Robust Real-World Image Super-Resolution against Adversarial Attacks

Jiutao Yue*
Sun Yat-sen University
yuejt@mail2.sysu.edu.cn

Haofeng Li*
Shenzhen Research Institute of Big Data, Guangdong Provincial Key Laboratory of Big Data Computing, The Chinese University of Hong Kong, Shenzhen
lhaof@foxmail.com

Pengxu Wei†
Sun Yat-sen University
weipx3@mail.sysu.edu.cn

Guanbin Li
Sun Yat-sen University
liguanbin@mail.sysu.edu.cn

Liang Lin
Sun Yat-sen University
linliang@ieee.org

ABSTRACT

Recently deep neural networks (DNNs) have achieved significant success in real-world image super-resolution (SR). However, adversarial image samples with quasi-imperceptible noises could threaten deep learning SR models. In this paper, we propose a robust deep learning framework for real-world SR that randomly erases potential adversarial noises in the frequency domain of input images or features. The rationale is that on the SR task clean images or features have a different pattern from the attacked ones in the frequency domain. Observing that existing adversarial attacks usually add high-frequency noises to input images, we introduce a novel random frequency mask module that blocks out high-frequency components possibly containing the harmful perturbations in a stochastic manner. Since the frequency masking may not only destroys the adversarial perturbations but also affects the sharp details in a clean image, we further develop an adversarial sample classifier based on the frequency domain of images to determine if applying the proposed mask module. Based on the above ideas, we devise a novel real-world image SR framework that combines the proposed frequency mask modules and the proposed adversarial classifier with an existing super-resolution backbone network. Experiments show that our proposed method is more insensitive to adversarial attacks and presents more stable SR results than existing models and defenses.

CCS CONCEPTS

• **Computing methodologies** → **Image processing**.

*Both authors contributed equally to this research.
†Corresponding author

MM '21, October 20–24, 2021, Virtual Event, China.

ACM ISBN 978-1-4503-8651-7/21/10…$15.00
https://doi.org/10.1145/3474085.3475627

KEYWORDS

Real-world image super-resolution, Adversarial robustness, Adversarial attack, Deep neural networks

ACM Reference Format:
Jiutao Yue, Haofeng Li, Pengxu Wei, Guanbin Li, and Liang Lin. 2021. Robust Real-World Image Super-Resolution against Adversarial Attacks. In *Proceedings of the 29th ACM Int'l Conference on Multimedia (MM '21), Oct. 20–24, 2021, Virtual Event, China.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3474085.3475627

1 INTRODUCTION

Single image super-resolution (SISR) is to recover high-resolution (HR) visual contents with clearer details and better fidelity from a degraded low-resolution (LR) image. Image super-resolution is a fundamental problem in the field of image processing and multimedia, and has been widely investigated for a long time. Image SR algorithms could play an essential role in a variety of applications, such as multi-media processing [8, 14, 29, 55], medical imaging [9, 40], depth imaging [49, 60], and remote sensing [7, 63]. In recent years, deep convolutional neural networks (CNNs) have demonstrated superior performance over traditional SR algorithms, due to the strong capacity of DNNs. Training CNNs usually requires a large number of paired samples which contain low-resolution input images and their corresponding high-quality version. For image SR, a straightforward way is to obtain the degraded input images by downsampling existing high-resolution images [4, 51]. Such simulated datasets fail to model the complicated blur kernels in practical applications, which severely drops the performance of learning-based SR methods [16, 58]. To address the issue, some recent works [5, 6, 54, 64] have raised the real-world image super-resolution (RealSR) task and built a realistic benchmark by zooming out and in the optical lens in DSLR cameras. Deep CNNs models trained with real-world datasets are more robust to practical noises and image degeneration.

However, deep neural networks have been notoriously threatened by adversarial attacks [3, 32, 41] that synthesize an adversarial sample by adding subtle noises to a natural image. Such adversarial noises are usually computed with a deliberately incorrect supervision, and the resulted adversarial sample could mislead the target

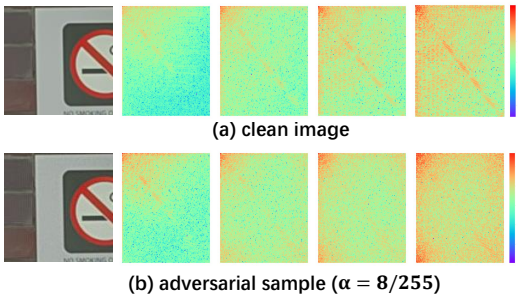

Figure 1: Visualization of 2D Discrete Cosine Transform (DCT) of the features extracted by a SR network [54]. DCT reflects the presence of signals with different frequency. In a DCT map, low-frequency signals are encoded at the upper left while high-frequency components are located at the bottom-left, bottom-right and top-right regions. In the above maps, a region in warmer color has larger values. (a) shows a clean image and DCT maps of its neural features while (b) shows the adversarial samples and its DCTs. It can be seen that attacked features have larger values in the high-frequency components than clean features.

neural model considerably. The same phenomenon also occurs in the single-image super-resolution task [10]. Adversarial samples could make a deep learning based SR model to predict undesirable artifacts, which implies that existing learning-based real SR methods may lack generalization and still suffer from unknown degradation kernels.

Most existing defenses are proposed for high-level image understanding tasks [23, 57, 61], which may be unpromising in the low-level image super-resolution task. In this paper, we investigate how adversarial perturbations affect SR models from a frequency perspective, by visualizing the frequency domain (*e.g.* Discrete Cosine Transform, DCT) of image features as shown in Figure 1. We surprisingly find that adversarial attacks did change the DCT pattern of an image feature map extracted by deep SR models. For a natural image, its frequency domain map usually contains the most significant values in low-frequency coefficients which encode the flat image regions of similar colors. The middle-to-high frequency components typically have the second largest values, which reflects the presence of sharp edges and corners. The smallest coefficients usually occur at the high-frequency component which is related to the highly repetitive elements in an image, such as noises, texture and artifacts. Interestingly, attacking a SISR model considerably increases the high-frequency coefficients from either vertical or horizontal direction in the attacked image features, which corresponds to densely distributed adversarial noises. The DCT difference between clean features and the attacked ones provides a hint to detect and resist adversarial samples.

