DIFFender: Diffusion-Based Adversarial Defense against Patch Attacks

Caixin Kang[1], Yinpeng Dong[2,5], Zhengyi Wang[2,6], Shouwei Ruan[1], Yubo Chen[1], Hang Su[2,4⋆], and Xingxing Wei[1,3⋆]

[1] Institute of Artificial Intelligence, Beihang University, Beijing 100191, China
[2] Dept. of Comp. Sci. & Tech., Institute for AI, BNRist Center, THBI Lab, Tsinghua-Bosch Joint ML Center, Tsinghua University, Beijing, 100084, China
[3] Hangzhou Innovation Institute, Beihang University, Hangzhou 311228, China
[4] Zhongguancun Laboratory, Beijing, 100080, China [5] RealAI [6] ShengShu
{caixinkang,xxwei}@buaa.edu.cn, {dongyinpeng,suhangss}@tsinghua.edu.cn

Abstract. Adversarial attacks, particularly patch attacks, pose significant threats to the robustness and reliability of deep learning models. Developing reliable defenses against patch attacks is crucial for real-world applications. This paper introduces DIFFender, a novel defense framework that harnesses the capabilities of a text-guided diffusion model to combat patch attacks. Central to our approach is the discovery of the Adversarial Anomaly Perception (AAP) phenomenon, which empowers the diffusion model to detect and localize adversarial patches through the analysis of distributional discrepancies. DIFFender integrates dual tasks of patch localization and restoration within a single diffusion model framework, utilizing their close interaction to enhance defense efficacy. Moreover, DIFFender utilizes vision-language pre-training coupled with an efficient few-shot prompt-tuning algorithm, which streamlines the adaptation of the pre-trained diffusion model to defense tasks, thus eliminating the need for extensive retraining. Our comprehensive evaluation spans image classification and face recognition tasks, extending to real-world scenarios, where DIFFender shows good robustness against adversarial attacks. The versatility and generalizability of DIFFender are evident across a variety of settings, classifiers, and attack methodologies, marking an advancement in adversarial patch defense strategies. Our code is available at https://github.com/kkkcx/DIFFender.

Keywords: Adversarial Robustness · Patches Attacks · Diffusion Model

1 Introduction

Deep neural networks are vulnerable to adversarial examples [12, 36], in which imperceptible perturbations are intentionally added to natural examples, leading to incorrect predictions with high confidence of the model [25, 45]. Most adversarial attacks and defenses are devoted to studying the ℓ_p-norm threat models [3, 8, 12, 27], which assume that the adversarial perturbations are restricted

⋆ Corresponding authors.

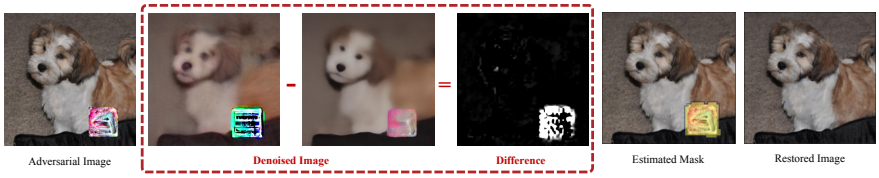

Fig. 1: The intriguing phenomenon of the diffusion model. A diffusion model is performed multiple times on the given adversarial image, and the differences between any two denoised images are pronounced within the adversarial patch regions, which can be leveraged to further pinpoint the location of adversarial patches.

by the ℓ_p norm to be imperceptible. However, the classic ℓ_p perturbations require modification of every pixel of the images, which is typically not practical in the physical world. On the other hand, adversarial patch attacks [2,20,22,39], which usually apply perturbations to a localized region of the objects, are more physically realizable. Adversarial patch attacks pose significant threats to real-world applications, such as face recognition [33,44], autonomous driving [7,19,21,53].

Although many adversarial defenses against patch attacks have been proposed in the past years, the defense performance is not satisfactory, which cannot meet the demands of the safety and reliability of real-world applications. Some methods employ adversarial training [31, 41, 42] and certified defenses [4, 13], which are only effective against specific attacks but generalize poorly to other forms of patch attacks in the real world [30]. Another category of patch defense is based on pre-processing techniques [14, 24, 29, 48], which usually destroy the patterns of adversarial patches by image completion or smoothing. However, these methods can hardly restore the images with high fidelity, leading to visual artifacts of the reconstructed images that impact recognition. They can also be evaded by stronger adaptive attacks due to gradient obfuscation [1].

Recently, diffusion models [17,34] have emerged as a powerful family of generative models, and have been successfully applied to improving adversarial robustness by purifying the input data [30,38,43]. Our initial intuition is to explore whether diffusion purification can defend against patch attacks. However, we find it fails to counter such attacks since it cannot remove the adversarial patches completely. In contrast, we discover the **Adversarial Anomaly Perception (AAP)** phenomenon, as shown in Fig. 1. The phenomenon suggests that we can calculate the difference between various denoised images to identify the region of adversarial patches. Subsequently, it facilitates targeted restoration of specific patch-affected areas. The reason behind the phenomenon may be that adversarial patches are often complexly crafted perturbations or contextually misplaced elements, significantly differing from the natural image distributions it is trained on. This phenomenon indicates progress in understanding how diffusion models can differently respond to adversarial patches and resolves the inherent trade-off between purifying the adversarial patches and preserving the image semantics.

Based on the AAP phenomenon, we propose **DIFFender**, a novel defense framework against adversarial patch attacks with pre-trained diffusion models.

