Robust Single Image Reflection Removal Against Adversarial Attacks

Zhenbo Song[1] Zhenyuan Zhang[1*] Kaihao Zhang[2] Wenhan Luo [3✉]
Zhaoxin Fan[4] Wenqi Ren[3] Jianfeng Lu[1]

[1]Nanjing University of Science and Technology [2]Australian National University
[3]Shenzhen Campus of Sun Yat-sen University [4]Renmin University of China

Abstract

This paper addresses the problem of robust deep single-image reflection removal (SIRR) against adversarial attacks. Current deep learning based SIRR methods have shown significant performance degradation due to unnoticeable distortions and perturbations on input images. For a comprehensive robustness study, we first conduct diverse adversarial attacks specifically for the SIRR problem, i.e. towards different attacking targets and regions. Then we propose a robust SIRR model, which integrates the cross-scale attention module, the multi-scale fusion module, and the adversarial image discriminator. By exploiting the multi-scale mechanism, the model narrows the gap between features from clean and adversarial images. The image discriminator adaptively distinguishes clean or noisy inputs, and thus further gains reliable robustness. Extensive experiments on Nature, SIR^2, and Real datasets demonstrate that our model remarkably improves the robustness of SIRR across disparate scenes.

1. Introduction

Single image reflection removal (SIRR) is a classic topic in the low-level image processing area, namely a kind of image restoration. When taking an image through a transparent surface, a reflection layer would be blended with the original photography (*i.e.* the transmission layer), resulting in imaging corruptions. The SIRR is devoted to recovering a clear transmission image by removing the reflection layer. However, the SIRR is fundamentally ill-posed [42] that there could be an infinite number of transmission and reflection decompositions from a blended image. Therefore, traditional methods often exploit manual priors to optimize the layer separation, such as gradient sparsity prior [25] and relative smoothness prior [27]. These priors are often violated when facing complex scenes. Recently, deep learning based methods [10, 16, 55] have attracted considerable attention to tackle the SIRR problem. By learning semantic and contextual features, deep SIRR methods have achieved much better quality of the recovered images.

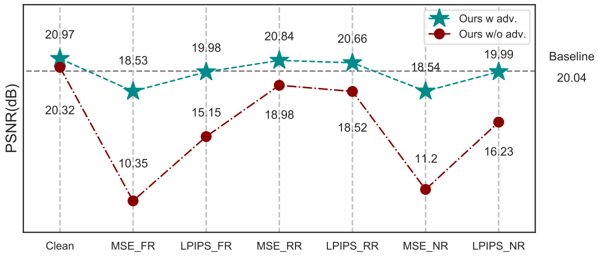


Figure 1. The PSNR measurements of our approach under different kinds of adversarial attacks. 'Clean' indicates no-attacks, 'MSE_FR' represents attacking on **F**ull **R**egion with MSE objective, and so on. The testing result is from Nature dataset [26].

However, deep neural networks are often vulnerable to visually imperceptible adversarial perturbations [14, 29]. The prediction can be totally invalid even with slight and unnoticeable attacks on inputs. Similarly, such vulnerability is also an important issue for the deep SIRR problem, and the robustness of current methods has not been thoroughly studied. There have been no benchmarks and evaluations for the robustness of deep SIRR models against intended attacks. Meanwhile, general defense methods [44] have not been applied to SIRR models. Accordingly, the robust SIRR model is still a crucial and desiderate research problem.

In this paper, we first investigate the robustness of deep SIRR methods. We apply the widely-used and powerful attack method PGD [29] to generate adversarial samples. For completeness of the robustness evaluation, we present various attack modes by referring [3, 48]. Specifically, we employ different attack objectives *i.e.* mean squared error (MSE) and learned perceptual image patch similarity (LPIPS) [52], as well as different attack regions, *i.e.* the

[*] Equal contribution
[✉] Corresponding Author: < whluo.china@gmail.com >.

full image region (FR), the reflection region (RR), and the non-reflection region (NR). Through a systematic analysis, the most effective attack mode and currently the most robust SIRR model are identified. Then we conduct adversarial training based on this model to enhance its robustness. In order to develop a furthermore robust SIRR model, we borrow the wisdom of multi-scale image processing [19] and adversarial discriminating [56] from previous defense methods. Consequently, we build a new robust SIRR model based on the image transformer [41], which integrates the cross-scale attention module, the multi-scale fusion module, and the adversarial image discriminator. The proposed method obtains significant improvements in robustness. Fig. 1 reveals the PSNR changes of our model prediction under distinct attack modes on the Nature dataset [26]. It is notable that our model yields limited degradations against perturbed images when compared with input clean images.

Overall, our main contributions can be summarized below. (1) We present a comprehensive evaluation of existing deep SIRR methods in terms of their robustness against various adversarial attacks on diverse datasets. Extensive experimental studies suggest presently the most effective attack and the most robust SIRR model. (2) We propose a novel transformer-based SIRR model, which integrates several relatively-robust modules to defend against adversarial attacks. The model can mitigate the effects of perturbations and distinguish clean or polluted inputs as well. (3) We carry out sufficient experiments to analyze the robustness of the proposed method, which achieves state-of-the-art stability against adversarial images. The model performs superior reflection removal robustness on distorted images while maintaining favorable accuracy on original clean images.

2. Related Work

2.1. Single-image reflection removal

Recently, deep learning-based methods have achieved remarkable success in image restoration [49–51], which also includes reflection removal [12, 40]. Existing methods can be divided into one-stage and multi-stage methods.

One-stage methods. Zhang *et al*. (ZN18) [53] apply conditional GAN [17] into the network and utilize perceptual information. Wei *et al*. (WY19) [42] introduce alignment-invariant loss to solve the training problem of unaligned real-world images. Wen *et al*. (WT19) [43] provide a network to predict the transmission layer, the reflection layer, and the alpha blending mask. Kim *et al*. (KH20) [22] propose a physically-based method for synthesizing images while taking into account the spatial variability of the visual effects of reflections.