Motivated by the above observations, we propose to improve the robustness of SR neural networks with a frequency mask module that reduces adversarial noises in the level of frequency domain. Considering that adversarial noises are parts of the high-frequency elements in an image, as shown in Figure 1. The proposed mask module fills zeros in the high-frequency components for the DCT of images or features to destroy the adversarial perturbations. On the other hand, an image may contain highly-repetitive textures, which are encoded as high-frequency parts as the noises. It is hard to separate the harmful noises from natural textures in a frequency domain. We find that a component of higher frequency is more likely to encode adversarial noises, and design a probability distribution to randomly determine if setting a coefficient as zero. Such an improvement could better preserve the original contents of the input image. For a clean image, it is undesirable to discard its high-frequency elements which may include finer details of the image. Thus we further devise an image classifier based on the observation that adversarial samples have a different distribution of the frequency domain from the original images. The proposed classifier takes a visualized frequency domain map as input and predicts if the corresponding image is an adversarial sample. If not, our proposed frequency mask module could be skipped so that the sharpness and fidelity of clean images are well maintained. Then we deploy the random frequency mask module and the adversarial sample classifier to an existing neural network to construct a robust super-resolution model. In summary, our main contributions have three folds:

- We propose a random frequency mask module that erases the high-frequency components of input images and features with a prior probability distribution to mitigate adversarial attacks.
- We introduce a frequency-based adversarial sample classifier which determines if applying the proposed mask module and helps maintain sharp details for clean images.
- We develop a robust image super-resolution network with the proposed frequency mask module and the adversarial sample classifier. The proposed method not only achieves satisfactory super-resolved results on real-world images, but also obtains the state-of-the-art performance against adversarial attacks.

2 RELATED WORK

2.1 Real-World Image Super-Resolution

Single-image super-resolution (SISR) requires to synthesize high-resolution contents from a single low-resolution image. SISR algorithms have been studied for a long time, and can be grouped into two categories: traditional and learning-based methods. Traditional SR methods are mainly based on edge priors [17, 28], image registration [27] and statistics [2, 30, 48]. Among learning-based SR methods [22, 31, 43, 59], DNN [13, 33, 62] is one of the most popular and effective models. According to the source of degraded images, SISR tasks could be divided into three groups: non-blind synthetic SR, blind image SR and real-world SR. Non-blind synthetic SR [12, 21] simply adopts bicubic or Gaussian downsampling to simulate the image degradation on both training and testing set, but the SR models trained with the synthetic dataset may fail to adapt to unknown blur kernels in practice. Blind image SR [19, 25] is based on a more realistic setting that the blur kernels during the testing are unavailable when training a SR model. Blind image SR algorithms could be evaluated with synthetic or real image sets, but these real datasets lack HR ground truths.

To fill the gap, real-world image SR task [5, 6, 54, 64] has been proposed, upsampling real degraded images by collecting many

pairs of real LR sample and its HR version. Cai *et al.* [5] employ two types of digital cameras to collect real SR pairs as a new dataset, RealSR. Wei *et. al.* [54] further provide a larger real-world SR benchmark DRealSR, which consists of well-aligned SR pairs captured by up to five different DSLR cameras. Particularly, Cai *et al.* [5] reveal that existing SR networks trained on a simulated dataset [51] do not show advantages with real-world samples. It is because real degradation includes lots of factors, *e.g.*, anisotropic blur, signal-dependent noise and cross-camera degradation processes. Thus real-world image SR is a more challenging and meaningful problem.

To solve this task, Cai *et. al.* [5] develop a laplacian pyramid kernel prediction network LP-KPN to explicitly learn a specific restoration kernel for each pixel. Wei *et. al.* [54] introduce a component divide-and-conquer model CDC that constructs three Component-Attentive Blocks (CAB) associated with flatten regions, edges, and corners. CDC infers super-resolved contents with the outputs of three CABs, and obtains the state-of-the-art results in real-world image super-resolution. However, these real-world SR networks may still be sensitive to adversarial noises, even though they are trained with complicated blur kernels. In this paper we focus on implementing a robust neural network for real-world SISR.

2.2 Adversarial Attacks and Defenses

Previous studies have shown that DNNs are vulnerable to adversarial attacks [41, 46, 50] which cheat a neural model by applying inconspicuous changes to an input image. Adversarial attacks can be categorized into two groups: black-box and white-box attacks, according to the knowledge acquired by attackers. Black-box attackers [26, 44] are allowed to access only limited knowledge of the data and the targeted network, such as the output of querying the targeted model. In this paper we only consider white-box attacks [24, 50] in which all parameters of the target model are exposed to the attacker. As for white-box attacks, Szegedy *et al.* [50] for the first time propose an adversarial attack for deep learning models, by maximizing the classification loss and minimizing the magnitude of the adversarial perturbation with a box-constrained L-BFGS. Goodfellow *et al.* [18] introduce an attack, fast gradient sign method (FGSM), which computes backward propagated gradients to maximize the classification loss and takes the sign of gradients to update an adversarial sample. Kurakin *et al.*[32] further extend FGSM to an iterative variant, I-FGSM. To attack an image SR network, Choi *et al.* [10] respectively implement basic, universal and partial attacks based on I-FGSM. The basic and the universal attacks can affect existing CNN-based image super-resolution models considerably. Following Choi *et al.* [10], we adopt the basic attack in this paper since it presents higher success rate of attack than the universal one.

Many defense methods and robust models [11, 38, 45, 68] which have been developed to resist adversarial attacks, attempt to remove, destroy or adapt to adversarial noises. Adversarial training [47, 52, 67] is a large group of defense methods that trains the target network with adversarial samples and are effective against white-box attacks. Some early defenses that apply image transformations [20, 37, 56] to disrupt or eliminate adversarial noises, are limited to gray-box setting [3] where the defense is unknown for attackers. Other improved methods that propose new neural modules to denoise or smooth image features [23, 35, 57], do show their robustness. Zhang *et al.* [66] propose to suppress high frequency in the discrete Fourier transform (DFT) of an input image, which is mostly related to our proposed method. Zhang *et al.* [66] utilize a fixed radius to erase noises, which discarded all the high-frequency elements including natural textures in an image. Differently, we attempt to maintain some parts of high-frequency components by randomly masking the perturbations, so that our proposed method has the chance to reconstruct the original high-frequency textures.

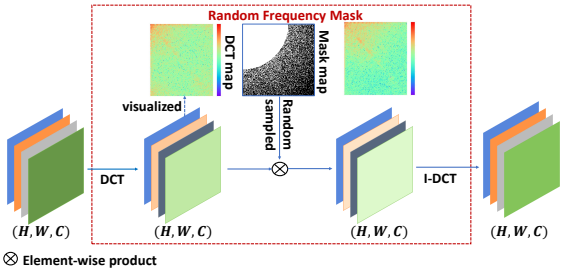

Figure 2: The architecture of our proposed random frequency mask module. The input of the mask module could be an image or a feature map extracted by super-resolution networks. First, the input image or feature is mapped into a frequency domain by applying discrete cosine transformation to each channel slice. Then a sector mask is synthesized based on a Bernoulli distribution, and is used to element-wisely multiply with the DCT maps to lower potential adversarial noises. Afterwards, the output feature of our proposed module is obtained by performing an inverse-DCT operation on the masked DCT maps.