DIFFender localizes the region of the adversarial patch by comparing the differences between various denoised images and then recovers the identified patch region in the image while preserving the integrity of the underlying content. Importantly, these two stages are carefully guided by a unified diffusion model, thus we can utilize the close interaction between them to improve the whole defense (i.e., an accurate localization will promote the following restoration, and a perfect restoration will help evaluate the performance of localization step in return). Specifically, we incorporate a text-guided diffusion model such that DIFFender can localize and recover the adversarial patches more accurately with textual prompts. Moreover, we design a few-shot prompt-tuning algorithm to facilitate simple and efficient tuning, enabling the pre-trained diffusion model to easily adapt to the adversarial defense task for improved robustness. The pipeline of DIFFender is illustrated in Fig. 2. In summary, our contributions are as follows:

- We uncover the intriguing Adversarial Anomaly Perception (AAP) phenomenon within the diffusion model, enabling it to leverage the distributional discrepancies between adversarial patches and natural images for accurate localization, thus overcoming the trade-off between purifying patches and preserving image semantics. This approach broadens the applicability of diffusion techniques, making it feasible to employ the diffusion model in countering adversarial patch attacks.
- Arising from the AAP phenomenon, we introduce DIFFender, an innovative diffusion-based defense framework. DIFFender employs one diffusion model throughout the entire process to both localize and restore adversarial patches, leveraging vision-language pre-training to implement efficient defense. To our knowledge, DIFFender stands as the first framework to leverage the diffusion model for comprehensive defense against patch attacks, marking a notable advancement in the field.
- Additionally, we develop an efficient prompt-tuning module and design three effective losses. The losses fine-tune the pre-trained diffusion model through the tuning process, enabling the model to co-optimize both localization and restoration modules, thereby achieving improved defense performance. This approach not only enhances the model's adaptability but also reduces the computational overhead associated with traditional retraining methods.
- We conduct extensive experiments on image classification, face recognition, and further in the physical world, demonstrating that DIFFender effectively reduces the attack success rate even under strong adaptive attacks. The results indicate that DIFFender can also generalize well to various scenarios, diverse classifiers, and multiple attack methods.

2 Related work

Adversarial attacks. Deep neural networks (DNNs) can be misled to produce erroneous outputs [9, 12, 36] by introducing perturbations to input examples. Most adversarial attacks [8, 12, 26–28] typically induce misclassification by

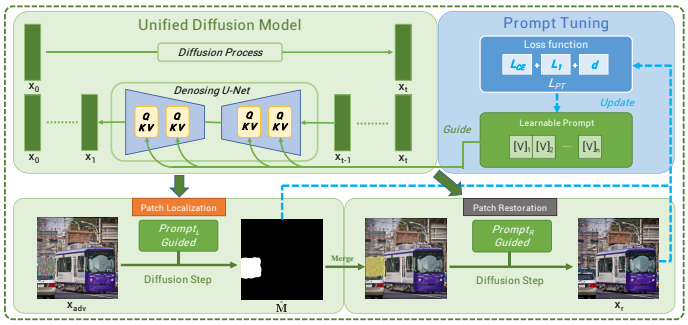

Fig. 2: Pipeline of DIFFender. DIFFender leverages a unified diffusion model to jointly guide the localization and restoration of adversarial patch attacks, and combines a prompt-tuning module to facilitate efficient tuning.

adding small perturbations to the pixels of input examples. However, while these methods can effectively generate adversarial examples in the digital world, they lack practicality in the real world. On the other hand, adversarial patch attacks [2, 20, 22, 39, 51] aim to deceive models by applying a pattern or a sticker to a localized region of the object, which are more realizable in the physical world.

Adversarial defenses. As attacks evolve, various defense methods have emerged. However, most existing defenses primarily focus on global perturbations with ℓ_p norm constraints, including former diffusion-based defenses [30, 38, 43], and defenses against patch attacks have not been extensively studied. Despite the effectiveness of adversarial training [31, 42, 50] and certified defenses [4, 13] against specific attacks, they have limited generalization to other patch attacks.

Therefore, most studies focus on pre-processing defenses. Digital Watermarking [14] utilizes saliency maps to detect adversarial regions and employs erosion operations to remove small holes. Local Gradient Smoothing [29] performs gradient smoothing on regions with high gradient amplitudes, taking into account the high-frequency noise introduced by patch attacks. Feature Normalization and Clipping [48] involves gradient clipping operations on images to reduce informative class evidence based on knowledge of the network structure. SAC [24] defends against patch attacks by detecting and removing patches. Jedi [37] utilizes entropy to obtain masks for patches. However, these methods can hardly reconstruct the original image and can be evaded by adaptive attacks [1]. In contrast, we introduce the AAP phenomenon and propose to leverage pre-trained diffusion models to localize and restore the adversarial patches. This method not only addresses the limitations of existing defenses but also opens new pathways for research in using diffusion models for patch defense.

3 Methodology

In this section, we first introduce the discovery of the AAP phenomenon within diffusion models. Following this understanding, we outline the whole framework of DIFFender and detail the improved techniques by prompt tuning.

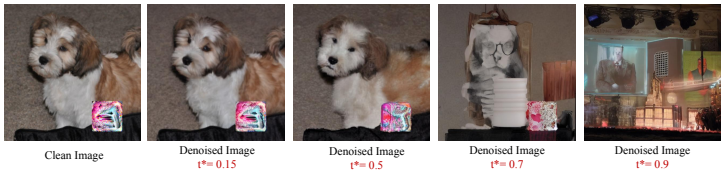

Fig. 3: Denoised results by diffusion model at different noise ratios. With small ratios ($t^* = 0.15/0.5$), the patch cannot be purified; conversely, the global structure becomes lost with large ratios ($t^* = 0.7/0.9$).

3.1 Discovery of the AAP Phenomenon

DiffPure [30] is a recent method that utilizes diffusion models to remove the imperceptible perturbations on the images by introducing Gaussian noise at a predetermined ratio t^* (ranging from 0 to 1) to adversarial images, followed by a denoising process via the reverse dynamics of diffusion models. Our study initially aimed to assess the applicability of DiffPure against patch attacks. The empirical investigations, illustrated in Fig. 3 demonstrate the inadequacy of DiffPure in countering patch attacks. This inefficacy stems from an inherent trade-off between purifying the adversarial perturbations (with a larger t^*) and preserving the image semantics (with a smaller t^*), making it impossible to find an appropriate noise ratio that can defend against adversarial patches.