Multi-stage methods. Fan *et al*. [12] are the first to use a two-stage deep learning network for estimating edges and reconstructing images respectively. Yang *et al*. (YG18) [46] present a multi-stage network that can sequentially estimate two layers. Li *et al*. (LY20) [26] introduce Long Short-Term Memory (LSTM) into the long cascade network to prevent gradient vanishing. Specifically, Chang *et al*. (CL21) [5] introduce a three-stage network, which first train the edge detector to obtain more edge information, then a reflection classifier for constraints and objectives, and final apply the decomposition mechanism to the final network training. Hu and Guo (HG21) [16] propose a general rule for deep interactive learning which means that the two branches should communicate with each other frequently by exchanging, rather than discarding useless information. Zheng *et al*. (ZS21) [55] propose a two-stage network considering the absorption effect in reflection removal, which first estimates the absorption effect and then takes it and a blended image as the input.

2.2. Adversarial Attacks and Defenses

Deep neural networks can misclassify images under the influence of imperceptible perturbations [1, 38, 54]. These perturbations are calculated by maximizing the prediction error of the network, which are called ”adversarial examples”. Szegedy *et al*. [38] propose the limited-memory BFGS (L-BFGS) attack, and convert the proposed optimization problem into an easily solvable box constraint form. Goodfellow *et al*. [14] propose the Fast Gradient Sign Method (FGSM) to generate adversarial examples that always tend to the direction of the gradient with a single gradient step. Unlike [14], Miyato *et al*. [31] calculate the adversarial perturbation based on the gradient value. Moosavi-Dezfooli *et al*. [32] propose a DeepFool algorithm to calculate the minimum necessary perturbations. Unlike the above one-step methods, Madry *et al*. [29] propose a multi-step solution method called Projected Gradient Descent (PGD).

To protect the security of deep learning models, there are many different strategies towards adversarial attacks, which can be basically divided into three categories, including gradient masking [2, 35], robust optimization [15, 24, 29, 34, 39, 56] and adversarial examples detection [4, 45]. Gradient masking is to hide the gradient, since most attacks leverage the gradient of the network. Robust optimization is to train the network on adversarial samples, which can learn how to restore the ground truth from adversarial perturbations and perform robustly. Adversarial examples detection is to learn to distinguish normal samples and adversarial samples and to prohibit the input of the latter.

Recently, the research about model robustness against adversarial attacks is thriving on the low-level computer vision applications, *e.g*. super-resolution [8, 33, 47], derain [48], deblurring [13]. Choi *et al*. [8] apply the adversarial attack to the super-resolution and evaluate all deep learning-based methods towards adversarial attack. Yu *et al*. [48] are the first to investigate adversarial attack in rain removal scenarios. They systematically evaluate the advantages of each module in existing methods for adversarial attacks and select effective modules to form their own model. Gandikota *et al*. [13] introduce adversarial attack in to image deblurring and evaluate te robustness of the models.

However, there is still no research on robust network towards adversarial attack in image reflection removal. An evaluation on adversarial attacks is necessary, and so are appropriate attack methods.

3. Attacks on Single Image Reflection Removal

The purpose of adversarial attacks is to generate slight perturbations on the input image, which are visually imperceptible but will deteriorate the prediction of deep SIRR networks. For adversarial perturbation generation, we utilize a classic optimization based approach PDG [29], which has been considered as a baseline attack method in some robustness evaluation works [48, 56].

Mathematically speaking, a captured image I is typically formulated as a linear addition of a transmission layer T and a reflection layer A, *i.e.* $I = T + A$. Given a pre-trained deep SIRR model $f(*)$, it recovers the transmission image T from I, *i.e.* $T \doteq f(I)$. In order to fool the deep model $f(*)$, slight perturbations δ are generated and added to I in a pixel-wise manner on the region of interest, deriving an attacking sample I'. The values of δ are within $[-\epsilon, \epsilon]$ to guarantee that perturbations are visually unnoticeable. Let R denote the attack region, where attacked regions are set to 1 (otherwise 0), and O represents the objective to measure the output degradation. Then the attacking image can be obtained by:

$$I' = I + R \cdot \delta, \tag{1}$$

where $\cdot$ denotes the element-wise multiplication. Subsequently, perturbations are optimized by maximizing the deviation of the attacked output from the original output.

$$\delta = arg \max_{\|\delta\| \leq \epsilon} O(f(I), f(I')). \tag{2}$$

PGD [29] is adopted to solve the optimization problem iteratively. After T times of updating, the final perturbations δ^* are derived.

For a more comprehensive attack analysis, we define two types of attack objectives and three different attack regions. Inspired by [48], one objective is from the aspect of pixel-wise image discrepancy, and the other objective focuses on the high-level perceptual similarity of output images. Details of the two objective functions are as follows.

- Mean Squared Error (MSE) directly measured on output images:

$$O = \|f(I) - f(I')\|_2 \,. \tag{3}$$

- Learned Perceptual Image Patch Similarity (LPIPS) measured by a neural network-based function ℓ_{lpips} [52]:

$$O = \ell_{lpips}(f(I), f(I')). \tag{4}$$

The attack regions are concerned with whether one pixel is mixed with reflection or not. If the discrepancy between the blended and transmission images exceeds an empirical tolerance θ, we consider this pixel into a reflection region. Thereby, the three attack regions are given specifically below.

- Full Region (FR) attack working on the whole image:

$$R = \mathbf{1}. \tag{5}$$

- Reflection Region (RR) attack working on pixels of great change:

$$R = abs(f(I) - I) > \theta. \tag{6}$$

- Non-reflection Region (NR) attack particularly working on unchanged pixels:

$$R = abs(f(I) - I) \leq \theta. \tag{7}$$

By combining the above objective attacks and regional attacks, we formulate six attack modes. In order to evaluate model performance, we calculate the PSNR and SSIM between predicted outputs and the ground truth transmission images T_{gt} from labeled datasets. The original and the degraded outputs are widely compared for better analysis of robustness. To summarize, the evaluation metrics are listed below.

$$\begin{aligned} m_1 = \text{PSNR}(f(I), T_{gt}), m_2 = \text{PSNR}(f(I'), T_{gt}), \\ m_3 = \text{SSIM}(f(I), T_{gt}), m_4 = \text{SSIM}(f(I'), T_{gt}), \end{aligned} \tag{8}$$