3 METHODOLOGY

In this section, we first propose a novel random frequency mask module on the basis of discrete cosine transform (DCT) and Bernoulli distribution. Then we introduce an image classifier to detect adversarial samples according to the DCT pattern of the input image. Subsequently, we discuss how to incorporate the mask module and the classifier into an existing super-resolution backbone to build our proposed robust SR network. At the end of this section, we brief how to tune the overall network with the frequency mask modules and the adversarial classifier via an adversarial training strategy.

3.1 Random Frequency Mask

We develop a novel random frequency mask module to mitigate adversarial attacks, since most of adversarial perturbations are encoded as high-frequency components in a frequency domain. Figure 2 illustrates the architecture of the proposed mask module that consists of three steps: transforming an input to the DCT frequency domain, masking the DCT representations with a sampled map, and converting the masked DCT back to the image or feature space.

Our proposed mask module could take an image or a feature map as input. Without loss of generality, the module input is denoted as a $H \times W \times C$ tensor. Let $X \in \mathcal{R}^{H \times W}$ denote one of the C channel slices in the $H \times W \times C$ input tensor. We adopt discrete cosine transform

to compute the frequency representations of X. Let $\hat{X} \in \mathcal{R}^{H \times W}$ stand for the DCT result of X. $\hat{X}$ can be calculated as:

$$\hat{X}(u,v) = c(u)c(v) \sum_{i=0}^{H-1} \sum_{j=0}^{W-1} \left\{ X(i,j) \cos[(i+0.5)\pi/H \cdot u] \cos[(j+0.5)\pi/W \cdot v] \right. \quad (1)$$

where $c(u)$ is a compensation coefficient. $c(u)$ is set as $\sqrt{1/H}$ for $u = 0$ and $\sqrt{2/H}$ for $u \neq 0$, and the definition of $c(v)$ is the same as $c(u)$.

To understand DCT representations, we normalize and take average of different channels in the DCT results that are computed from a clean or attacked input. The averaged DCT maps are visualized in Figure 1 where the first row is for clean inputs and the second is for the attacked ones. In a DCT map, low-frequency coefficients are located nearby the upper-left corner. Figure 1 shows that the attacked DCT maps have larger high-frequency components, and smaller middle-frequency coefficients than the clean ones. We argue that adversarial attacks harm the original middle-frequency details to some degree, and add more high-frequency noises to the attacked image. Thus we propose to reduce the adversarial perturbations by multiplying the DCT $\hat{X}$ with a binary mask $\mathcal{M} \in \mathcal{R}^{H \times W}$: $\hat{X}_m = \hat{X} \odot \mathcal{M}$, where $\odot$ is element-wise multiplication. Then the masked DCT $\hat{X}_m$ is translated to the same image or feature space as the module input X via inverse discrete cosine transform. The whole process of our proposed mask module is formulated as: $X_m = \mathcal{F}^{-1}(\mathcal{M} \odot \mathcal{F}(X))$, where $\mathcal{F}(\cdot)$ stands for DCT, and $\mathcal{F}^{-1}$ denotes the Inverse-DCT. The module output X_m has the same shape as the input X.

We discuss how to determine the binary mask $\mathcal{M}$. Considering that the distance from the lowest-frequency component corresponds to the frequency degree of a component. For each coefficient of position (u, v), We set its corresponding weight in the binary mask according to its normalized distance from $(0, 0)$:

$$r_{(u,v)} = \sqrt{u^2 + v^2} / r_{max} \quad (2)$$

where r_{max} equals to $\sqrt{(H-1)^2 + (W-1)^2}$ and denotes the maximum radius for a DCT map of size $H \times W$. For a DCT component whose $r_{(u,v)}$ is smaller than a threshold r_t, the component probably contains the information of mean colors, smooth regions or sharp edges, which are image elements of low-to-middle frequency. To preserve the original image contents, we keep such a DCT coefficient unchanged by setting $\mathcal{M}(u, v)$ as 1. Since the boundary between middle-frequency details and high-frequency noises is uncertain, we uniformly sample r_t from $[r_l, r_u]$. r_l and r_u are manually set lower and upper bounds respectively by visualizing the difference of DCT maps between clean and attacked samples. If the $r_{(u,v)}$ value of a DCT coefficient is larger than the threshold, the coefficient might still encode normal contents such as highly repetitive textures. Since a higher-frequency coefficient more possibly contains adversarial noises, we adopt a Bernoulli distribution with the probability $p = r_{(u,v)}$ to decide if masking the current component. The distribution returns 1 with probability p and 0 with probability 1 - p. The binary mask $\mathcal{M}(u, v) = Bernoulli(p = r_{(u,v)})$ is formally defined as:

$$\mathcal{M}_{(u,v)} = \begin{cases} 1, & 0 \leq r_{(u,v)} \leq r_t \\ Bernoulli(p = r_{(u,v)}), & r_{(u,v)} > r_t \end{cases} \quad (3)$$

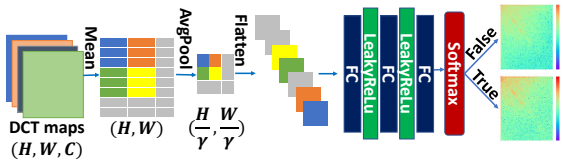

Figure 3: The architecture of our proposed adversarial sample classifier. The classifier takes DCT maps of an image as input, pools and reshapes them into a vector that is fed to a two-class classifier with three fully-connected layers. *True* denotes the input sample is attacked while *False* means a clean input. γ is a scaling factor.

The proposed mask module has two strengths. First, it supports backward propagation and could be placed at arbitrary positions of a SR network. Second, the mask module has no learnable parameters and is a lightweight module.

3.2 Adversarial Sample Classifier

Although we have developed a stochastic strategy to eliminate adversarial noises and preserve original contents at the same time, the proposed mask module still inevitably discards some fine details and degrades the performance on clean images. Thus we devise a two-class classifier shown in Figure 3 to predict if the input image is adversarial. The frequency mask modules are skipped for a clean input so that high-frequency elements of clean images are not affected. Considering the pattern difference between clean DCT maps and the attacked ones, we take DCT results as the classifier input.

