Universal robustness via median randomized smoothing for real-world super resolution

Zakariya Chaouai, Mohamed Tamaazousti

Universal Robustness via Median Randomized Smoothing for Real-World Super-Resolution

Zakariya Chaouai
zakariya.chaouai@cea.fr

Mohamed Tamaazousti
mohamed.tamaazousti@cea.fr

Université Paris-Saclay, CEA, List, F-91120, Palaiseau, France

Abstract

Most of the recent literature on image Super-Resolution (SR) can be classified into two main approaches. The first one involves learning a corruption model tailored to a specific dataset, aiming to mimic the noise and corruption in low-resolution images, such as sensor noise. However, this approach is data-specific, tends to lack adaptability, and its accuracy diminishes when faced with unseen types of image corruptions. A second and more recent approach, referred to as Robust Super-Resolution (RSR), proposes to improve real-world SR by harnessing the generalization capabilities of a model by making it robust to adversarial attacks. To delve further into this second approach, our paper explores the universality of various methods for enhancing the robustness of deep learning SR models. In other words, we inquire: "Which robustness method exhibits the highest degree of adaptability when dealing with a wide range of adversarial attacks ?". Our extensive experimentation on both synthetic and real-world images empirically demonstrates that median randomized smoothing (MRS) is more general in terms of robustness compared to adversarial learning techniques, which tend to focus on specific types of attacks. Furthermore, as expected, we also illustrate that the proposed universal robust method enables the SR model to handle standard corruptions more effectively, such as blur and Gaussian noise, and notably, corruptions naturally present in real-world images. These results support the significance of shifting the paradigm in the development of real-world SR methods towards RSR, especially via MRS.

1. Introduction

The aim of single-image super-resolution (SISR) is to improve the resolution of a given low-resolution (LR) image, by producing a high-resolution (HR) image that is clear and without artifacts. SISR is widely used in a range of real-world applications, such as oceanography [9], surveillance [35], and medical images [12]. However, super-resolving an image poses a considerable challenge due to the ill-posed nature of the problem, since multiple HR solutions can correspond to a single LR image. There are several well-known methods for scaling high-resolution images, such as linear interpolation methods [17] or the estimation of covariance or correlation in LR data [2, 23]. Unfortunately, these methods often produce results that appear blurred, noisy and have difficulty in faithfully capturing high-frequency image details.

In recent years, SISR methods based on deep neural networks (DNNs) have made considerable progress [8, 21, 29, 33, 37] and offer much better quality for the upscaled image. Despite this progress, DNNs have been shown to be vulnerable to adversarial attacks, whether in classification [11, 27, 31] or in SR [6, 7] (see Figure 1). The inevitability and universality of adversarial examples is rooted in their definition. It is possible to systematically introduce additive perturbations into the input, causing the model to misclassify an example. The susceptibility to adversarial inputs poses a potential issue, hindering the application of deep learning methods in security and safety-critical contexts. It is important to note that even state-of-the-art SR models [24, 37] tend to perform poorly on real-world images that contain some corruption or amount of sensor noise. Since the majority of SR models are trained in a supervised way, requiring matching pairs of HR and LR images, LR images are typically generated from HR images by using bicubic downscaling.

The recognition of this constraint spurred the investigation of real-world SR on datasets with synthetic and natural corruptions. Several benchmarks [25, 26] design real-world artifacts and corruptions under different assumptions or from varying sensors. Consequently, some methods in real-world SR [10, 14] generate photo-realistic results only when they are evaluated on a specific dataset for which they were trained, but they fail to generalize to new datasets with unseen corruptions. A more recent approach, Castillo et al. [4], referred to as Robust Super-Resolution (RSR), proposes to improve real-world SR by harnessing the generalization capabilities of a model, making it robust to unseen noise by

using adversarial training, see Subsection 3.2. To the best of our knowledge, it is the only work that has attempted to create a generalized real-world SR model that achieves state-of-the-art results without training or fine-tuning on real-world datasets.