4. Robust SIRR Model

The overall architecture of the proposed network is illustrated in Fig. 2. It consists of a cross-scale image encoding stream, a multi-scale feature decoding stream, and an adversarial image discriminator (AID) controlling three dynamic convolution [6] modules (D-Conv). A blended image is first downsampled to its 1/2 scale and its 1/4 scale. Then using the cross-scale attention module, multi-input images are gradually taken into the encoding stream to obtain multi-scale deep features. Afterward, the decoding stream exploits multi-scale fusion modules to aggregate these feature maps to recover the output transmission images. Based on the perturbations of the input image, the AID generates the weights to compute dynamic convolution kernels. Hence, the main network could be adversarially-aware, and

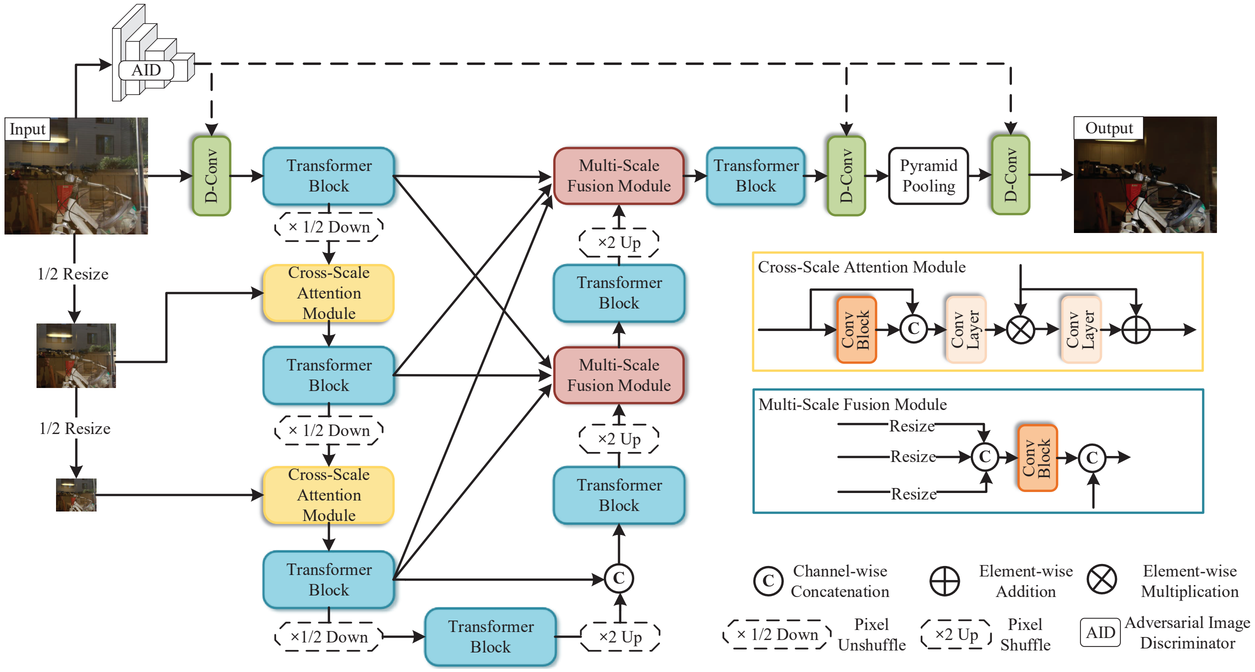


Figure 2. The overall architecture of our approach for robust single image reflection removal.

it is able to adaptively handle clean and corrupted inputs. Our model is mainly motivated by the robustness consideration. Inspired by previous research on visual attacks and defenses, there are three key ideas for robustness enhancement.

The first idea aims at the multi-scale image processing strategy, which has been verified as robust against adversarial attacks [19, 44]. We exploit the multi-input single encoder structure [7], *i.e.* the cross-scale attention module, which combines downsized high-level features and low-level features from downsized inputs to handle diverse image perturbations. Then instead of the typical skip connection, we utilize the multi-scale fusion module for further aggregating multi-scale complementary information. Implementation details of these two modules are introduced in the following subsections.

The second idea is to use transformer blocks as feature extractors. It is reported that attention operations can indeed improve the robustness of deep derain models [48]. Since SIRR is the same type of low-level vision task as derain, a similar insight could be migrated to the robust SIRR model. Therefore, we apply efficient RestoreFormer blocks [41] to construct feature extractors at different stages.

The third idea comes from the adversarial examples detection strategy, which learns to distinguish normal and adversarial samples. Furthermore, AID [56] upgraded this idea by generating adaptive convolution kernels with respect to different samples. We also employ the AID to offer dynamic convolutions on clean and corrupted images. Notably, existing adversarial training methods suffer from the bottleneck in that there could be a significant degradation on clean inputs while improving limited adversarial robustness [56]. AID is able to alleviate the issue to a certain extent.

4.1. Cross-Scale Attention Module

As aforementioned, this module combines two kinds of image features, *i.e.* the high-level feature from the deep main network, and the low-level feature from a shallow branch network. The structure of this module is demonstrated by the yellow boxes in Fig. 2. The left input stream takes the downsized image to extract features using a convolution block (CB). Then the features are concatenated with the input, and further refined by a convolution layer (CL_1) to derive low-level features. We denote the downsized image as I_{down}, then the low-level features F_{low} is calculated by:

$$F_{low} = CL_1(I_{down} \copyright CB(I_{down})), \tag{9}$$

where © denotes the channel-wise concatenation.

The top input stream takes features F_{high} from the previous feature level. F_{high} is first multiplied with F_{low} to obtain an attention map, following a convolution layer (CL_2) to smooth the attention map. The attention map is added to the original high-level features to derive the final output features F_{out}. This process is described below.

$$F_{out} = CL_2(F_{high} * F_{low}) + F_{high}. \tag{10}$$

4.2. Multi-Scale Fusion Module

For better aggregating encoding features, the multi-scale fusion module is conducted throughout the decoder by densely exchanging the information across the multi-resolution features. The structure of this module is shown by the blue boxes in Fig. 2. Here, we describe an example of fusing three-scale feature maps, as illustrated in Fig. 2. Fusion two scales or four scales can be easily derived.