For an input image $X \in \mathcal{R}^{H \times W \times C}$ that could be an adversarial sample or a clean one, we first compute its DCT of the same shape of $H \times W \times C$. Then we take average of different channel slices to obtain a single-channel DCT map of size $H \times W$, which corresponds to the visualization in Figure 1. The DCT map is further processed by a $\gamma \times \gamma$ 2D average pooling layer with stride 3 and padding 0. The shape of the pooled map becomes $H/\gamma \times W/\gamma$. Afterwards, we flatten the pooled map as a vector of size $1 \times HW/\gamma^2$ and feed the vector into 3 consecutive fully-connected (FC) layers. The parameters of these FC layers are denoted as $\theta = \{(\mathcal{W}_i, b_i), i = 1, 2, 3\}$. The overall classifier is denoted as a mapping function: $\mathcal{G} : X \rightarrow \mathcal{G}(X)$ and computed as:

$$\mathcal{G}(X) = \hat{\sigma}\left(\mathcal{W}_3 \sigma\left(\mathcal{W}_2 \sigma\left(\mathcal{W}_1 \phi(X) + b_1\right) + b_2\right) + b_3\right) \quad (4)$$

where $\phi(\cdot)$ stands for the combination of the channel averaging, the 2D spatial average pooling and the flattening operation. σ is an activation function LeakyReLU and $\hat{\sigma}$ is the Softmax function. $\mathcal{G}(X) \in \{False, True\}$, *True* indicates that the input image is an adversarial sample while *False* means a clean input. The prediction of classifier determines whether we conduct the subsequent mask operations. If *False*, we skip all the random mask modules. It is feasible because the input and the output of the mask module have the same shape, and are in the same image or feature space. Otherwise, we go through each mask module to mitigate the adversarial perturbations. Note that the input shape in the inference stage might be inconsistent with that in the training stage. To align

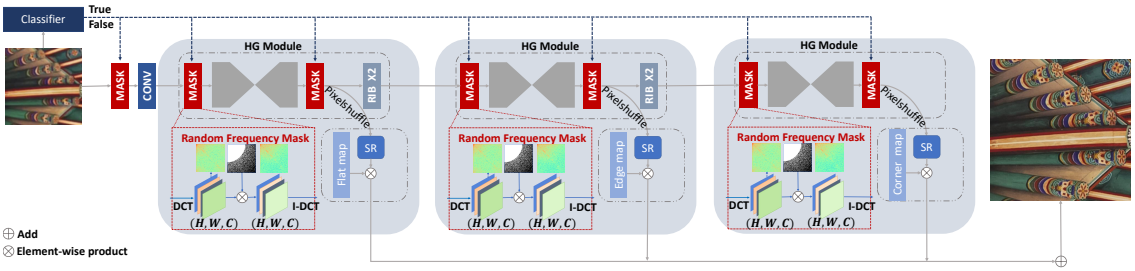

Figure 4: The overview framework of our proposed SR network. We equip a existing baseline CDC [54] with our proposed defenses. The proposed classifier and a mask module are located at the head of the network. Two mask modules are placed at the head and the tail of each hourglass (HG) module. Each HG module has a encoder-decoder architecture with skipped residual connections. Only if the classifier predicts the input as an adversarial sample, the proposed mask modules are activated.

Figure 5: Super-resolved results of our proposed method and existing SR methods on the RealSR dataset. The leftmost column shows the ground truth. For other columns, the left of the first row is a clean image, while the right image is the adversarial sample with the intensity $\alpha = 8/255$. The middle row is the super-resolved output of the clean images. The bottom row corresponds to the super-resolved results of the adversarial samples.

the inconsistent shapes, we simply insert a spatial adaptive pooling layer after the channel averaging step in the testing stage.

3.3 Network Architecture

In this section we first introduce a super-resolution baseline network, and then describe how to incorporate our proposed mask module and classifier into the baseline. Inspired by Newell *et al.* [42], we implement an image SR backbone by stacking six hourglass (HG) modules in a sequential manner. For simplicity, Figure 4 shows a network architecture with only three HG modules. The hourglass modules have the same architecture but do not share their weights. Two Residual Inception Blocks (RIB) are located in between each two adjacent HG modules. Following Wei *et al.* [54], we deploy three component-attentive blocks (CAB) at the end of the 2nd, 4th and 6th hourglass modules respectively. An input image is decomposed of three components: smooth regions, edges and corners. Each CAB only focuses on one of the three components by weighting its SR output with a 2D attentive map. The intermediate output of the three CABs are aggregated to yield the final super-resolved image.

In our proposed method, each hourglass module contains two random frequency mask modules: one at the beginning of the HG module, and the other before the residual blocks. Our proposed adversarial classifier is placed at the very beginning to predict if the input is adversarial. The classification result is sent to all mask modules and decides if passing through the random mask modules.

3.4 Adversarial Training

We employ a stagewise adversarial training strategy to tune our proposed SR network. First, we train the SR backbone network without the mask modules or the classifier. Second, we train the proposed classifier using clean images and adversarial samples that are synthesized with the basic attack [10] and the SR backbone trained at the first step. Third, we equip the SR backbone with the trained classifier and all the frequency mask modules to construct our complete SR network. The overall SR network is trained after freezing the classifier and randomly initializing the backbone. Following Wei *et al.* [54], We adopt two loss functions, an intermediate one and a gradient-weighted one. To achieve adversarial training, we maximize these loss functions to yield adversarial samples for training the overall proposed SR network. Note that the adversarial sample classifier is not attacked in the adversarial training.

Table 1: Comparison between our proposed method and existing super-resolution methods on clean samples and adversarial samples of different intensity. '0/255' denotes clean samples.

Method	Scale	0/255		1/255		2/255		4/255		6/255		8/255	
		PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM	PNSR	SSIM
CARN	×4	28.94	0.816	**28.72**	0.807	27.74	0.765	24.86	0.626	21.90	0.482	20.06	0.389
SRResNet		29.12	0.823	28.17	0.796	25.38	0.690	20.23	0.428	17.25	0.259	15.69	0.172
EDSR		29.22	0.825	28.47	0.805	26.18	0.722	21.09	0.458	18.47	0.298	16.77	0.201
RRDB		29.22	0.826	28.19	0.809	25.85	0.742	21.43	0.540	18.18	0.352	16.12	0.229
RCAN		**29.34**	0.827	28.43	0.808	25.43	0.699	20.26	0.450	17.88	0.314	16.25	0.223
LP-KPN		29.13	0.823	28.16	0.790	25.20	0.663	20.40	0.417	17.68	0.274	16.12	0.198
CDC		29.33	**0.828**	28.37	**0.813**	26.32	0.763	23.05	0.634	20.70	0.501	18.98	0.383
Ours		29.30	0.826	28.61	0.809	**27.94**	**0.785**	**27.65**	**0.771**	**27.27**	**0.762**	**26.88**	**0.753**

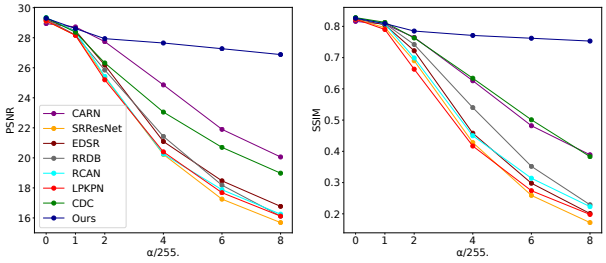

Figure 6: Comparison with the state-of-the-art SR methods against the attacks of different intensity α.