In this paper, we delve further into this latter approach. We recall that the adversarial learning employed in [4] relies on using the Projected Gradient Descent (PGD) attack 3.1 as a form of attack on LR images during the training phase. However, we will show that this type of defense is sensitive to other types of perturbations, and it is not the most effective generalized real-world SR model. In response to this limitation, we employ the Median Randomized Smoothing (MRS) approach, a scalable technique providing certified robustness for neural network-based models. This technique, initially applied in the context of object detection [5], transforms any DNN into a new smoothed one with certifiable l_2-norm robustness guarantees, as described in Lemma 4.1. The transformation is defined as follows: let $f_\theta : [0,1]^n \to [0,1]^m$, $f_\theta = (f_\theta^1, ..., f_\theta^m)$, be a SR neural network, and x be an input. Then, the median smoothing of f_θ is defined as $q_{0.5}(x) = (q_{0.5}^1(x), ..., q_{0.5}^m(x))$, where $q_{0.5}^i(x) = \inf\{y \in \mathbb{R} | \mathbb{P}(f_\theta^i(x+G) \leq y) \leq 0.5\}$ and $G \sim N(0, \sigma^2 I)$ follows a Gaussian distribution. The estimation of $q_{0.5}(x)$ can be approximated empirically through Monte Carlo (MC) sampling, as explained in [5]. The advantage of using the median on SR over the mean, commonly used in the classification field [28], stems from the fact that the median is nearly unaffected by outliers present in LR images. Unlike the median, the mean tends to smooth out the areas where predictions are locally constant, [5], which is disadvantageous for images as they often contain textures. Moreover, it is important to mention that the MRS method is known to require a large number of samples (of order 2000 [5]) with the MC procedure for classification and regression in object detection tasks. However, we discovered that in the context of SR, the MSR is well-suited because pixel-wise variations in predicted images are not large. We can easily control this instability with a few samples (of order 21). Finally, we will need to fine-tune the SR model on noisy LR images using different Gaussians samples to make it insensitive to this type of noise, as we are certifying our model with this type of noise, for our SR model to be insensitive to this type of noise.

Our main contributions are as follows:

1. We extend the use of adversarial attacks in SR. Until now, only the PGD attack presented in [4] has been applied in the context of real-world SR based on the perceptual loss. In this paper, we adapt other commonly used attacks from the classification literature. Specifically, we adapt the Fast Gradient Sign Method (FGSM), the Basic Iteration Method (BIM), and the Carlini and Wagner (CW) attack to the perceptual and pixel level of the image. We apply adversarial training using these attacks to create RSR models.
2. We propose a novel use of MRS to create a real-world SR model named CertSR that achieves state-of-the-art results, particularly for the Learned Perceptual Image Patch Similarity (LPIPS) metric.
3. Finally, we show that MRS is more universal in terms of robustness compared to all the previously mentioned adversarial training techniques.

2. Related works

It has been shown by Choi et al. [6] that state-of-the-art deep learning-based SR methods are highly susceptible to adversarial attacks. This vulnerability is primarily attributed to the propagation of the perturbation through the convolutional operation. In the SR domain, adversarial examples can be represented as follows: an original LR image x is perturbed by adding a small value δ to generate an adversarial LR image x_{adv}. Consequently, x_{adv} is slightly different from x. However, the prediction of x_{adv} deteriorates significantly compared to the prediction of x.

We note that adversarial attacks and robust models are applied to SR for the first time by Choi et al. [6, 7]. Notably, Choi et al. [6] explored target and non-targeted attacks, originally developed for classification tasks by Kurakin et al. [20]. They adapted these attacks to SR with the goal of maximizing the pixel degradation of super-resolved images. In [7], Choi et al. proposed a defense method formulated as an entropy regularization loss for model training, against the adversarial attacks constructed in [6], thus improving the robustness of the original SR model. However, as explained in [4], these last works focused on evaluating the methods based on pixel-wise metrics and did not concentrate their study on real-world SR.