In this example, the left input stream takes three scales of encoding feature maps, *e.g.* $F_{\times 2}$, $F_{\times 1}$, and $F_{1/2}$, and the corresponding output feature map is $F^{out}_{\times 1}$. Firstly, the feature map at scale $1/2$ is resized to the current scale by bilinear upsampling, while the feature map at scale 2 is downsampled half-scale using mean pooling. The feature map at the proper scale as the output is copied for later fusion. The fusion step concatenates re-sampled feature maps, and then several convolutional layers are adopted for fusing across multi-resolution features. The fused feature map is concatenated with features F_{btm} from the bottom input stream to obtain the final output. This fusion process can be formulated as follows,

$$\begin{aligned} F^{out}_{\times 1} =& CB(Up(F_{1/2}) \copyright F_{\times 1} \copyright Down(F_{\times 2})) \\ & \copyright F_{btm} \end{aligned}, \tag{11}$$

where the $Up(\cdot)$ and $Down(\cdot)$ denote the upsampling and downsampling operation respectively. The $CB(\cdot)$ represents convolutional blocks.

4.3. Adversarial Image Discriminator

In order to make our model discriminative to clean and adversarial inputs, the AID is employed to control several convolution kernels of the SIRR network. Before extracting features from the input image, the AID could first generate K probability vectors of the image. Then these vectors are devoted to calculating the convolution kernels of specific dynamic convolution layers. In the proposed SIRR model, we use three such dynamic layers. One dynamic layer is inserted before the encoder, one is embedded after the decoder, and the last one is used to generate the output. This arrangement balance the network's ability to learn shared or distinctive features.

4.4. Adversarial Training

We train the proposed method with both clean and adversarial samples. The training loss of the proposed network consists of one online triplet loss [36] to discriminate clean and adversarial inputs, along with three supervised losses to minimize the discrepancy between prediction outputs and the ground truth.

AID Loss. According to [56], the AID learns to distinguish the clean and adversarial inputs by measuring the distance between probability distributions of two kinds of image outputs. Let D_{in} denote the AID model. Given an anchor instance I_a, a positive instance I_p (*i.e.* same type of clean or adversarial image with I_a), and a negative instance I_n, the AID loss is defined as follows,

$$\begin{aligned} \mathcal{L}_{aid} = clip_{(0,+\infty)}(& JS(D_{in}(I_a)||D_{in}(I_p)) - \\ & JS(D_{in}(I_a)||D_{in}(I_n)) + \gamma) \end{aligned}, \tag{12}$$

where $JS(*||*)$ is the Jensen-Shannon divergence [28], and γ is an empirical margin.

Euclidean Loss. Basically, the SIRR problem is modeled as an end-to-end image translation problem, namely using a generator to estimate the transmission image from a single blended image. For training the generator, the pixel-level discrepancy between the network output and the ground truth is usually the fundamental loss function. Given the network output $T_{pred} = f(I) \ or \ f(I')$ and the corresponding ground truth image T_{gt}, this Euclidean loss is defined as

$$\mathcal{L}_1 = ||T_{pred} - T_{gt}||_1. \tag{13}$$

GAN Loss. In order to restore realistic transmission images, the GAN loss is also widely used. Thus, we suggest building a discriminator referring to the LSGAN [30]. The discriminator D_{out} is trained along with the generator f by applying alternating gradient updates. For training stability, we adopt the relativistic cost function [21], which is defined as:

$$\begin{aligned} \mathcal{L}_{adv} =& \mathbb{E}[(D_{out}(T_{gt}) - \mathbb{E}[D_{out}(T_{pred})] - 1)^2] + \\ & \mathbb{E}[(D_{out}(T_{pred}) - \mathbb{E}[D_{out}(T_{gt})] + 1)^2] \end{aligned}, \tag{14}$$

where $\mathbb{E}$ measures the average of the discriminative value of real and fake image pools. The GAN loss for optimizing the generator is to minimize Eq. (14). Additionally, it is observed in the PatchGAN [18] that sharper results are produced by using the discriminator on 'local' image patches rather than on the 'global' full image. Correspondingly, we adopt a combination of the local and global discriminator, and derive a double-scale objective [23] as:

$$\mathcal{L}_{adv} = \alpha \mathcal{L}_{adv1} + \mathcal{L}_{adv2}, \tag{15}$$

where $\mathcal{L}_{adv1}$ optimizes the generator using Eq. (14) on local patches, α is the corresponding local weights, and $\mathcal{L}_{adv2}$ measures over the full image.

Perceptual Loss. We utilize the perceptual loss [20], which measures high-level perceptual and semantic feature distances between images. This loss is computed over different activation maps from VGG-19 [37], which is usually pretrained on ImageNet [9]. The VGG-19 model is denoted as ϕ and the perceptual loss is computed using

$$\mathcal{L}_{feat} = \sum_{i=1}^{5} ||\phi_i(T_{pred}) - \phi_i(T_{gt})||_1, \tag{16}$$

where ϕ_i calculates the feature map after layers `relu1_1`, `relu2_1`, `relu3_1`, `relu4_1`, and `relu5_1` respectively.

The overall loss of our robust SIRR model is a linear combination of the above objectives. We empirically set the coefficients of each loss term, and then the final loss function is given as follows,

$$\mathcal{L} = \lambda_{\ell_1}\mathcal{L}_1 + \lambda_{adv}\mathcal{L}_{adv} + \lambda_{feat}\mathcal{L}_{feat} + \lambda_{aid}\mathcal{L}_{aid}. \quad (17)$$

5. Experiments

We first conduct adequate attacking experiments to evaluate the robustness of existing deep SIRR models. Then, based on the thorough analysis, we select a baseline model among these models. After adversarial training, we carry out sufficient experiments to compare the performance of the proposed method and the baseline method. Key components of our model are also evaluated via an ablation study.

5.1. Robustness Evaluation

Dataset. For the training dataset, we use both synthetic and real data, following [16, 42]. In terms of the synthetic data, we synthesize $7,643$ cropped images from PASCAL VOC 2012 dataset [11]. In terms of the real data, we adopt 90 real-world training images from [53]. For the testing dataset, we use the public dataset, *e.g.* Nature [26], SIR^2 [40], and Real dataset [53].

SIRR methods. We consider six state-of-the-art deep learning SIRR methods with various parametric quantities and modules, *e.g.* WY19 [42], WT19 [43], LY20 [26], CL21 [5], HG21 [16], and ZS21 [55]. Tab. 1 shows their characteristics in terms of the number of parameters and the modules.