4 EXPERIMENTS

4.1 Implementation Details

We adopt the real-world SR dataset RealSR [5] for evaluation. RealSR contains 595 pairs of LR and HR images. These image pairs are captured by adjusting the focal length of digital cameras, and have been well aligned. Following Cai *et al.* [5], 495 pairs are selected for training and 100 pairs are used for testing. Their image sizes are in the range of [700, 3100] and [600, 3500]. 48×48 image patches are cropped for training SR models. We use the Adam optimizer, exponential decay rate of 0.9, batch size of 16, the initial learning rate of 2e-4. The learning rate is reduced by half after every 1e5 iterations. The maximum number of iterations is 4e5. For the proposed mask module, we set the lower and upper bounds $[r_l, r_u]$ as $[0.43, 0.5]$. For the proposed adversarial classifier, γ is set as 3. For adversarial training, we use the basic attack [10] with 2 iterations and the attack intensity as 6/255. For evaluation, we use the basic attack with 10 iterations, and the intensity $\alpha \in \{1, 2, 4, 6, 8\}/255$. Peak signal-to-noise ratio (PSNR) and structural similarity (SSIM) are used as the evaluation metrics. Different from Choi *et al.* [10], PSNR and SSIM are calculated between the HR ground truths and the super-resolved results of adversarial images.

4.2 Comparison with the State-of-the-art

We compare our proposed method with existing SR models based on deep CNNs, including SRResNet [34], EDSR [39], RRDB [53], RCAN [65], CARN [1], LP-KPN [5], CDC [54]. CARN is a model designed for lightweight image SR, LP-KPN and CDC are developed for real-world image SR. We train these methods on the RealSR dataset and evaluate them from two perspectives below.

Qualitative Evaluation Figure 5 visually compares the super-solved results of our proposed model with those of existing models. The upper row contains the cropped patches of input images, including a clean LR image on the left and the adversarial LR sample on the right. These adversarial samples are synthesized with the intensity $\alpha = 8/255$ and the iteration number $T = 10$. The middle row is the cropped super-resolved results of clean inputs, while the lower one is the outputs with adversarial samples. Figure 5 shows that our method is able to reconstruct sharp textures and maintain high-level clarity with clean samples. Under the adversarial attack, existing SR models usually produce unreasonable and structured artifacts, which severely affect the quality of super-resolved images. However, such highly repetitive artifacts are not seen in the results produced by our proposed method. This confirms to a certain extent that the proposed mask modules can effectively reduce high-frequency noises.

Quantitative Evaluation Table 1 shows the quantitative results in terms of PSNR & SSIM. Under the attack intensity of 8/255, the proposed network achieves the best value of PSNR and SSIM: 26.88 dB and 0.753 respectively. Compared with our baseline CDC [54], our proposed network obtains an improvement of 7.9 dB in PSNR and 0.37 in SSIM. For the clean images, the proposed model shows a small performance gap of 0.03 dB in PSNR and 0.002 in SSIM by comparing with the state-of-the-art CDC. Besides, it can be seen in Figure 6 that, as the intensity of adversarial perturbations increases, our proposed model presents smaller decreases than other state-of-the-art methods. Therefore, the above results suggest that our proposed network not only achieves more robust defense than existing real-world SR methods when resisting the white-box attack, but also maintains a competitive performance with clean images.

4.3 Comparison with Existing Defenses

We further verify the effectiveness of our proposed modules by comparing with other existing defenses, including image compression (JPEG) [15], image Random Resizing (RR) [56], a High-Frequency components Suppressing method (HFS) [66] and a Feature Denoising method (FD) [36, 57]. For the fairness, we adopt CDC [54] as the baseline that is combined with each of the above defenses respectively in this subsection. We place the JPEG Compression module, the Random Resizing and the High-Frequency Suppressing method in the front of the SR network, following their original setting. We insert four feature denoising blocks at the head of network, the tail

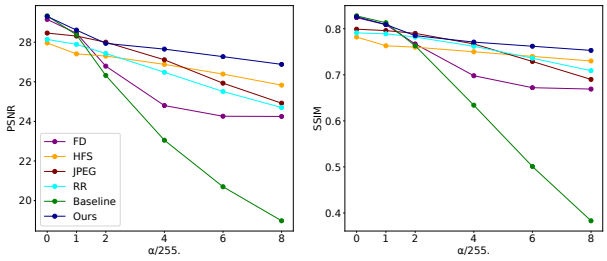

Figure 7: Comparison with existing defenses: Feature Denoising (FD), High-Frequency Suppressing (HFS), JPEG, Random Resizing (RR).

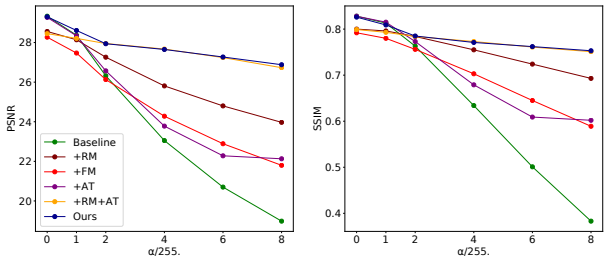

Figure 8: Ablation study on the proposed random frequency mask, the proposed adversarial sample classifier and the adversarial training of our method.

of the 2nd, 4th and 6th HG module. Adversarial training strategy is utilized for High-Frequency Suppressing and Feature Denoising, which corresponds to their most effective settings [57, 66]. Figure 7 shows the robustness of all the above defense methods with CDC. It could be observed that our proposed defense obtains the most significant performance at almost every adversarial intensity. With the highest intensity $\alpha = 8/255$, the proposed defense exceeds the second best method by 1.05 dB. The baseline and Feature Denoising could obtains approximate results with our proposed method on clean images. But these two models are far worse than our defense on adversarial samples.

4.4 Ablation Studies

Table 2: Numerical results of the ablation study on our proposed mask module and adversarial classifier in terms of PSNR and SSIM.

