the range $[0.0, 5.0]$. As shown in Figure. 6(a), we found the module of robust representation restoration can significantly improve the adversarial accuracy compared to adversarial training alone (*i.e.*, $\alpha = 0$). We observed that our model performed relatively well in both adversarial accuracy and natural accuracy when $\alpha = 2.0$, and we thus set α to 2.0 by default.

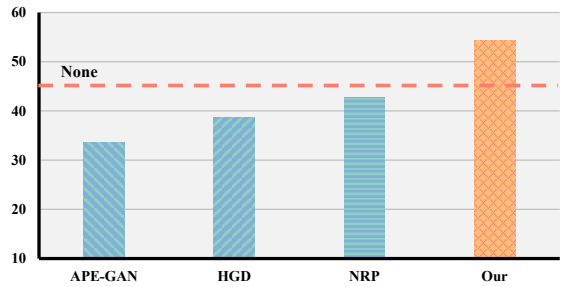
Excluding the influence of adversarial training on the proposed denoising model. To more clearly present the role of information discard and robust representation restoration, we performed adversarial training on the denoising models of baselines and compared our model with them. The results in Figure. 6(b) showed that our method still presented better performances.

The pre-training procedure. We canceled the pre-training procedure and introduce the mean square error between restored representations and natural samples into $\mathcal{L}_{den}$. The adversarial accuracy was reduced from 54.27% to 47.21% and more time consumption (about 1.1 times) was requested.

Depth of the denoising model. We found that the number of encoder and decoder layers had less impact on the natural accuracy. When the number of encoder layers was less than 8 or the number of decoder layers was less than 2, the robust accuracy decreased significantly. Considering that more layers required more hardware and time consumption, we chose an encoder of 8 layers and a decoder of 2 layers. The results were presented in Appendix. C.



(a) Robust and natural accuracy (percentage) at different α



(b) Robust accuracy of adversarially trained defenses

Figure 6. Ablation studies on *CIFAR-10*. The adversarial samples were crafted by PGD with a step number of 20. *None* denoted the robust accuracy of the target model trained by standard AT.

5. Conclusion

Denoising model-based defenses aim to remove adversarial noise from input adversarial samples. They have shown strong potential in defending against adversarial attacks. However, defensive denoising models are easily broken and even cause negative effects on adversarially trained target models in white-box adversarial environments. To handle this issue, we destroy the spatial characteristics of adversarial noise and focus on restoring the latent robust features from natural information retained in adversarial samples during the denoising process. We proposed a method based on information discard and robust representation restoration. These two modules were implemented by introducing complementary masks on input samples and by performing a perturbation-related game against the target model on restored representations, respectively. Empirical results demonstrated the effectiveness of the proposed method. The limitations of this work are the additional time consumption required due to the introduced game and the lack of evaluation on larger datasets. We will address these issues in future work, such as using fast adversarial training (Andriushchenko & Flammarion, 2020). Overall, our work aims to explore and improve the performance of defensive denoising models in worst-case scenarios to help counter complex potential threats in the real world.

Acknowledgements

The authors greatly appreciate all reviewers. NNW was supported in part by the National Key Research and Development Program of China under Grant 2018AAA0103202; in part by the National Natural Science Foundation of China under Grant U22A2096; in part by the Technology Innovation Leading Program of Shaanxi under Grant 2022QFY01-15; in part by Open Research Projects of the Zhejiang Laboratory under Grant 2021KG0AB01; in part by the Fundamental Research Funds for the Central Universities under Grant QTZX23042; in part by the Youth Innovation Team of Shaanxi University. XBG was supported by the National Natural Science Foundation of China under Grant 62036007. DCL was supported in part by the Fundamental Research Funds for the Central Universities under Grand XJS221502 and in part by the Natural Science Basis Research Plan in Shaanxi Province of China under Grant 2022JQ-696; DWZ and YKC were supported in part by the Fundamental Research Funds for the Central Universities and in part by the Innovation Fund of Xidian University under Grant YJSJ23012.

References

Andrew, G. and Gao, J. Scalable training of l 1-regularized log-linear models. In *Proceedings of the 24th international conference on Machine learning*, pp. 33–40, 2007.

Andriushchenko, M. and Flammarion, N. Understanding and improving fast adversarial training. *Advances in Neural Information Processing Systems*, 33:16048–16059, 2020.