In contrast, we find at a critical noise ratio of t^*, a distinct pattern emerged: while the adversarial patches exhibited resistance to denoising, they also struggled to be restored, resulting in variable outcomes. Meanwhile, the remainder of the image retained its semantic integrity unscathed. This suggests that we can calculate the difference between various denoised images to identify the region of adversarial patches. This observation, depicted through various examples in Fig. 4, leads to the **Adversarial Anomaly Perception (AAP)** phenomenon. The reason behind this phenomenon may be that adversarial patches are often intentionally crafted perturbations with a complexity far exceeding the noise present in real image datasets. Alternatively, it could be some meaningful sticker placed in an inappropriate location, signifying that the patch is out of context within the scenario. Diffusion models are trained to learn the probability distribution

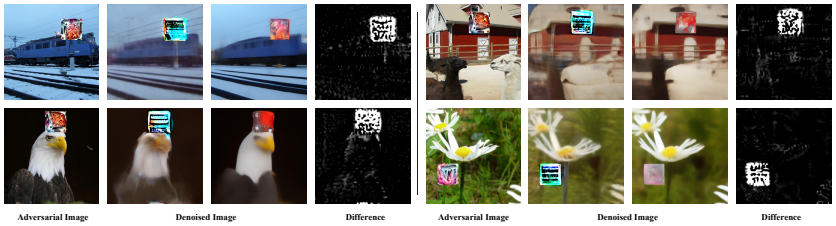

Fig. 4: In the analysis of ImageNet images, we find a pronounced difference specifically within regions affected by adversarial patches. This observation provides empirical evidence supporting the AAP phenomenon.

of real images, thus they struggle to fully adapt to the distribution of adversarial examples in its latent space, leading to the difference.

The discovery of AAP provides insight into understanding how diffusion models can differentially respond to adversarial patches, which empowers the diffusion model to detect and localize adversarial patches through the analysis of distributional discrepancies, and further facilitates targeted restoration of specific patch-affected areas. It resolves the inherent trade-off between purging adversarial patches and preserving image authenticity. Leveraging AAP, we propose a unified diffusion-based defense framework **DIFFender**, employing a single diffusion model that both locates and restores patch attacks.

3.2 DIFFender

Patch localization. DIFFender first performs accurate patch localization based on the above phenomenon of the diffusion model. Given the adversarial image $\mathbf{x}_{adv}$, we first add Gaussian noise to create a noisy image $\mathbf{x}_t$ with a certain noise ratio t^* (chosen as 0.5 in the experiments). Next, inspired by [5], we apply a text-guided diffusion model to obtain a denoised image $\mathbf{x}_p$ from $\mathbf{x}_t$ with a textual prompt $prompt_L$, and $\mathbf{x}_e$ with empty text. We can estimate the mask region $\hat{\mathbf{M}}$ by taking the difference between the denoised images $\mathbf{x}_p$ and $\mathbf{x}_e$. However, the diffusion model incurs a significant time cost due to the time steps T required. To address this issue, we directly predict the image $\mathbf{x}_0$ from the noisy image $\mathbf{x}_t$ with only one step, which saves T times the processing time.

Although the one-step predicted results often exhibit discrepancies and increased blurriness compared to the original one, the differences between one-step predictions still align with the AAP phenomenon. In practice, we perform one-step denoising twice, obtaining two results: $\mathbf{x}_a$, the one guided by $prompt_L$, and $\mathbf{x}_b$, the one guided by empty text to calculate the difference and binarize it, as:

$$\hat{\mathbf{M}} = \text{Binarize}\left(\frac{1}{m}\sum_{i=0}^{m}(\mathbf{x}_a^i - \mathbf{x}_b^i)\right), \tag{1}$$

where we calculate the difference for m times (set to 3 in the experiments) to enhance stability and eliminate randomness. The $prompt_L$ can be hand-designed (e.g., "adversarial") or automatically tuned as shown in Sec. 3.3.

Mask refining. As shown in Fig. 5, directly obtained averaged difference may sometimes result in minor inaccuracies. Therefore, we binarize the difference to gain the initial mask and then refine it by sequentially applying Gaussian smoothing and dilation operations, leading to a precise estimation of the patch region. The processed mask edges may slightly extend beyond the patch area, which helps maintain consistency in the patch restoration, thereby enhancing the overall performance of the defense pipeline.

Patch restoration. After locating the patch region, DIFFender then restores it to eliminate the adversarial effects, while also considering preserving the overall coherence and quality of the image. In particular, we combine the estimated mask $\hat{\mathbf{M}}$ and $\mathbf{x}_{adv}$ as inputs to the text-guided diffusion model with prompt

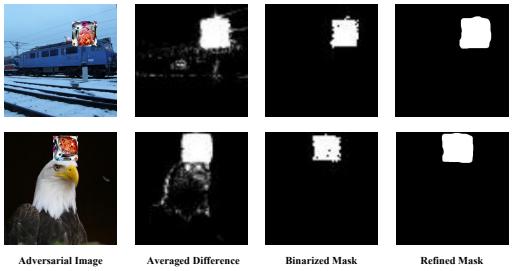

Fig. 5: To gain the final refined mask, the estimated differences are binarized, applied Gaussian smoothing and dilation operations.

$prompt_R$ to obtain a restored image $\mathbf{x}_r$. We follow the inpainting pipeline in Stable Diffusion [32] to process the mask, where a UNet is used with an additional five input channels to incorporate the estimated mask $\hat{\mathbf{M}}$. Similarly, $prompt_R$ can be manually set (e.g., "clean") or automatically tuned.

Unified defense model. The aforementioned two stages have been meticulously integrated into one unified diffusion model (e.g., Stable Diffusion), driven by the critical AAP phenomenon. This intentional fusion allows us to harness the tight interplay between these stages, thereby enhancing the defense mechanism. As a direct consequence of our insights, we also introduced the prompt-tuning module, which encompasses the joint optimization of the entire pipeline.

3.3 Prompt Tuning

Following the aforementioned pipeline, leveraging vision-language pre-training, DIFFender is capable of efficiently performing zero-shot localization and restoration. While it is accurate in locating and restoring in most cases, subtle discrepancies may occur in certain challenging situations. Given that vision-language pre-training takes advantage of large-capacity text encoders to explore a vast semantic space [52], to facilitate the effective adaptation of learned representations into adversarial defense tasks, we introduce the algorithm of prompt tuning.