Method	Scale	0/255		4/255		8/255	
		PSNR	SSIM	PSNR	SSIM	PSNR	SSIM
Baseline	×4	**29.33**	**0.828**	23.05	0.634	18.98	0.383
+RM		28.56	0.800	25.81	0.755	23.97	0.693
+FM		28.27	0.792	24.28	0.703	21.80	0.589
+AT		29.27	0.828	23.78	0.679	22.13	0.602
+RM+AT		28.45	0.799	**27.67**	**0.773**	26.73	0.751
Ours		29.30	0.826	27.65	0.771	**26.88**	**0.753**

We study the effectiveness of the proposed random frequency mask module, the proposed adversarial sample classifier and the adversarial training in our model. Our proposed network is compared with the baseline (described in Section 3.3), the baseline with the Random frequency Mask (+RM), the baseline with the Fixed frequency Mask (+FM), the baseline with Adversarial Training (+AT), the baseline with Fixed frequency Mask and Adversarial Training (+FM+AT), and the one with Random frequency Mask and Adversarial Training (+RM+AT). The line charts of the above comparison with different attack strengths are plotted in Figure 8. The +RM setting outperforms the +FM on all values of the attack intensity, which means that the proposed random mask module is more robust than the one with a fixed pre-defined mask. Note that the +RM+AT model is equivalent to removing the proposed classifier from our finally proposed method (denoted as 'Ours' in Figure 8). By comparing the line charts of the +RM+AT model with our method, we find that our proposed model is superior to the +RM+AT model without classifier on clean images or low-intensity adversarial samples. It indicates that the proposed classifier could detect adversarial samples and avoid unnecessary masking to preserve the original image details.

Table 2 shows the numerical results of the ablation study. As the table displays, the +RM model surpasses the +FM model by 2.17 dB in PSNR and 0.104 in SSIM against the attack of intensity 8/255. It suggests that the proposed stochastic strategy in our random mask module is effective to improve the super-resolved results, in comparison to the fixed mask module. We verify that adversarial training could enhance the PSNR of the baseline by 3.15 dB with the adversarial samples of intensity 8/255. Besides, our proposed mask module could further boost the baseline with adversarial training by 4.6 dB in PSNR, by comparing the +RM+AT model with the +AT model. Figure 9 is the qualitative results of the ablation study, which shows that our proposed method successfully reduces the highly repetitive artifacts in the SR result of the baseline.

We investigate if our proposed defenses including the random frequency mask and the adversarial classifier are compatible with other super-resolution models. Three state-of-the-art SR networks (including LP-KPN, SRResNet and EDSR) are selected to combine with our proposed modules. For these SR networks, we incorporate a random frequency mask module at their head, and insert a mask module for every four residual blocks. The proposed classifier predicts if the input is adversarial, and determines whether to apply the mask modules. In Figure 10, the results of three SR networks are shown as solid lines, while the results of adding our proposed modules are dotted lines. As the figure shows, our proposed defenses significantly enhance the performance against adversarial attacks for the three SR networks. Even with the increase of the attack intensity, the networks with our method still present a stable performance. It indicates that our proposed modules form a general defense framework that has the potential to work with arbitrary SR neural networks.

5 CONCLUSION

In this paper, we first propose a random frequency mask module that improves the robustness of real-world image super-resolution models. The proposed mask module randomly erases high-frequency components in the discrete cosine transform domain, which is calculated from an image or a convolutional feature map. Considering that the frequency masking operation might be harmful for the original repetitive textures in an input image, we further develop an adversarial sample classifier that avoids unnecessary masking

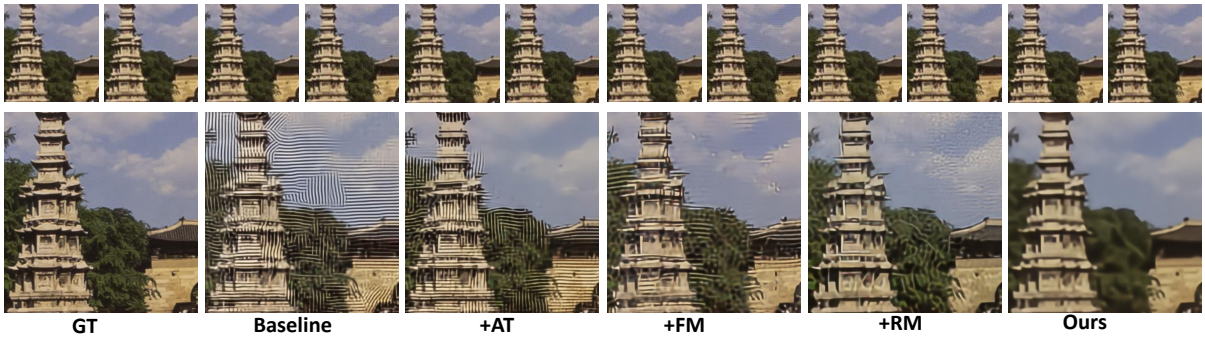

Figure 9: Super-resolved results of the ablation study on adversarial training and our proposed mask module. The first column is the ground truth while the others are for different variants. In each of these columns, the upper left is a clean image while the upper right is the adversarial sample of intensity $\alpha = 8/225$. The bottom row presents the super-resolved results of the adversarial images.

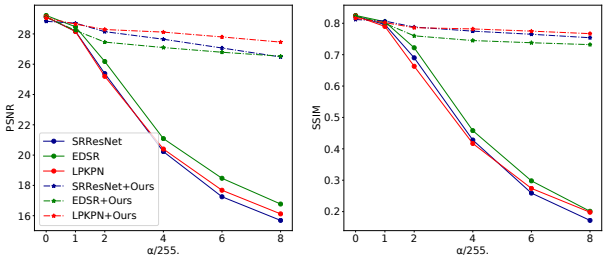

Figure 10: Comparison of robustness with and without our proposed modules in other super-resolution networks. The dotted lines denote results after combining our method.

the high-frequency details by detecting adversarial inputs. The proposed mask modules are only activated when the input image is very likely to be attacked. We experimentally show that our proposed method not only defends a wide range of existing SR networks against white-box attacks, but also maintains competitive performance with clean images. Overall, we introduce a general defense framework for real-world image super-resolution and the proposed framework may be extended to other robust applications in future.

ACKNOWLEDGMENTS

This work is supported by Key-Area Research and Development Program of Guangdong Province [2020B0101350001], National Natural Science Foundation of China under Grant No.61976250, No.U1811463 and No.62006253, and Guangzhou Science and technology project under Grant No.202102020633.

REFERENCES

[1] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. 2018. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 252–268.

[2] H. A. Aly and E. Dubois. 2005. Image up-sampling using total-variation regularization with a new observation model. *IEEE Transactions on Image Processing* 14, 10 (2005), 1647–1659. https://doi.org/10.1109/TIP.2005.851684

[3] Anish Athalye, Nicholas Carlini, and David Wagner. 2018. Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 274–283. http://proceedings.mlr.press/v80/athalye18a.html

[4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. 2012. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. (2012).

[5] Jianrui Cai, Hui Zeng, Hongwei Yong, Zisheng Cao, and Lei Zhang. 2019. Toward real-world single image super-resolution: A new benchmark and a new model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3086–3095.