It is worth mentioning that in the context of SR, the primary objective is to obtain perceptually well-resolved HR images. In pursuit of this objective, Castillo et al. [4] recently employed an adversarial attack based on pixel-wise and perceptual losses to construct a robust model. This type of attack was originally introduced by Madry et al. [27] for classification tasks. To the best of our knowledge, this work is the only one that reports the study of adversarial training for real-world SR problems, where the evaluation was done on perceptual metrics. In this study, we will show that our method performs much better, is more robust, and generalizes better to real-world SR problems, achieving state-of-the-art results without training or fine-tuning on corrupt datasets.

3. Adversarial attacks and training on SR

In this section, we present novel adversarial attacks tailored for SR tasks. It is noteworthy that these attacks are drawn

from the most relevant and widely employed techniques in the classification literature [3, 11, 20, 27]. The visual effect of these adversarial attacks is revealed in Figure 1. Subsequently, we will provide a general overview of adversarial learning, regardless of the specific adversarial attack used. These adversarial attacks, as well as the RSR based on these attacks, will be used in our experiments to assess the universality of the robustness of our certified SR approach. This evaluation encompasses various adversarial attacks, perturbations existing in the literature, and synthetic perturbations representative of those encountered in real-world images (as detailed in Section 5).

3.1. Adversarial attacks

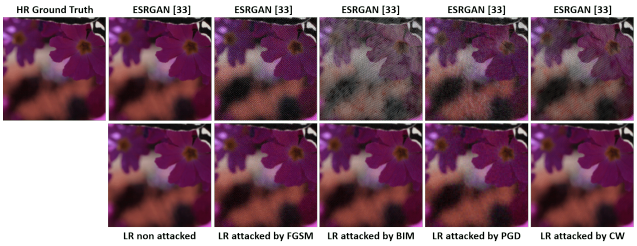

Figure 1. Visualization of both non-attacked and the corresponding attacked LR image subjected to various types of attacks, which we presented above, along with their predictions using ESRGAN [33], is provided in the first row. The top-left corner displays the ground truth image from the validation dataset of DIV2K [1], while the clean LR image is shown below it. The LR image was attacked using FGSM, BIM, and PGD with perturbations bounded within a ball of radius $\epsilon = 10/255$. For the CW attack, we utilized Adam [18] optimization to solve the problem in (2) with a learning rate of 10^{-2} for 6 iterations and $c = 0.01$.

Fast Gradient Sign Method (FGSM) is primarily designed to be a fast algorithm for generating adversarial LR images. Moreover, it is an attack that uses the gradient of the loss function to determine the direction in which pixel intensities should be changed to find the most efficient input perturbation. The adversarial LR image is mathematically calculated as follows:

$$x_{adv} = x + \epsilon \operatorname{sign}(\nabla_x \mathcal{L}(f_\theta(x), y)), \tag{1}$$

where $\mathcal{L}$ is composed of the L_{percep} perceptual and L_1 pixel-wise loss functions of the generator. Here, x represents the LR image, y represents the HR ground truth, and ϵ is the step size for the allowed perturbation. As ϵ increases, it becomes easier to degrade the network's predictions.

Basic Iterations Method (BIM) represents a simple refinement of the FGSM attack. Instead of taking a single step of size ϵ in the direction of the gradient sign, multiple smaller steps α are taken. Specifically, begin by setting $x_0 = x$ as a clean LR image used for initialization in iteration,

$$x_t = x_{t-1} + \alpha \operatorname{sign}(\nabla_{x_{t-1}} \mathcal{L}(f_\theta(x_{t-1}), y)).$$

Here, $\alpha = \frac{\epsilon}{T}$, where T represents the number of iterations. This approach is convenient because it provides extra control over the attack.

Projected Gradient Descent (PGD) [4] is considered as a generalization of the BIM attack that doesn't require the condition $\alpha = \frac{\epsilon}{T}$. Moreover, the initialization begins with perturbed LR images following a uniform distribution $U(-\epsilon, \epsilon)$. The perturbation is computed by taking multiple steps of gradient ascent with a small step size α and then projecting the perturbation onto the ϵ-ball around the input. Specifically, start by setting $x_0 = x + u$, where x is a clean LR image and $u \sim U(-\epsilon, \epsilon)$ is used for initialization in iteration,

$$x_t = \operatorname{clip}_{x,\epsilon}(x_{t-1} + \alpha \operatorname{sign}(\nabla_{x_{t-1}} \mathcal{L}(f_\theta(x_{t-1}), y))).$$

Here, $\operatorname{clip}_{x,\epsilon}$ denotes the clipping of the values of the adversarial sample so that they fall within an ϵ-neighborhood of the original sample x.