Learnable prompts. First, the textual vocabulary is replaced by learnable vectors. Thus, $prompt_L$ and $prompt_R$ can be transformed into vectors as follows:

$$\begin{aligned} prompt_L &= [V_L]_1[V_L]_2 \dots [V_L]_n; \\ prompt_R &= [V_R]_1[V_R]_2 \dots [V_R]_n, \end{aligned} \tag{2}$$

where each $[V_L]_i$ or $[V_R]_i$ ($i \in \{1, \dots, n\}$) is a vector of the same dimensionality as word embeddings. n is a hyperparameter that specifies the number of context tokens, we set $n = 16$ by default. The text content used to initialize $prompt_L$ and $prompt_R$ can be manually provided or randomly initialized.

Tuning process. After obtaining the learnable vectors, we design three loss functions for prompt tuning, which jointly optimize the vectors to capture the characteristics of the adversarial regions and improve the defense performance.

First, to accurately identify the adversarial regions, we employ cross-entropy loss by comparing estimated mask $\hat{\mathbf{M}}$ with ground-truth mask $\mathbf{M}$.

$$L_{CE}(\mathbf{M}, \hat{\mathbf{M}}) = -\sum_{i=1}^{d} \mathbf{M}_i \log(\hat{\mathbf{M}}_i), \tag{3}$$

where i indicates the i-th element of the mask. Next, in the patch restoration module, our objective is to restore the mask region while eliminating the adversarial effect of the image. To ensure effective defense, we calculate the ℓ_1 distance between restored image $\mathbf{x}_r$ and clean image $\mathbf{x}$ as:

$$L_1(\mathbf{x}_r, \mathbf{x}) = |\mathbf{x}_r - \mathbf{x}|\,. \tag{4}$$

Lastly, to verify that the adversarial effects have been eliminated, we draw inspiration from [23] and [49] to make the high-level feature representations of the downstream classifiers between the restored image $\mathbf{x}_r$ and the clean image $\mathbf{x}$ close to each other. Specifically, we compute the ℓ_2 distance between their feature representations weighted by a layer-wise hyperparameter as

$$d\left(\mathbf{x}_r, \mathbf{x}\right) = \sum_{l} \frac{1}{H_l W_l} \sum_{h,w} \left\| w_l \odot \left(\hat{y}^l_{rhw} - \hat{y}^l_{chw}\right)\right\|_2^2, \tag{5}$$

where l denotes a certain layer in the network, $\hat{y}^l_r, \hat{y}^l_c \in \mathcal{R}^{H_l \times W_l \times C_l}$ are the unit-normalize results in the channel dimension, and vector $w^l \in \mathcal{R}^{C_l}$ is used for scaling activation channels.

By summing three losses, the overall prompt tuning loss L_{PT} is:

$$L_{PT} = L_{CE}(\mathbf{M}, \hat{\mathbf{M}}) + L_1(\mathbf{x}_r, \mathbf{x}) + d\left(\mathbf{x}_r, \mathbf{x}\right). \tag{6}$$

We perform gradient descent to minimize L_{PT} w.r.t. $prompt_L$ and $prompt_R$ for prompt tuning. The design of continuous representations also enables thorough exploration in the embedding space.

Few-shot learning. During prompt tuning, we utilize a limited number of images for few-shot tuning. Specifically, DIFFender is tuned on a limited set of attacked images (8-shot in the experiments) from a single attack, but can learn optimal prompts that generalize well to other scenarios and attacks, which makes the tuning module effective and straightforward.

4 Experiments

4.1 Experimental settings.

Datasets and Baselines. We consider ImageNet [6] for evaluation against eight state-of-the-art defense methods: Image smoothing-based defenses, including JPEG [11] and Spatial Smoothing [46], image completion-based defenses such as DW [14], LGS [29] and SAC [24], feature-level suppression defense FNC [48],

Table 1: Accuracy (%) against attacks on ImageNet by Inception-v3 and Swin-S.

Models	Inception-v3					Swin-S				
		Adaptive		Non-adaptive			Adaptive		Non-adaptive	
Defense \ Attack	Clean	AdvP	LaVAN	GDPA	RHDE	Clean	AdvP	LaVAN	GDPA	RHDE
Undefended	100.0	0.0	8.2	64.8	39.8	100.0	1.6	3.5	78.1	51.6
JPEG [11]	48.8	0.4	15.2	64.8	13.3	85.2	0.8	5.9	77.0	38.7
SS [46]	72.7	1.2	14.8	57.8	16.4	86.3	2.3	5.5	68.8	34.8
DW [14]	87.1	1.2	9.4	62.5	28.5	88.3	0.0	5.1	77.3	66.0
LGS [29]	87.9	55.5	50.4	67.2	49.6	89.8	65.6	59.8	82.0	69.1
FNC [48]	91.0	61.3	64.8	66.4	46.5	91.8	6.3	7.4	77.0	63.7
DiffPure [30]	65.2	10.5	15.2	67.6	44.9	74.6	18.4	26.2	77.7	62.3
SAC [24]	**92.8**	84.2	65.2	68.0	41.0	93.6	92.8	84.6	79.3	54.9
Jedi [37]	92.2	67.6	20.3	74.6	47.7	93.4	89.1	12.1	78.1	67.6
DIFFender	91.4	**88.3**	**71.9**	**75.0**	**53.5**	**93.8**	**94.5**	**85.9**	**82.4**	**70.3**

alongside Jedi [37], a defense based on entropy. Additionally, we assess the diffusion purification defense, DiffPure [30]. For classifiers, we consider two advanced classifiers: CNN-based Inception-v3 [35] and Transformer-based Swin-S [35].

Adversarial attacks. We employ AdvP [2] and LaVAN [20], which randomly select positions and optimize patches. GDPA [22], which optimizes the patch's position and content to execute attacks, and natural-looking attack RHDE [39], which utilizes realistic stickers and searches for their optimal positions to launch attacks. To implement adaptive attacks, we approximate gradients using the BPDA [1] to conduct BPDA+AdvP and BPDA+LaVAN attacks, which implies that the defense methods are white-box against the attacks. The number of iterations for the attacks is set to 100 with patch size 5% of the input image. For adapting the attack on DIFFender, we use an additional Straight-Through Estimator (STE) [47] during backpropagation through thresholding operations.

Evaluation metrics. We evaluate the performance of defenses under standard accuracy and robust accuracy. Due to the computational cost of adaptive attacks, unless otherwise specified, we assess the robust accuracy on a subset of 512 sampled images from the test set. To facilitate the observation, we ensure that the selected subset consists of images correctly classified.