Attack Levels. We utilize the PGD [29] to solve the optimization problem. For a more comprehensive evaluation, we set the perturbation level $\epsilon \in \{1/255, 2/255, 4/255, 8/255\}$. We set the iteration number $T = 20$ in MSE attack objective and $T = 30$ in LPIPS attack objective.

Fig. 3 compares the PSNR performance of deep SIRR methods under different attack objectives and different attack regions. As the perturbation level ϵ increases, the output quality degrades clearly. In some cases, *e.g.* HG21 [16] in subfigure (a), the PSNR has halved from 20 to 10 when $\epsilon = 8/255$. The visualization results are more intuitive as shown in Fig. 4. For HG21, the perturbed input image (a) has no visual difference from the original input image, but the prediction result (c) is totally corrupted compared to the original prediction (b). These results indicate that the attack method is effective in terms of various attack regions and different attack objectives. More visualization results can be found in the supplementary material. Although all methods show similar degradation trends against adversarial attacks, there are still differences with respect to attack modes. Besides, existing models show relatively superior or inferior robustness.

Table 1. Properties of deep SIRR methods. GAN denotes that the GAN loss is used. AT, RB denote attention module and residual blocks respectively. LSTM means using the LSTM mechanism.

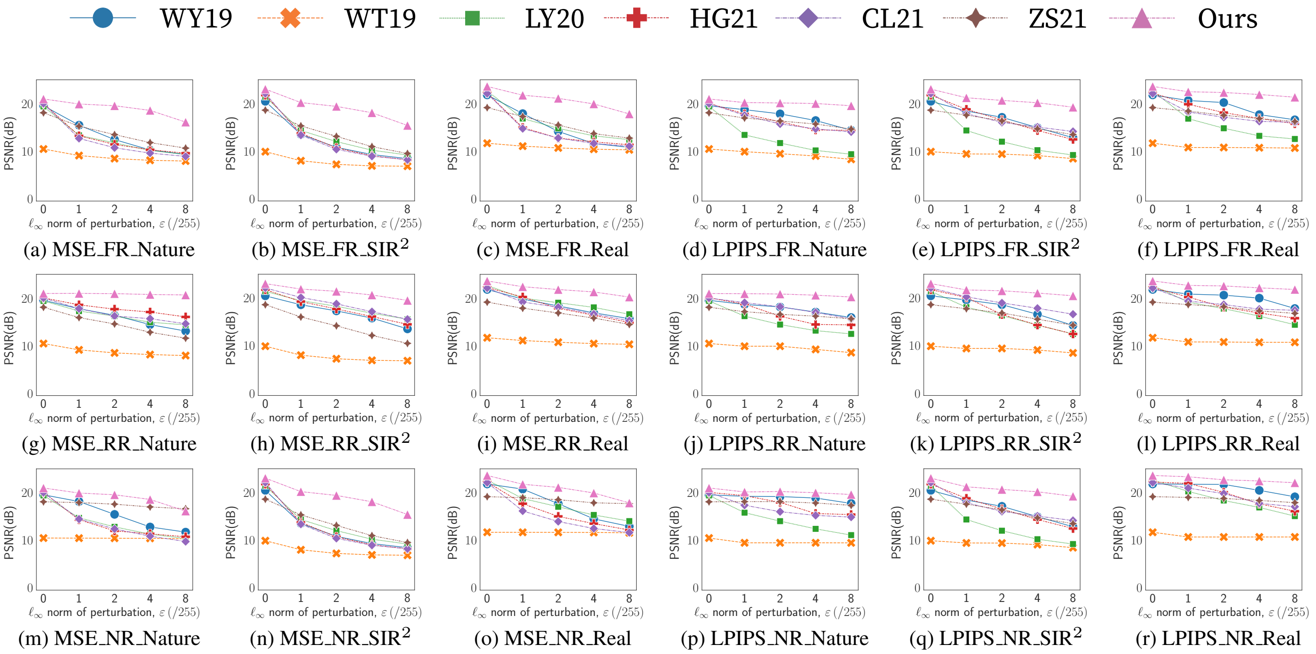
Method	Para.	GAN	AT	RB	LSTM
WY19 [42]	18M	✓	✓	✓	
WT19 [43]	65M				
LY20 [26]	24M	✓		✓	✓
CL21 [5]	275M				
HG21 [16]	39M	✓	✓		
ZS21 [55]	38M			✓	
Ours	25M	✓	✓	✓	

Attack mode comparison. By comparing the first three columns with the corresponding last three columns in Fig. 3, *e.g.* (a) and (d), it is observed that the MSE attack shows shaper decreasing trends than the LPIPS attack. In terms of attack regions, we notice that attack on the full region (the first row) or non-reflection region (the last row) is much more stable than on reflection regions (the middle row). SIRR models natively consider reflection regions as distortions and aim to remove the reflection effects. Perturbations on reflection regions may be judged to be reflection itself, which makes attacks on these regions much harder.

Robustness comparison. Generally, the robustness of SIRR methods is unsatisfactory because there is no adversarial training conducted. Among these SIRR methods, ZS21 [55] shows relatively robust in some conditions, *e.g.* in Fig. 3 (j), (m), (o), (p). However, its PSNR performance is not competitive. By contrast, WY19 [42] maintains high image quantity and exhibits robustness in Fig. 3 (d), (h), (k), (p). According to Tab. 1, WY19 utilizes the attention mechanism, residual blocks, and the GAN loss, which are summarized as the key components for robustness in [48]. It is also worth noticing that our model (the triangle pink line in Fig. 3) presents superior robustness under these attacks.