[6] Chang Chen, Zhiwei Xiong, Xinmei Tian, Zheng-Jun Zha, and Feng Wu. 2019. Camera lens super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1652–1660.

[7] Hang Chen, Hongyan Zhang, Juan Du, and Bin Luo. 2020. Unified framework for the joint super-resolution and registration of multiangle multi/hyperspectral remote sensing images. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing* 13 (2020), 2369–2384.

[8] Peilin Chen, Wenhan Yang, Long Sun, and Shiqi Wang. 2020. When Bitstream Prior Meets Deep Prior: Compressed Video Super-Resolution with Learning from Decoding. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 1000–1008. https://doi.org/10.1145/3394171.3413504

[9] Zhen Chen, Xiaoqing Guo, Chen Yang, Bulat Ibragimov, and Yixuan Yuan. 2020. Joint Spatial-Wavelet Dual-Stream Network for Super-Resolution. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*. Springer, 184–193.

[10] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. 2019. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 303–311.

[11] Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee. 2020. Adversarially Robust Deep Image Super-Resolution using Entropy Regularization. In *Proceedings of the Asian Conference on Computer Vision (ACCV)*.

[12] T. Dai, J. Cai, Y. Zhang, S. Xia, and L. Zhang. 2019. Second-Order Attention Network for Single Image Super-Resolution. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 11057–11066. https://doi.org/10.1109/CVPR.2019.01132

[13] C. Dong, C. C. Loy, K. He, and X. Tang. 2016. Image Super-Resolution Using Deep Convolutional Networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 2 (2016), 295–307. https://doi.org/10.1109/TPAMI.2015.2439281

[14] Hao Dou, Chen Chen, Xiyuan Hu, Zuxing Xuan, Zhisen Hu, and Silong Peng. 2020. PCA-SRGAN: Incremental Orthogonal Projection Discrimination for Face Super-Resolution. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 1891–1899. https://doi.org/10.1145/3394171.3413590

[15] Gintare Karolina Dziugaite, Zoubin Ghahramani, and Daniel M Roy. 2016. A study of the effect of jpg compression on adversarial images. *arXiv preprint arXiv:1608.00853* (2016).

[16] N. Efrat, D. Glasner, A. Apartsin, B. Nadler, and A. Levin. 2013. Accurate Blur Models vs. Image Priors in Single Image Super-resolution. In *2013 IEEE International Conference on Computer Vision*. 2832–2839. https://doi.org/10.1109/ICCV.2013.352

[17] Raanan Fattal. 2007. Image Upsampling via Imposed Edge Statistics. In *ACM SIGGRAPH 2007 Papers* (San Diego, California) *(SIGGRAPH '07)*. Association for Computing Machinery, New York, NY, USA, 95–es. https://doi.org/10.1145/1275808.1276496

[18] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. 2015. Explaining and harnessing adversarial examples. In *ICLR*.

[19] J. Gu, H. Lu, W. Zuo, and C. Dong. 2019. Blind Super-Resolution With Iterative Kernel Correction. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1604–1613. https://doi.org/10.1109/CVPR.2019.00170

[20] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. 2018. Countering Adversarial Images using Input Transformations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SyJ7ClWCb

[21] Y. Guo, J. Chen, J. Wang, Q. Chen, J. Cao, Z. Deng, Y. Xu, and M. Tan. 2020. Closed-Loop Matters: Dual Regression Networks for Single Image Super-Resolution. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5406–5415. https://doi.org/10.1109/CVPR42600.2020.00545

[22] H. He and W. Siu. 2011. Single image super-resolution using Gaussian process regression. In *CVPR 2011*. 449–456. https://doi.org/10.1109/CVPR.2011.5995713

[23] Xiang He, Sibei Yang, Guanbin Li, Haofeng Li, Huiyou Chang, and Yizhou Yu. 2019. Non-Local Context Encoder: Robust Biomedical Image Segmentation against Adversarial Attacks. *Proceedings of the AAAI Conference on Artificial Intelligence* 33, 01, 8417–8424. https://doi.org/10.1609/aaai.v33i01.33018417

[24] Zhezhi He, Adnan Siraj Rakin, and Deliang Fan. 2019. 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 (CVPR)*.

[25] L. Huang and Y. Xia. 2021. Fast Blind Image Super Resolution Using Matrix-Variable Optimization. *IEEE Transactions on Circuits and Systems for Video Technology* 31, 3 (2021), 945–955. https://doi.org/10.1109/TCSVT.2020.2996592

[26] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. 2018. Black-box Adversarial Attacks with Limited Queries and Information. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 2137–2146.

[27] Michal Irani and Shmuel Peleg. 1991. Improving resolution by image registration. *CVGIP: Graphical Models and Image Processing* 53, 3 (1991), 231–239.

[28] Jian Sun, Zongben Xu, and Heung-Yeung Shum. 2008. Image super-resolution using gradient profile prior. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. 1–8. https://doi.org/10.1109/CVPR.2008.4587659

[29] Jing Jin, Junhui Hou, Jie Chen, Sam Kwong, and Jingyi Yu. 2020. Light Field Super-Resolution via Attention-Guided Fusion of Hybrid Lenses. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 193–201. https://doi.org/10.1145/3394171.3413585

[30] Jinggang Huang and D. Mumford. 1999. Statistics of natural images and models. In *Proceedings. 1999 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (Cat. No PR00149)*, Vol. 1. 541–547 Vol. 1. https://doi.org/10.1109/CVPR.1999.786990

[31] K. I. Kim and Y. Kwon. 2010. Single-Image Super-Resolution Using Sparse Regression and Natural Image Prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 6 (2010), 1127–1133. https://doi.org/10.1109/TPAMI.2010.25

[32] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. 2017. Adversarial machine learning at scale. In *ICLR*.

[33] W. Lai, J. Huang, N. Ahuja, and M. Yang. 2017. Deep Laplacian Pyramid Networks for Fast and Accurate Super-Resolution. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5835–5843. https://doi.org/10.1109/CVPR.2017.618

[34] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, et al. 2017. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4681–4690.

[35] Guanlin Li, Shuya Ding, Jun Luo, and Chang Liu. 2020. Enhancing Intrinsic Adversarial Robustness via Feature Pyramid Decoder. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[36] Haofeng Li, Guanbin Li, Binbin Yang, Guanqi Chen, Liang Lin, and Yizhou Yu. 2020. Depthwise nonlocal module for fast salient object detection using a single thread. *IEEE Transactions on Cybernetics* (2020).