4.2 Evaluation on ImageNet

Experimental results. Tab. 1 presents the experimental results, where the highest accuracy is highlighted in bold. Based on these results, we draw the following conclusions:

(1) DIFFender outperforms in defense effectiveness. Under adaptive attacks utilizing gradients, such as the BPDA+AdvP and BPDA+LaVAN, DIFFender exhibits exceptional performance, even only involving an 8-shot process. Other attacks like GDPA may not achieve strong attack effectiveness, but DIFFender still attains the highest robust accuracy. This can be attributed to that DIFFender is built upon the unified diffusion framework. Empowered by the AAP

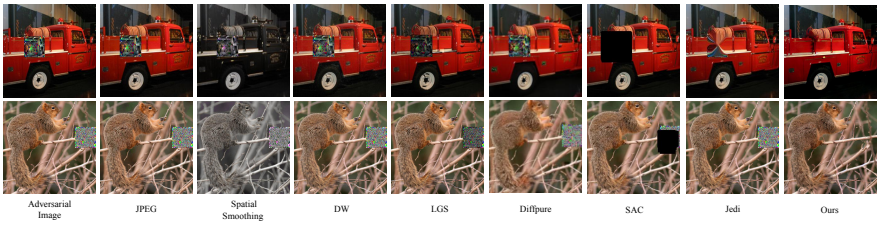

Fig. 6: Visualization with examples from ImageNet. The restored images of DIFFender exhibit no residual signs of the adversarial patch, and the restored details are remarkable (e.g., the recovery of tree branches in the second column of images).

phenomenon, the diffusion model can effectively locate and remove adversarial areas while ensuring a high-quality and diverse restoration that closely follows the underlying distribution of clean data. Additionally, the inherent stochasticity in the diffusion model allows for robust stochastic defense mechanisms [16], which makes it a well-suited "defender" for adaptive attacks.

(2) Image processing defense methods, such as JPEG, SS, and DW, experience a significant decrease in robust accuracy under adaptive attacks. This can be attributed to the algorithms' gradients can be easily obtained. Other methods, such as LGS, FNC, SAC and Jedi, consider the robustness against adaptive attacks. For instance, FNC achieves respectable robust accuracy on Inception-v3. However, its defense effectiveness diminishes when applied to the Swin-S. This is because the feature norm clipping layer proposed is specifically designed for handling CNN feature maps, while DIFFender exhibits excellent generalization capabilities that can extend to different classifiers.

(3) DIFFender demonstrates generalizability to unseen attacks. In the experiments, DIFFender only undergoes 8-shot prompt tuning specifically for the AdvP method yet achieves promising results under other attacks as well. For Jedi, it has strong robustness against several attack methods, such as AdvP, but its robust accuracy has significantly decreased under other methods, like LaVAN. This might be because the autoencoder used by Jedi is trained under a specific style and cannot generalize well.

(4) For the naturalistic attack RHDE, it poses a lesser threat to classifiers compared to adaptive meaningless attacks. However, it introduces a more significant challenge to defense methods, likely due to its utilization of irregular, more naturally-appearing patches. Nevertheless, DIFFender still achieves the best defense results without prior exposure to RHDE patches. Moreover, DIFFender is adaptable; with the prompt tuning module, a few-shot tuning can be employed to enhance performance specifically against naturalistic patch attacks.

(5) When defending against global perturbations with ℓ_p-norm constraints, DiffPure achieves excellent results. However, it performs poorly when facing patch attacks. Specifically, in Tab. 1, when tested against AdvP and LaVAN, the Inception-v3 model purified by DiffPure maintains robust accuracy rates of 10.5% and 15.2%, respectively. This aligns with our observations in Sec. 3.1.

Table 2: Ablation study for different loss functions of DIFFender.

			Inception-v3				
L_{CE}	L_1	d	Clean	AdvP	LaVAN	GDPA	RHDE
	✓	✓	**91.8**	76.2	66.0	72.3	49.2
✓		✓	88.3	87.1	69.5	73.8	52.7
✓	✓		90.2	87.1	69.1	73.0	52.0
✓	✓	✓	91.4	**88.3**	**71.9**	**75.0**	**53.5**

Table 3: Accuracy against attacks of varying patch sizes by Inception-v3.

Size	0.5%	1.0%	5.0%	10.0%	15.0%
Undefended	64.3	50.8	0.0	0.0	0.0
SAC	81.8	83.8	84.2	60.9	34.8
Jedi	61.7	56.4	67.6	42.2	33.8
DIFFender	**86.1**	**87.3**	**88.3**	**70.5**	**56.6**

Visualization. Fig. 6 presents the defense results of the defense methods against patch attacks. Since FNC suppresses the feature maps during the inference stage, it is not shown in Fig. 6. Other methods, such as JPEG and DW, only exhibit minor changes in the reconstructed images and fail to defend against adaptive attacks. After Spatial Smoothing defense, the images show color distortion and are still vulnerable to attacks. In the case of the LGS method, the patch area is visibly suppressed, which improves the robust accuracy to some extent, but the patch is not completely eliminated. For Jedi and SAC, their localization algorithm fails under certain scenarios, as the second line in Fig. 6. And the restored results of Jedi cannot achieve complete recovery. On the other hand, the restored images of DIFFender no longer show any traces of the patch, and the restored details are remarkable.

4.3 Ablation studies and additional results

Impact of loss functions. To evaluate the effectiveness of different losses, we conduct separate tuning experiments by removing loss functions L_{CE}, L_1, and d separately, as presented in Tab. 2, where we observe that the robust accuracy significantly decreases when optimizing only the Restoration module without optimizing L_{CE} due to the performance loss in the localization, although it led to an improvement in clean accuracy. On the other hand, removing L_1 results in a noticeable decrease in clean accuracy, as images cannot be well restored. Eliminating either the d or L_1 causes a slight drop in robust accuracy. Finally, DIFFender, which incorporates all three loss functions, achieves the highest robust accuracy, demonstrating the importance of joint optimization and close interaction between the two modules for the overall performance of DIFFender.
Impact of patch size. We conducted experiments under patch attacks of varying sizes, using patches generated by AdvP ranging from 0.5% to 15% in size, and compared them with the state-of-the-art SAC and Jedi. As shown in Tab. 3, DIFFender exhibits better generalization to patches of various sizes, benefitting from vision-language pre-training, whereas Jedi and SAC are more sensitive to patch size. Notably, DIFFender was only prompt-tuned for patches of 5.0% size.
Impact of restoration module. To verify the necessity of restoration, we remove the patch restoration and set the value in the $\hat{\mathbf{M}}$ region to zero. Experimental results in Tab. 4 show that the inclusion of patch restoration ensures better DIFFender performance. This is because patches may occasionally obscure crucial areas of an image, resulting in a loss of semantic information. The

Table 4: Ablation study for restoration modules in DIFFender. "NR" denotes "No Restoration Process".