5.2. Robust SIRR Results

Since WY19 [42] is relatively the most robust model, we use it as the baseline model to compare with our model. For making a fair comparison, we train the WY19 model adversarially, thus bringing its robustness into full play. As a multi-stage pipeline, WY19 utilizes a VGG-19 pre-trained on ImageNet to extract fundamental features as input to the main network. WY19 may benefit from this way when training on small datasets. Nevertheless, we still bring the pre-trained VGG-19 into adversarial training. We also train another model of our method, which only uses clean training samples accordingly. Specifically, we set the AID loss Eq. (12) to constant zero and keep the rest loss func-

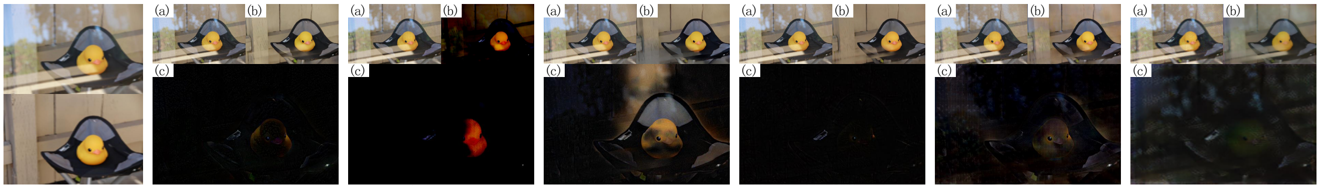


(a) MSE_FR_Nature (b) MSE_FR_SIR^2 (c) MSE_FR_Real (d) LPIPS_FR_Nature (e) LPIPS_FR_SIR^2 (f) LPIPS_FR_Real

(g) MSE_RR_Nature (h) MSE_RR_SIR^2 (i) MSE_RR_Real (j) LPIPS_RR_Nature (k) LPIPS_RR_SIR^2 (l) LPIPS_RR_Real

(m) MSE_NR_Nature (n) MSE_NR_SIR^2 (o) MSE_NR_Real (p) LPIPS_NR_Nature (q) LPIPS_NR_SIR^2 (r) LPIPS_NR_Real

Figure 3. Comparison of the PSNR values with respect to perturbation levels ϵ for different attacks on various datasets. 'MSE_FR_Nature' represents attacking on **F**ull **R**egion with MSE objective on the Nature [26] dataset, and so the others. Best view by zooming in.



(1) Input/GT (2) WY19 (3) WT19 (4) LY20 (5) HG21 (6) CL21 (7) ZS21

Figure 4. Visual comparison by attacking with $\epsilon = 8/255$. The attack is on the full region with MSE objectives. The left column is the original blended image and the ground truth transmission image. Subfigures (a), (b), and (c) are perturbed input images, the output of inputting the original image, and the output of inputting the perturbed images, respectively. Best view by zooming in.

tions. We analyze the model robustness from two perspectives, *i.e.* the output degradation of inputting adversarial images over inputting clean ones, and the output performance of inputting clean images before and after adversarial training. Besides, we further conduct the ablation study to evaluate key components of our model.

Robustness analysis. The quantitative results of the baseline and our method are reported in Tab. 2. Without adversarial training, our method gets better PSNR and SSIM on clean images. Yet, under the MSE or LPIPS attack, the performance of our model decreases more significantly than WY19. Taking PSNR on the Nature dataset as an example, our method outperforms the baseline, saying 20.33 vs. 19.54. Whereas, when attacked by the MSE objective, our result drops by 9.98 to 10.35, while the baseline only decreases to 11.86. The same issue is widely verified on the other two datasets regardless of which attack mode is adopted. As mentioned above, it is understandable that WY19 is based on features from a pre-trained model. Since many data augmentation methods have been used to train it, the model can be stable to limited input perturbations. By contrast, adversarial training alleviates the effects of perturbation inputs. When the MSE attack is applied to WY19 and our method, the reductions of PSNR on the Nature dataset become 2.26 and 1.79 respectively. This indicates the robustness improvement, and our method predicts the output more steadily. According to testing results in Tab. 2, our method shows great robustness over WY19 on the Nature and Real dataset. The degradation of our method is more significant on the SIR^2 dataset, but the absolute metric values are much higher than WY19.

Besides, it is observed that the PSNR of WY19 on the Nature dataset is decreasing along with adversarial training, while our method keeps high PSNR accuracy. In this regard, the robustness is mainly contributed by the AID [56]. Notably, in terms of absolute PSNR and SSIM values, our

Table 2. Comparison of different training strategies on three benchmark datasets. ‘w/’ and ‘w/o adv.’ mean training with or without adversarial images. MSE and LPIPS denote corresponding attacks over full regions. ↓ and ↑ represent the degradation and degradation performance compared to the original prediction inputting clean images.

		Nature		SIR^2		Real	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
WY19 [42] w/o adv.	Clean	19.54	0.738	20.45	0.853	21.82	0.812
	MSE	11.86$^{\downarrow 7.68}$	0.361$^{\downarrow 0.377}$	10.49$^{\downarrow 9.97}$	0.410 $^{\downarrow 0.442}$	13.61 $^{\downarrow 8.21}$	0.388 $^{\downarrow 0.424}$
	LPIPS	16.85$^{\downarrow 2.69}$	0.588 $^{\downarrow 0.149}$	15.81 $^{\downarrow 4.65}$	0.677 $^{\downarrow 0.176}$	18.79$^{\downarrow 3.04}$	0.639 $^{\downarrow 0.173}$
WY19 [42] w/ adv.	Clean	17.28$^{\downarrow 2.26}$	0.670$^{\downarrow 0.067}$	17.97$^{\downarrow 2.49}$	0.832$^{\downarrow 0.021}$	19.23$^{\downarrow 2.59}$	0.752$^{\downarrow 0.060}$
	MSE	16.08$^{\downarrow 3.46}$	0.613 $^{\downarrow 0.125}$	16.54 $^{\downarrow 3.92}$	0.769 $^{\downarrow 0.083}$	18.61 $^{\downarrow 3.21}$	0.718$^{\downarrow 0.094}$
	LPIPS	17.01 $^{\downarrow 2.53}$	0.633 $^{\downarrow 0.105}$	17.49 $^{\downarrow 2.96}$	0.779 $^{\downarrow 0.074}$	16.64 $^{\downarrow 5.18}$	0.702 $^{\downarrow 0.110}$
Ours w/o adv.	Clean	20.33	0.758	**23.43**	**0.894**	22.26	0.826
	MSE	10.35 $^{\downarrow 9.98}$	0.264 $^{\downarrow 0.494}$	9.18 $^{\downarrow 14.24}$	0.317 $^{\downarrow 0.577}$	11.92 $^{\downarrow 10.34}$	0.274$^{\downarrow 0.552}$
	LPIPS	15.15 $^{\downarrow 5.18}$	0.560 $^{\downarrow 0.198}$	14.84 $^{\downarrow 8.59}$	0.645 $^{\downarrow 0.250}$	16.38 $^{\downarrow 5.88}$	0.573$^{\downarrow 0.253}$
Ours w/ adv.	Clean	**20.97**$^{\uparrow 0.64}$	**0.764**$^{\uparrow 0.006}$	23.02$^{\downarrow 0.41}$	0.892$^{\downarrow 0.002}$	**23.61**$^{\uparrow 1.35}$	**0.835**$^{\uparrow 0.009}$
	MSE	18.53$^{\downarrow 1.79}$	0.726$^{\downarrow 0.032}$	18.25 $^{\downarrow 5.17}$	0.821 $^{\downarrow 0.073}$	20.15 $^{\downarrow 2.11}$	0.752$^{\downarrow 0.074}$
	LPIPS	19.98 $^{\downarrow 0.35}$	0.732$^{\downarrow 0.026}$	20.31 $^{\downarrow 3.12}$	0.830 $^{\downarrow 0.064}$	22.02 $^{\downarrow 0.24}$	0.768 $^{\downarrow 0.058}$