Athalye, A. and Carlini, N. On the robustness of the cvpr 2018 white-box adversarial example defenses. *arXiv preprint arXiv:1804.03286*, 2018.

Athalye, A., Carlini, N., and Wagner, D. A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning*, 2018a.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In *International conference on machine learning*, pp. 284–293. PMLR, 2018b.

Aydin, A., Sen, D., Karli, B. T., Hanoglu, O., and Temizel, A. Imperceptible adversarial examples by spatial chroma-shift. In *Proceedings of the 1st International Workshop on Adversarial Learning for Multimedia*, pp. 8–14, 2021.

Bastani, V., Helfroush, M. S., and Kasiri, K. Image compression based on spatial redundancy removal and image inpainting. *Journal of Zhejiang University SCIENCE C*, 11(2):92–100, 2010.

Carlini, N. and Wagner, D. Magnet and "efficient defenses against adversarial attacks" are not robust to adversarial examples. *arXiv preprint arXiv:1711.08478*, 2017a.

Carlini, N. and Wagner, D. Towards evaluating the robustness of neural networks. In *2017 Ieee Symposium on Security and Privacy (sp)*, pp. 39–57. IEEE, 2017b.

Croce, F. and Hein, M. Minimally distorted adversarial examples with a fast adaptive boundary attack. In *International Conference on Machine Learning*, pp. 2196–2205. PMLR, 2020a.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, pp. 2206–2216. PMLR, 2020b.

Ding, G. W., Lui, K. Y. C., Jin, X., Wang, L., and Huang, R. On the sensitivity of adversarial robustness to input data distributions. In *ICLR (Poster)*, 2019.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Goodfellow, I. J., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Guo, C., Rana, M., Cissé, M., and van der Maaten, L. Countering adversarial images using input transformations. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *Conference on Computer Vision and Pattern Recognition*, pp. 770–778, 2016.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16000–16009, 2022.

Hill, M., Mitchell, J. C., and Zhu, S.-C. Stochastic security: Adversarial defense using long-run dynamics of energy-based models. In *International Conference on Learning Representations*, 2021.

Hu, Y., Yang, W., and Liu, J. Coarse-to-fine hyper-prior modeling for learned image compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 11013–11020, 2020.

Ilyas, A., Santurkar, S., Tsipras, D., Engstrom, L., Tran, B., and Madry, A. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

Jin, G., Shen, S., Zhang, D., Dai, F., and Zhang, Y. APE-GAN: adversarial perturbation elimination with GAN. In *International Conference on Acoustics, Speech and Signal Processing*, pp. 3842–3846, 2019.

Kim, J., Lee, B.-K., and Ro, Y. M. Distilling robust and non-robust features in adversarial examples by information bottleneck. *Advances in Neural Information Processing Systems*, 34:17148–17159, 2021.

Kos, J., Fischer, I., and Song, D. Adversarial examples for generative models. In *2018 ieee security and privacy workshops (spw)*, pp. 36–42. IEEE, 2018.

Krizhevsky, A., Hinton, G., et al. Learning multiple layers of features from tiny images. 2009.

Li, S., Xia, X., Zhang, H., Zhan, Y., Ge, S., and Liu, T. Estimating noise transition matrix with label correlations for noisy multi-label learning. In *NeurIPS*, 2022.

Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., and Zhu, J. Defense against adversarial attacks using high-level representation guided denoiser. In *Conference on Computer Vision and Pattern Recognition*, pp. 1778–1787, 2018.

Ma, X., Li, B., Wang, Y., Erfani, S. M., Wijewickrema, S. N. R., Schoenebeck, G., Song, D., Houle, M. E., and Bailey, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *International Conference on Learning Representations*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations*, 2018.

McInnes, L., Healy, J., and Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Naseer, M., Khan, S., Hayat, M., Khan, F. S., and Porikli, F. A self-supervised approach for adversarial robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 262–271, 2020.

Netzer, Y., Wang, T., Coates, A., Bissacco, A., Wu, B., and Ng, A. Y. Reading digits in natural images with unsupervised feature learning. 2011.

Qian, S. and Tolkova, I. Detection and properties of adversarial noise.