Defense	Inception-v3					Swin-S				
	Clean	AdvP	LaVAN	GDPA	RHDE	Clean	AdvP	LaVAN	GDPA	RHDE
DIFFender (NR)	86.3	84.0	66.8	69.5	48.0	88.7	92.2	81.8	78.9	69.1
DIFFender	**91.4**	**88.3**	**71.9**	**75.0**	**53.5**	**93.8**	**94.5**	**85.9**	**82.4**	**70.3**

Table 5: Ablation study for different prompt forms. "EP" and "MP" represent "Empty Prompt" and "Manual Prompt".

Defense	Inception-v3					Swin-S				
	Clean	AdvP	LaVAN	GDPA	RHDE	Clean	AdvP	LaVAN	GDPA	RHDE
DIFFender (EP)	89.1	76.4	66.8	71.1	47.0	93.2	89.8	81.4	79.3	65.7
DIFFender (MP)	87.3	77.9	68.2	70.3	47.8	92.2	91.2	82.4	77.0	67.6
DIFFender	**91.4**	**88.3**	**71.9**	**75.0**	**53.5**	**93.8**	**94.5**	**85.9**	**82.4**	**70.3**

Table 6: Transferability of DIFFender on ResNet50 and ViT-B-16 for ImageNet.

Defense	ResNet-50					ViT-B-16				
	Clean	AdvP	LaVAN	GDPA	RHDE	Clean	AdvP	LaVAN	GDPA	RHDE
Undefended	100.0	0.0	14.8	73.8	37.1	100.0	1.2	2.0	76.2	52.0
Jedi	82.8	46.9	8.6	70.7	45.5	89.8	83.8	14.8	76.8	59.8
DIFFender	**83.6**	**83.2**	**55.9**	**76.2**	**53.5**	**91.0**	**88.3**	**85.2**	**78.9**	**68.0**

restoration step can address this issue by recovering lost semantics, aiding classifiers in overcoming challenging scenarios. Furthermore, longer diffusion steps introduce more randomness, which preserves accuracy against adaptive attacks. Consequently, we conclude that the patch restoration is indeed necessary.

Impact of Prompt Tuning. In Tab. 5, DIFFender with prompt-tuning is compared with the "Empty prompt" and "Manual prompt" versions of DIFFender. For DIFFender with manual prompts, we set $prompt_L$ = "adversarial" and $prompt_R$ = "clean". The prompt-tuned DIFFender shows a notable improvement in robust accuracy compared to the other two zero-shot versions, despite exposure to only a few attacked images, underscoring the effectiveness of prompt-tuning.

Cross-model transferability. We apply only 8-shot prompt-tuning exclusively on Inception-v3. Subsequently, the transferability of DIFFender is tested on diverse classifiers, including the CNN-based ResNet50 [15] and Transformer-based ViT-B-16 [10]. The results are detailed in Tab. 6, underscore DIFFender's ability to maintain robust accuracy across novel classifiers, demonstrating its potent generalization capacity.

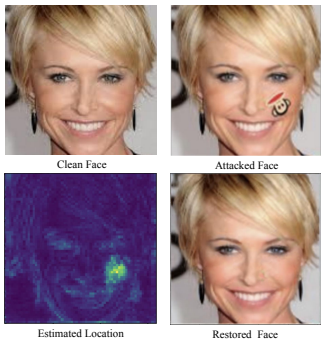

Fig. 7: Visualization with examples from LFW attacked naturally by RHDE, localized and restored by DIFFender.

Table 7: Accuracy against patch attacks on LFW by FaceNet.

Defense	FaceNet		
	Clean	GDPA	RHDE
Undefended	100.0	56.3	42.8
JPEG [11]	44.1	16.8	17.8
SS [46]	19.9	8.2	3.5
DW [14]	37.1	15.2	7.2
LGS [29]	60.9	71.9	53.5
FNC [48]	100.0	39.8	39.3
SAC [24]	100.0	77.3	43.2
Jedi [37]	100.0	74.2	43.9
DIFFender (EP)	100.0	79.3	57.2
DIFFender (MP)	100.0	77.0	57.2
DIFFender	**100.0**	**81.1**	**60.7**

4.4 Extension in Face Recognition.

Experimental settings. Facial expressions in human faces introduce a rich diversity, together with external factors such as lighting conditions and viewing angles, making face recognition a challenging task. We conducted experiments on the LFW [18], and employed two patch attacks: RHDE [39] and GDPA [22].
Experimental results. The results on the LFW dataset are presented in Tab. 7. DIFFender achieves the highest robust accuracy under both the GDPA and RHDE while ensuring clean accuracy. It is important to note that DIFFender is not tuned specifically for facial recognition. This further demonstrates the generalizability of DIFFender across different scenarios and attack methods. In contrast, JPEG, SS, and the FNC method obtained low robust accuracies. This is because in the specific context of facial recognition, the classifier focuses more on crucial local features, and preprocessing the entire image can disrupt these important features. Fig. 7 illustrates the results of DIFFender against face attacks. It can be observed that DIFFender accurately identifies the location of the natural patch and achieves excellent restoration.

4.5 Extension in Physical World.

We additionally conduct further experiments in the physical world, where we select 10 common object categories from ImageNet and perform two types of patch attacks (natural and meaningless) [40]. Our approach involves generating digital-world attack results first, then placing stickers on real-world objects in the same positions. We test DIFFender under various conditions, including different angles (rotations) and distances. Qualitative results are depicted in Fig. 8, while quantitative results are presented in Tab. 8, where each configuration is based on 256 frames successfully classified images from the 10 objects selected. Based on the results, we see that DIFFender demonstrates substantial defensive capabilities across various physical alterations, maintaining its efficacy in real-world scenarios.