Table 3. Comparison of different model settings on the Real [53] dataset. ‘Ours w/o CSA’ represents removing multi-resolution inputs and cross-scale attention modules. ’Ours w/o MSF’ represents removing multi-scale fusion modules and using traditional skip connections instead. ‘Ours w/o AID’ represents removing dynamic convolutions and the adversarial image discriminator.

		PSNR	SSIM
Ours w/o CSA	Clean	22.09	0.814
	MSE	19.38	0.608
	LPIPS	20.90	0.741
Ours w/o MSF	Clean	22.30	0.814
	MSE	19.29	0.732
	LPIPS	21.02	0.762
Ours w/o AID	Clean	21.86	0.813
	MSE	19.77	0.739
	LPIPS	21.11	0.757
Ours	Clean	**23.61**	**0.835**
	MSE	20.15	0.752
	LPIPS	22.02	0.768

method achieves state-of-the-art performance on both clean and adversarial images.

Ablation study. To verify the effectiveness of our network design, we compare four architectures with or without the key modules described in Sec. 4. Specifically, to construct three new networks, we remove cross-scale attention modules (CSA), multi-scale fusion modules (MSF), and the adversarial image discriminator (AID), while maintaining other structures unchanged. All networks are trained adversarially and with the same hyper-parameters, such as learning rate, batch size, and the number of epochs. In Tab. 3, we report PSNR and SSIM values of ablated networks with respect to clean and adversarial inputs. For clean inputs, the CSA and MSF contribute equally, which makes limited accuracy improvements. The AID is the most important assembly for guaranteeing the prediction accuracy of clean inputs. For adversarial inputs, the CSA and MSF modules show their importance in keeping robustness. Without CSA modules or MSF modules, PSNR values under the MSE attack will decrease by about 0.5 worse than that without the AID. These results suggest that the network components are carefully customized toward robustness enhancement.

6. Conclusion

This paper has evaluated the robustness of deep learning based SIRR methods against adversarial attacks, for which the PGD method is used to optimize attack perturbation on three regions and toward two objectives. Benchmark results show that current deep SIRR methods all inevitably degrade under adversarial attacks. We propose a robust transformer-based SIRR model, which integrates cross-scale attention modules, multi-scale fusion modules, and the adversarial image discriminator. Extensive experiments show state-of-the-art robustness over current methods.

7. Acknowledgement

This work was supported in part by the Jiangsu Funding Program for Excellent Postdoctoral Talent (No.2022ZB268), in part by the Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012839), in part by Shenzhen Science and Technology Program (No. JSGG20220831093004008)

References

[1] Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020. 2

[2] Anish Athalye, Nicholas Carlini, and David Wagner. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *International Conference on Machine Learning*, pages 274–283. PMLR, 2018. 2

[3] Nicholas Carlini, Anish Athalye, Nicolas Papernot, Wieland Brendel, Jonas Rauber, Dimitris Tsipras, Ian J. Goodfellow, Aleksander Madry, and Alexey Kurakin. On evaluating adversarial robustness. *CoRR*, abs/1902.06705, 2019. 1

[4] Nicholas Carlini and David Wagner. Adversarial examples are not easily detected: Bypassing ten detection methods. In *ACM Workshop on Artificial Intelligence and Security*, pages 3–14, 2017. 2

[5] Ya-Chu Chang, Chia-Ni Lu, Chia-Chi Cheng, and Wei-Chen Chiu. Single image reflection removal with edge guidance, reflection classifier, and recurrent decomposition. pages 2033–2042, 2021. 2, 6

[6] Yinpeng Chen, Xiyang Dai, Mengchen Liu, Dongdong Chen, Lu Yuan, and Zicheng Liu. Dynamic convolution: Attention over convolution kernels. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 11030–11039, 2020. 3

[7] Sung-Jin Cho, Seo-Won Ji, Jun-Pyo Hong, Seung-Won Jung, and Sung-Jea Ko. Rethinking coarse-to-fine approach in single image deblurring. In *IEEE International Conference on Computer Vision*, pages 4641–4650, 2021. 4

[8] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 303–311, 2019. 2

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

[10] Zheng Dong, Ke Xu, Yin Yang, Hujun Bao, Weiwei Xu, and Rynson WH Lau. Location-aware single image reflection removal. In *IEEE International Conference on Computer Vision*, pages 5017–5026, 2021. 1

[11] Mark Everingham, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *International Journal of Computer Vision*, 88(2):303–338, 2010. 6

[12] Qingnan Fan, Jiaolong Yang, Gang Hua, Baoquan Chen, and David Wipf. A generic deep architecture for single image reflection removal and image smoothing. In *IEEE International Conference on Computer Vision*, pages 3238–3247, 2017. 2

[13] Kanchana Vaishnavi Gandikota, Paramanand Chandramouli, and Michael Moeller. On Adversarial Robustness of Deep Image Deblurring. In *IEEE International Conference on Image Processing*, pages 3161–3165. IEEE, 2022. 2, 3

[14] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. 1, 2

[15] Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel. On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017. 2

[16] Qiming Hu and Xiaojie Guo. Trash or Treasure? An Interactive Dual-Stream Strategy for Single Image Reflection Separation. *Advances in Neural Information Processing Systems*, 34:24683–24694, 2021. 1, 2, 6