Qin, Y., Frosst, N., Sabour, S., Raffel, C., Cottrell, G., and Hinton, G. Detecting and diagnosing adversarial images with class-conditional capsule reconstructions. *arXiv preprint arXiv:1907.02957*, 2019.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Rony, J., Hafemann, L. G., Oliveira, L. S., Ayed, I. B., Sabourin, R., and Granger, E. Decoupling direction and norm for efficient gradient-based L2 adversarial attacks and defenses. In *Conference on Computer Vision and Pattern Recognition*, pp. 4322–4330, 2019.

Ru, B., Cobb, A., Blaas, A., and Gal, Y. Bayesopt adversarial attack. In *International Conference on Learning Representations*, 2019.

Sun, C., Chen, S., Cai, J., and Huang, X. Type i attack for generative models. In *2020 IEEE International Conference on Image Processing (ICIP)*, pp. 593–597. IEEE, 2020.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tramer, F., Carlini, N., Brendel, W., and Madry, A. On adaptive attacks to adversarial example defenses. *arXiv preprint arXiv:2002.08347*, 2020.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, G., Wei, X., and Yan, H. Improving adversarial transferability with spatial momentum. *arXiv preprint arXiv:2203.13479*, 2022.

Wang, X. and He, K. Enhancing the transferability of adversarial attacks through variance tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1924–1933, 2021.

Wang, Y., Deng, X., Pu, S., and Huang, Z. Residual convolutional ctc networks for automatic speech recognition. *arXiv preprint arXiv:1702.07793*, 2017.

Wang, Y., Zou, D., Yi, J., Bailey, J., Ma, X., and Gu, Q. Improving adversarial robustness requires revisiting misclassified examples. In *International Conference on Learning Representations*, 2019.

Wu, D., Xia, S.-T., and Wang, Y. Adversarial weight perturbation helps robust generalization. *Advances in Neural Information Processing Systems*, 33, 2020a.

Wu, K., Wang, A. H., and Yu, Y. Stronger and faster wasserstein adversarial attacks. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 10377–10387, 2020b.

Wu, S., Xia, X., Liu, T., Han, B., Gong, M., Wang, N., Liu, H., and Niu, G. Class2simi: A noise reduction perspective on learning with noisy labels. In *ICML*, pp. 11285–11295, 2021.

Xia, X., Liu, T., Wang, N., Han, B., Gong, C., Niu, G., and Sugiyama, M. Are anchor points really indispensable in label-noise learning? In *NeurIPS*, 2019.

Xia, X., Liu, T., Han, B., Gong, M., Yu, J., Niu, G., and Sugiyama, M. Sample selection with uncertainty of losses for learning with noisy labels. In *ICLR*, 2022.

Xie, Z., Zhang, Z., Cao, Y., Lin, Y., Bao, J., Yao, Z., Dai, Q., and Hu, H. Simmim: A simple framework for masked image modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9653–9663, 2022.

Xu, W., Evans, D., and Qi, Y. Feature squeezing: Detecting adversarial examples in deep neural networks. *arXiv preprint arXiv:1704.01155*, 2017.

Yang, Y., Zhang, G., Katabi, D., and Xu, Z. Me-net: Towards effective adversarial robustness with matrix estimation. *arXiv preprint arXiv:1905.11971*, 2019.

Zhang, H., Yu, Y., Jiao, J., Xing, E., El Ghaoui, L., and Jordan, M. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning*, pp. 7472–7482. PMLR, 2019.

Zhou, D., Wang, N., Gao, X., Han, B., Yu, J., Wang, X., and Liu, T. Improving white-box robustness of pre-processing defenses via joint adversarial training. *arXiv preprint arXiv:2106.05453*, 2021.

A. Proof-of-concept experiment in Section. 3.2.

To perform information discard, we randomly constructed masks with different spatial ratios. That is, we replace the masked regions in the adversarial noise or natural samples with zero-valued pixels. For adversarial noise crafted by PGD with a step number of 20 and a step size of 2 / 255, we added masked noise to the original natural samples and calculated the robust accuracy. For natural samples, we directly discarded pixels and calculated the natural accuracy. The used target model is adversarially trained by standard AT.