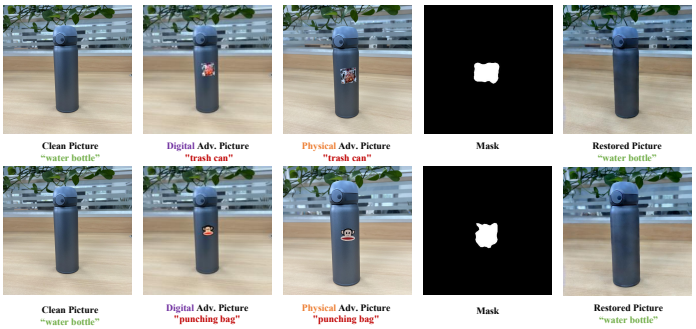

Fig. 8: Physical world defense demonstrations of DIFFender against meaningless and natural patch attacks. The mask edges may extend slightly beyond the patch region, aid in restoring the patch, and help maintain consistency in the restored image.

Table 8: Quantitative result of meaningless physical attacks on the Inception-v3 at different angles and distances.

	0°	yaw ±15°	yaw ±30°	pitch ±15°	distance
Undefended	28.9	34.8	41.8	36.7	35.9
Jedi	61.7	57.8	66.4	63.3	62.1
DIFFender	**80.9**	**76.6**	**77.7**	**75.4**	**73.8**

5 Discussion and Conclusion

We propose **DIFFender**, a novel defense framework harnessing a pre-trained unified diffusion model for dual roles in the localization and restoration of patch attacks, empowered by the discovery of the Adversarial Anomaly Perception (AAP) phenomenon. Additionally, we design a few-shot prompt-tuning algorithm to facilitate simple and efficient tuning, thus eliminating the need for extensive retraining. To validate the robust performance of DIFFender, we conduct experiments on image classification, face recognition, and further the physical world scenarios. Our findings demonstrate that DIFFender exhibits exceptional robustness even under adaptive attacks and extends the generalization capability of pre-trained large models to various scenarios, diverse classifiers, and multiple attack methods, requiring only a few-shot prompt-tuning. We prove that DIFFender effectively reduces the success rate of patch attacks while producing realistic restored images, promising a wide spectrum of diffusion model applications and inspiring future explorations in the domain.

There are several avenues for further exploration. While we have designed acceleration techniques to expedite diffusion-based methods, further acceleration can be achieved by adopting advanced acceleration sampling methods. Another potential direction for exploration is to extend the DIFFender framework to other modalities of adversarial attack defenses, such as video data.

Acknowledgement

This work was supported by the Project of the National Natural Science Foundation of China (No. 62076018, 92370124, 62276149, 92248303), the Fundamental Research Funds for the Central Universities, and Tsinghua-Alibaba Joint Research Program. Y. Dong was also supported by the China National Postdoctoral Program for Innovative Talents and CCF-BaiChuan-Ebtech Foundation Model Fund.

References

1. Athalye, A., Carlini, N., Wagner, D.: Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In: International conference on machine learning. pp. 274–283. PMLR (2018)
2. Brown, T.B., Mané, D., Roy, A., Abadi, M., Gilmer, J.: Adversarial patch. arXiv preprint arXiv:1712.09665 (2017)
3. Carlini, N., Wagner, D.: Towards evaluating the robustness of neural networks. In: 2017 ieee symposium on security and privacy (sp). pp. 39–57. Ieee (2017)
4. Chiang, P.y., Ni, R., Abdelkader, A., Zhu, C., Studer, C., Goldstein, T.: Certified defenses for adversarial patches. arXiv preprint arXiv:2003.06693 (2020)
5. Couairon, G., Verbeek, J., Schwenk, H., Cord, M.: Diffedit: Diffusion-based semantic image editing with mask guidance. arXiv preprint arXiv:2210.11427 (2022)
6. Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. pp. 248–255. Ieee (2009)
7. Dong, Y., Kang, C., Zhang, J., Zhu, Z., Wang, Y., Yang, X., Su, H., Wei, X., Zhu, J.: Benchmarking robustness of 3d object detection to common corruptions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1022–1032 (2023)
8. Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., Li, J.: Boosting adversarial attacks with momentum. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 9185–9193 (2018)
9. Dong, Y., Ruan, S., Su, H., Kang, C., Wei, X., Zhu, J.: Viewfool: Evaluating the robustness of visual recognition to adversarial viewpoints. arXiv preprint arXiv:2210.03895 (2022)
10. 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)
11. Dziugaite, G.K., Ghahramani, Z., Roy, D.M.: A study of the effect of jpg compression on adversarial images. arXiv preprint arXiv:1608.00853 (2016)
12. Goodfellow, I.J., Shlens, J., Szegedy, C.: Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572 (2014)
13. Gowal, S., Dvijotham, K.D., Stanforth, R., Bunel, R., Qin, C., Uesato, J., Arandjelovic, R., Mann, T., Kohli, P.: Scalable verified training for provably robust image classification. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 4842–4851 (2019)