[17] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 2

[18] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1125–1134, 2017. 5

[19] Jiahuan Ji, Baojiang Zhong, and Kai-Kuang Ma. Multi-Scale Defense of Adversarial Images. In *IEEE International Conference on Image Processing*, pages 4070–4074. IEEE, 2019. 2, 4

[20] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision*, pages 694–711. Springer, 2016. 5

[21] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. In *International Conference on Learning Representations*, 2018. 5

[22] Soomin Kim, Yuchi Huo, and Sung-Eui Yoon. Single image reflection removal with physically-based training images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5164–5173, 2020. 2

[23] Orest Kupyn, Tetiana Martyniuk, Junru Wu, and Zhangyang Wang. Deblurgan-v2: Deblurring (orders-of-magnitude) faster and better. In *IEEE International Conference on Computer Vision*, pages 8878–8887, 2019. 5

[24] Alexey Kurakin, Ian J. Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *International Conference on Learning Representations*, 2017. 2

[25] Anat Levin, Assaf Zomet, and Yair Weiss. Learning to perceive transparency from the statistics of natural scenes. *Advances in Neural Information Processing Systems*, 15, 2002. 1

[26] Chao Li, Yixiao Yang, Kun He, Stephen Lin, and John E. Hopcroft. Single image reflection removal through cascaded refinement. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3565–3574, 2020. 1, 2, 6, 7

[27] Yu Li and Michael S. Brown. Single image layer separation using relative smoothness. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2752–2759, 2014. 1

[28] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Trans. Inf. Theory*, 37(1):145–151, 1991. 5

[29] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learn-

ing models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018. 1, 2, 3, 6

[30] Xudong Mao, Qing Li, Haoran Xie, Raymond YK Lau, Zhen Wang, and Stephen Paul Smolley. Least squares generative adversarial networks. In *IEEE International Conference on Computer Vision*, pages 2794–2802, 2017. 5

[31] Takeru Miyato, Andrew M. Dai, and Ian J. Goodfellow. Adversarial training methods for semi-supervised text classification. In *International Conference on Learning Representations*, 2017. 2

[32] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard. Deepfool: A simple and accurate method to fool deep neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2574–2582, 2016. 2

[33] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *IEEE Transactions on Image Processing*, 29:1711–1724, 2019. 2

[34] Nicolas Papernot, Patrick McDaniel, Arunesh Sinha, and Michael Wellman. Towards the science of security and privacy in machine learning. *arXiv preprint arXiv:1611.03814*, 2016. 2

[35] Nicolas Papernot, Patrick McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy*, pages 582–597. IEEE, 2016. 2

[36] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 815–823, 2015. 5

[37] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015. 5

[38] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014. 2

[39] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems*, 2020. 2

[40] Renjie Wan, Boxin Shi, Ling-Yu Duan, Ah-Hwee Tan, and Alex C. Kot. Benchmarking single-image reflection removal algorithms. In *IEEE International Conference on Computer Vision*, pages 3922–3930, 2017. 2, 6

[41] Zhouxia Wang, Jiawei Zhang, Runjian Chen, Wenping Wang, and Ping Luo. Restoreformer: High-quality blind face restoration from undegraded key-value pairs. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 17512–17521, 2022. 2, 4

[42] Kaixuan Wei, Jiaolong Yang, Ying Fu, David Wipf, and Hua Huang. Single image reflection removal exploiting misaligned training data and network enhancements. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 8178–8187, 2019. 1, 2, 6, 8

[43] Qiang Wen, Yinjie Tan, Jing Qin, Wenxi Liu, Guoqiang Han, and Shengfeng He. Single image reflection removal beyond linearity. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3771–3779, 2019. 2, 6

[44] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018. 1, 4

[45] Weilin Xu, David Evans, and Yanjun Qi. Feature squeezing: Detecting adversarial examples in deep neural networks. In *Network and Distributed System Security Symposium*. The Internet Society, 2018. 2

[46] Jie Yang, Dong Gong, Lingqiao Liu, and Qinfeng Shi. Seeing deeply and bidirectionally: A deep learning approach for single image reflection removal. In *European Conference on Computer Vision*, pages 654–669, 2018. 2

[47] Minghao Yin, Yongbing Zhang, Xiu Li, and Shiqi Wang. When deep fool meets deep prior: Adversarial attack on super-resolution network. In *ACM Multimedia*, pages 1930–1938, 2018. 2

[48] Yi Yu, Wenhan Yang, Yap-Peng Tan, and Alex C. Kot. Towards Robust Rain Removal Against Adversarial Attacks: A Comprehensive Benchmark Analysis and Beyond. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6013–6022, 2022. 1, 2, 3, 4, 6

[49] Kaihao Zhang, Wenhan Luo, Yanjiang Yu, Wenqi Ren, Fang Zhao, Changsheng Li, Lin Ma, Wei Liu, and Hongdong Li. Beyond monocular deraining: Parallel stereo deraining network via semantic prior. *International Journal of Computer Vision*, 130(7):1754–1769, 2022. 2

[50] Kaihao Zhang, Wenhan Luo, Yiran Zhong, Lin Ma, Bjorn Stenger, Wei Liu, and Hongdong Li. Deblurring by realistic blurring. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 2737–2746, 2020. 2

[51] Kaihao Zhang, Wenqi Ren, Wenhan Luo, Wei-Sheng Lai, Björn Stenger, Ming-Hsuan Yang, and Hongdong Li. Deep image deblurring: A survey. *International Journal of Computer Vision*, 130(9):2103–2130, 2022. 2

[52] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018. 1, 3

[53] Xuaner Zhang, Ren Ng, and Qifeng Chen. Single image reflection separation with perceptual losses. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 4786–4794, 2018. 2, 6, 8

[54] Zhengyu Zhao, Zhuoran Liu, and Martha A. Larson. On success and simplicity: A second look at transferable targeted attacks. In *Advances in Neural Information Processing Systems*, pages 6115–6128, 2021. 2

[55] Qian Zheng, Boxin Shi, Jinnan Chen, Xudong Jiang, Ling-Yu Duan, and Alex C. Kot. Single image reflection removal with absorption effect. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 13395–13404, 2021. 1, 2, 6

[56] Liang Lin Ziyi Dong, Pengxu Wei. Adversarially-aware robust object detector. In *European Conference on Computer Vision*, 2022. 2, 3, 4, 5, 7