[37] Haofeng Li, Guanbin Li, and Yizhou Yu. 2020. ROSA: Robust Salient Object Detection Against Adversarial Attacks. *IEEE Transactions on Cybernetics* 50, 11 (2020), 4835–4847. https://doi.org/10.1109/TCYB.2019.2914099

[38] Haofeng Li, Yirui Zeng, Guanbin Li, Liang Lin, and Yizhou Yu. 2020. Online Alternate Generator Against Adversarial Attacks. *IEEE Transactions on Image Processing* 29 (2020), 9305–9315. https://doi.org/10.1109/TIP.2020.3025404

[39] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. 2017. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 136–144.

[40] Qing Lyu, Hongming Shan, Cole Steber, Corbin Helis, Chris Whitlow, Michael Chan, and Ge Wang. 2020. Multi-contrast super-resolution mri through a progressive network. *IEEE transactions on medical imaging* 39, 9 (2020), 2738–2749.

[41] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. 2017. Universal adversarial perturbations. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1765–1773.

[42] Alejandro Newell, Kaiyu Yang, and Jia Deng. 2016. Stacked Hourglass Networks for Human Pose Estimation. In *14th European Conference on Computer Vision, ECCV 2016*. 483–499.

[43] K. S. Ni and T. Q. Nguyen. 2007. Image Superresolution Using Support Vector Regression. *IEEE Transactions on Image Processing* 16, 6 (2007), 1596–1610. https://doi.org/10.1109/TIP.2007.896644

[44] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. 2017. Practical Black-Box Attacks against Machine Learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security* (Abu Dhabi, United Arab Emirates) *(ASIA CCS '17)*. Association for Computing Machinery, New York, NY, USA, 506–519. https://doi.org/10.1145/3052973.3053009

[45] A. Prakash, N. Moran, S. Garber, A. DiLillo, and J. Storer. 2018. Deflecting Adversarial Attacks with Pixel Deflection. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 8571–8580. https://doi.org/10.1109/CVPR.2018.00894

[46] A. Rahmati, S. M. Moosavi-Dezfooli, P. Frossard, and H. Dai. 2020. GeoDA: A Geometric Framework for Black-Box Adversarial Attacks. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8443–8452. https://doi.org/10.1109/CVPR42600.2020.00847

[47] Ali Shafahi, Mahyar Najibi, Amin Ghiasi, Zheng Xu, John Dickerson, Christoph Studer, Larry S Davis, Gavin Taylor, and Tom Goldstein. 2019. Adversarial training for free!. In *Advances in Neural Information Processing Systems*, Vol. 32.

[48] Qi Shan, Zhaorong Li, Jiaya Jia, and Chi-Keung Tang. 2008. Fast Image/Video Upsampling. *ACM Trans. Graph.* 27, 5, Article 153 (Dec. 2008), 7 pages. https://doi.org/10.1145/1409060.1409106

[49] Xibin Song, Yuchao Dai, Dingfu Zhou, Liu Liu, Wei Li, Hongdong Li, and Ruigang Yang. 2020. Channel attention based iterative residual learning for depth map super-resolution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5631–5640.

[50] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2014. Intriguing properties of neural networks. In *ICLR*.

[51] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. 2017. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*. 114–125.

[52] Florian Tramer, Alexey Kurakin, Nicolas Papernot, Ian Goodfellow, Dan Boneh, and Patrick Mcdaniel. 2018. Ensemble Adversarial Training: Attacks and Defenses. In *ICLR*.

[53] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. 2018. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*. 0–0.

[54] Pengxu Wei, Ziwei Xie, Hannan Lu, Zongyuan Zhan, Qixiang Ye, Wangmeng Zuo, and Liang Lin. 2020. Component Divide-and-Conquer for Real-World Image Super-Resolution. In *European Conference on Computer Vision*. Springer, 101–117.

[55] Zeyu Xiao, Zhiwei Xiong, Xueyang Fu, Dong Liu, and Zheng-Jun Zha. 2020. Space-Time Video Super-Resolution Using Temporal Profiles. In *Proceedings of the 28th ACM International Conference on Multimedia* (Seattle, WA, USA) *(MM '20)*. Association for Computing Machinery, New York, NY, USA, 664–672. https://doi.org/10.1145/3394171.3413667

[56] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. 2018. Mitigating Adversarial Effects Through Randomization. In *International Conference on Learning Representations*.

[57] Cihang Xie, Yuxin Wu, Laurens van der Maaten, Alan L. Yuille, and Kaiming He. 2019. Feature Denoising for Improving Adversarial Robustness. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[58] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. 2014. Single-Image Super-Resolution: A Benchmark. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 372–386.

[59] J. Yang, J. Wright, T. S. Huang, and Y. Ma. 2010. Image Super-Resolution Via Sparse Representation. *IEEE Transactions on Image Processing* 19, 11 (2010), 2861–2873. https://doi.org/10.1109/TIP.2010.2050625

[60] Xinchen Ye, Baoli Sun, Zhihui Wang, Jingyu Yang, Rui Xu, Haojie Li, and Baopu Li. 2020. Pmbanet: Progressive multi-branch aggregation network for scene depth super-resolution. *IEEE Transactions on Image Processing* 29 (2020), 7427–7442.

[61] Haichao Zhang and Jianyu Wang. 2019. Towards Adversarially Robust Object Detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

[62] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang. 2017. Beyond a Gaussian Denoiser: Residual Learning of Deep CNN for Image Denoising. *IEEE Transactions on Image Processing* 26, 7 (2017), 3142–3155. https://doi.org/10.1109/TIP.2017.2662206

[63] Shu Zhang, Qiangqiang Yuan, Jie Li, Jing Sun, and Xuguo Zhang. 2020. Scene-adaptive remote sensing image super-resolution using a multiscale attention

network. *IEEE Transactions on Geoscience and Remote Sensing* 58, 7 (2020), 4764–4779.

[64] Xuaner Zhang, Qifeng Chen, Ren Ng, and Vladlen Koltun. 2019. Zoom to Learn, Learn to Zoom. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[65] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. 2018. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European conference on computer vision (ECCV)*. 286–301.

[66] Zhendong Zhang, Cheolkon Jung, and Xiaolong Liang. 2019. Adversarial Defense by Suppressing High-frequency Components.. In *IJCAI workshop*.

[67] H. Zheng, Z. Zhang, J. Gu, H. Lee, and A. Prakash. 2020. Efficient Adversarial Training With Transferable Adversarial Examples. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 1178–1187. https://doi.org/10.1109/CVPR42600.2020.00126

[68] Hong-Yu Zhou, Chengdi Wang, Haofeng Li, Gang Wang, Shu Zhang, Weimin Li, and Yizhou Yu. 2021. SSMD: Semi-Supervised Medical Image Detection with Adaptive Consistency and Heterogeneous Perturbation. *Medical Image Analysis* (2021), 102117.