14. Hayes, J.: On visible adversarial perturbations & digital watermarking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. pp. 1597–1604 (2018)
15. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
16. He, Z., Rakin, A.S., Fan, D.: Parametric noise injection: Trainable randomness to improve deep neural network robustness against adversarial attack. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 588–597 (2019)
17. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems **33**, 6840–6851 (2020)
18. Huang, G.B., Mattar, M., Berg, T., Learned-Miller, E.: Labeled faces in the wild: A database forstudying face recognition in unconstrained environments. In: Workshop on faces in'Real-Life'Images: detection, alignment, and recognition (2008)
19. Jing, P., Tang, Q., Du, Y., Xue, L., Luo, X., Wang, T., Nie, S., Wu, S.: Too good to be safe: Tricking lane detection in autonomous driving with crafted perturbations. In: Proceedings of USENIX Security Symposium (2021)
20. Karmon, D., Zoran, D., Goldberg, Y.: Lavan: Localized and visible adversarial noise. In: International Conference on Machine Learning. pp. 2507–2515. PMLR (2018)
21. Kong, L., Xie, S., Hu, H., Niu, Y., Ooi, W.T., Cottereau, B.R., Ng, L.X., Ma, Y., Zhang, W., Pan, L., et al.: The robodrive challenge: Drive anytime anywhere in any condition. arXiv preprint arXiv:2405.08816 (2024)
22. Li, X., Ji, S.: Generative dynamic patch attack. BMVC (2021)
23. Liao, F., Liang, M., Dong, Y., Pang, T., Hu, X., Zhu, J.: Defense against adversarial attacks using high-level representation guided denoiser. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 1778–1787 (2018)
24. Liu, J., Levine, A., Lau, C.P., Chellappa, R., Feizi, S.: Segment and complete: Defending object detectors against adversarial patch attacks with robust patch detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 14973–14982 (2022)
25. Ma, K., Xu, Q., Zeng, J., Cao, X., Huang, Q.: Poisoning attack against estimating from pairwise comparisons. IEEE Transactions on Pattern Analysis and Machine Intelligence **44**(10), 6393–6408 (2021)
26. Ma, K., Xu, Q., Zeng, J., Li, G., Cao, X., Huang, Q.: A tale of hodgerank and spectral method: Target attack against rank aggregation is the fixed point of adversarial game. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(4), 4090–4108 (2022)
27. Madry, A., Makelov, A., Schmidt, L., Tsipras, D., Vladu, A.: Towards deep learning models resistant to adversarial attacks. arXiv preprint arXiv:1706.06083 (2017)
28. Moosavi-Dezfooli, S.M., Fawzi, A., Frossard, P.: Deepfool: a simple and accurate method to fool deep neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2574–2582 (2016)
29. Naseer, M., Khan, S., Porikli, F.: Local gradients smoothing: Defense against localized adversarial attacks. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). pp. 1300–1307. IEEE (2019)
30. Nie, W., Guo, B., Huang, Y., Xiao, C., Vahdat, A., Anandkumar, A.: Diffusion models for adversarial purification. arXiv preprint arXiv:2205.07460 (2022)

31. Rao, S., Stutz, D., Schiele, B.: Adversarial training against location-optimized adversarial patches. In: Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part V 16. pp. 429–448. Springer (2020)
32. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
33. Sharif, M., Bhagavatula, S., Bauer, L., Reiter, M.K.: Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition. In: Proceedings of the 2016 acm sigsac conference on computer and communications security. pp. 1528–1540 (2016)
34. Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., Ganguli, S.: Deep unsupervised learning using nonequilibrium thermodynamics. In: International Conference on Machine Learning. pp. 2256–2265 (2015)
35. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J., Wojna, Z.: Rethinking the inception architecture for computer vision. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2818–2826 (2016)
36. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
37. Tarchoun, B., Ben Khalifa, A., Mahjoub, M.A., Abu-Ghazaleh, N., Alouani, I.: Jedi: Entropy-based localization and removal of adversarial patches. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4087–4095 (2023)
38. Wang, J., Lyu, Z., Lin, D., Dai, B., Fu, H.: Guided diffusion model for adversarial purification. arXiv preprint arXiv:2205.14969 (2022)
39. Wei, X., Guo, Y., Yu, J.: Adversarial sticker: A stealthy attack method in the physical world. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(3), 2711–2725 (2023)
40. Wei, X., Pu, B., Lu, J., Wu, B.: Physically adversarial attacks and defenses in computer vision: A survey. arXiv preprint arXiv:2211.01671 (2022)
41. Wei, X., Zhao, S., Li, B.: Revisiting the trade-off between accuracy and robustness via weight distribution of filters. IEEE Transactions on Pattern Analysis and Machine Intelligence (2024)
42. Wu, T., Tong, L., Vorobeychik, Y.: Defending against physically realizable attacks on image classification. arXiv preprint arXiv:1909.09552 (2019)
43. Xiao, C., Chen, Z., Jin, K., Wang, J., Nie, W., Liu, M., Anandkumar, A., Li, B., Song, D.: Densepure: Understanding diffusion models for adversarial robustness. In: The Eleventh International Conference on Learning Representations (2022)
44. Xiao, Z., Gao, X., Fu, C., Dong, Y., Gao, W., Zhang, X., Zhou, J., Zhu, J.: Improving transferability of adversarial patches on face recognition with generative models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11845–11854 (2021)
45. Xu, Q., Yang, Z., Zhao, Y., Cao, X., Huang, Q.: Rethinking label flipping attack: From sample masking to sample thresholding. IEEE Transactions on Pattern Analysis and Machine Intelligence **45**(6), 7668–7685 (2022)
46. Xu, W., Evans, D., Qi, Y.: Feature squeezing: Detecting adversarial examples in deep neural networks. arXiv preprint arXiv:1704.01155 (2017)
47. Yin, P., Lyu, J., Zhang, S., Osher, S., Qi, Y., Xin, J.: Understanding straight-through estimator in training activation quantized neural nets. arXiv preprint arXiv:1903.05662 (2019)

48. Yu, C., Chen, J., Xue, Y., Liu, Y., Wan, W., Bao, J., Ma, H.: Defending against universal adversarial patches by clipping feature norms. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 16434–16442 (2021)
49. Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
50. Zhao, S., Wang, X., Wei, X.: Mitigating accuracy-robustness trade-off via balanced multi-teacher adversarial distillation. IEEE Transactions on Pattern Analysis & Machine Intelligence (01), 1–14 (2024)
51. Zhong, Y., Liu, X., Zhai, D., Jiang, J., Ji, X.: Shadows can be dangerous: Stealthy and effective physical-world adversarial attack by natural phenomenon. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15345–15354 (2022)
52. Zhou, K., Yang, J., Loy, C.C., Liu, Z.: Learning to prompt for vision-language models. International Journal of Computer Vision **130**(9), 2337–2348 (2022)
53. Zhu, Z., Zhang, Y., Chen, H., Dong, Y., Zhao, S., Ding, W., Zhong, J., Zheng, S.: Understanding the robustness of 3d object detection with bird's-eye-view representations in autonomous driving. arXiv preprint arXiv:2303.17297 (2023)