B. Examples of restored representations

We presented some representations restored by our method and a baseline (*e.g.*, HGD) in Figure. 7. Our method had less apparent anomalies compared with HGD and its adversarially trained vision. This indicated the higher reliability of our denoising model. In addition, we found that there were local differences in our restored representations from the original natural samples. The color textures in many patches were altered. This might be because these information contains more non-robust features and was therefore more likely to be perturbed by residual noise to mislead the target model. Our denoising model was guided to focus on latent robust features and might thus modify these information to restore robust representations. Of course, we noted that the representations restored by our method had raster-like textures, which we would further address in future work. However, this issue did not seem to significantly affect the classification of the restored representations by the target model. Objectives in these representations were usually classified normally in human vision. Quantitative results also verified the effectiveness of our method.

C. Depth of the denoising model

We explored the influence of the depth of the denoising model on adversarial accuracy. We found that the number of encoder and decoder layers in a suitable range had less impact on the natural accuracy. When the number of encoder layers was less than 8 or the number of decoder layers was less than 2, the robust accuracy decreased significantly. Considering that more layers required more hardware and time consumption, we chose an encoder of 8 layers and a decoder of 2 layers to construct the denoising model in this work. The results were presented in Table. 5.

Table 5. The influence of the depth of the denoising model on *CIFAR-10*. We compute the natural accuracy and the adversarial accuracy against adversarial samples crafted by PGD.

Encoder layer	Decoder layer	Natural accuracy	Adversarial accuracy
4	1	84.20	48.23
4	2	84.21	49.97
6	1	84.25	52.01
6	2	84.27	52.63
8	2	85.11	54.27
12	4	85.37	54.32

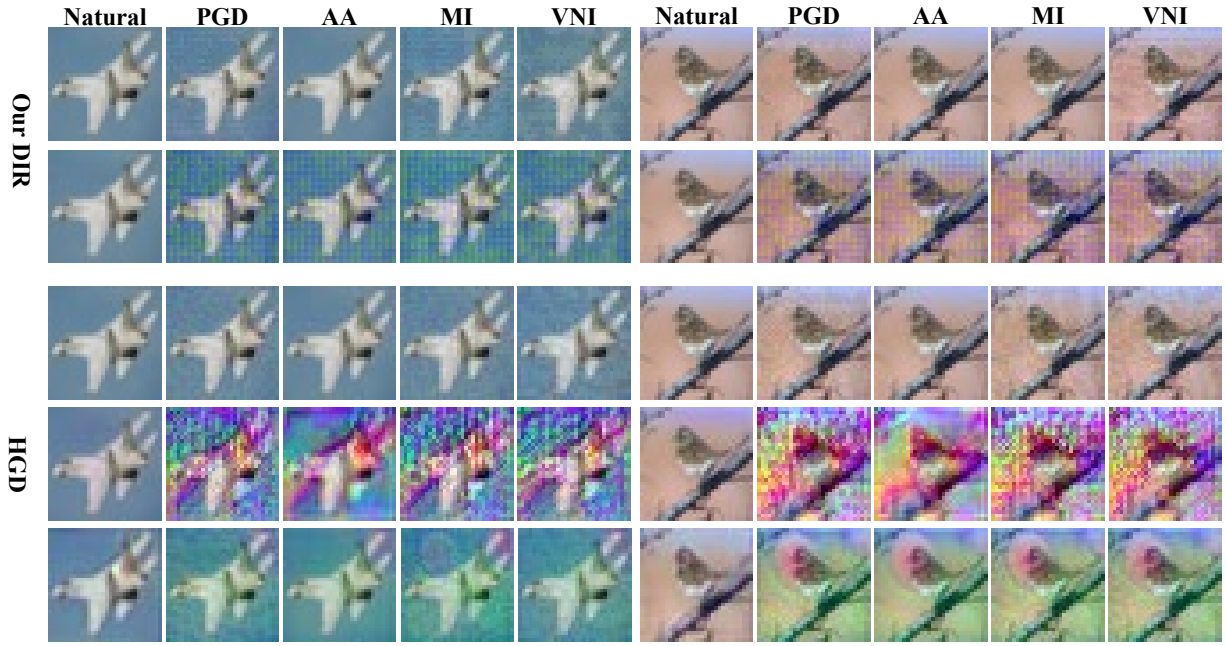


Figure 7. The examples of restored representations. The images from top to bottom are natural samples or white-box adversarial samples against our model, representations restored by our method, natural samples or white-box adversarial samples against a baseline (*e.g.*, HGD), representations restored by HGD and representations restored by the adversarially trained vision of HGD. It can be seen that our method effectively reduces the apparent anomalies in the restored representations.