Carlini and Wagner attack (CW) is an optimization-based adversarial attack. In this attack, the perturbation is not constrained by the ϵ-ball in the infinite norm but aims to be minimal for the L_2 norm. The goal of this attack is to maximize the loss function by attacking images with the optimal perturbation. The optimization problem is given by:

$$\min_\delta (\|\delta\|_2 - c \cdot \mathcal{L}(f_\theta(x), y)), \text{ such that } x + \delta \in [0,1]^n, \tag{2}$$

where c is a hyperparameter. To ensure that $x + \delta \in [0,1]^n$, which means that $x + \delta$ yields a valid image, it introduces a new variable w to substitute as follows

$$\delta = \frac{1}{2}(\tanh(w) + 1) - x.$$

3.2. Adversarial training

Roughly speaking, adversarial training consists of using adversarial examples generated from the training data set to increase robustness locally around the training samples. In this paper, in addition to our main method, which will be presented in Section 4, we will employ this technique to create robust models for comparison.

Adversarial learning typically takes the form of a robust min-max optimization problem, that is given as follows,

$$\theta^*_{\text{adv}} = \operatorname{argmin}_{\theta \in \Theta} \frac{1}{N} \sum_{(x^{(i)}, y^{(i)}) \in \mathcal{D}} \max_{\|\delta\|_2 \leq \epsilon} \mathcal{L}(f_\theta(x^{(i)} + \delta), y^{(i)}),$$

where $\mathcal{D}$ is a batch of LR and HR images. The training is usually processed using an optimization algorithm based on gradient descent on mini-batches. It is important to note that at each iteration of the optimization process, the DNN parameters are updated, and it is necessary to compute the adversarial perturbations with respect to these new parameters at each iteration. This step requires a huge additional computation time compared to classical learning.

4. The Main Method

4.1. Median Randomized Smoothing (MRS)

The MRS is a scalable approach to obtain certified robustness guarantees for any super-resolution neural network. The main principle of this method is to create from one LR image a sample of images by adding Gaussian noise with a certain standard deviation. Then, we get the median of all the predictions pixel-by-pixel. Consequently, we obtain a smoothed model that is certified in an interval of percentiles depending on the perturbation that exists in the input image of the model. More precisely, let $G \sim N(0, \sigma^2 I)$, a Gaussian random variable. The percentile smoothing of a DNN $g_\theta : \mathbb{R}^n \to \mathbb{R}$ is defined as follows

$$\overline{q}_p(x) = \inf\{y \in \mathbb{R} | \mathbb{P}(g_\theta(x+G) \leq y) \geq p\},$$

$$\underline{q}_p(x) = \sup\{y \in \mathbb{R} | \mathbb{P}(g_\theta(x+G) \leq y) \leq p\}.$$

We denote $q_p(x)$ as the percentile-smoothed function when either definition is applicable. When $p = 0.5$ these percentiles are equivalent to the median $q_{0.5}(x)$. Therefore, from [5] we have the following Lemma:

Lemma 4.1 *A percentile-smoothed function q_p with adversarial perturbation δ can be bounded as follows*

$$\underline{q}_{\underline{p}}(x) \leq q_p(x+\delta) \leq \overline{q}_{\overline{p}}(x), \quad \forall \|\delta\|_2 < \epsilon \tag{3}$$

such that $\overline{p} = \Phi(\Phi^{-1}(p) + \frac{\epsilon}{\sigma})$ and $\underline{p} = \Phi(\Phi^{-1}(p) - \frac{\epsilon}{\sigma})$, where Φ is the standard Gaussian CDF.

Here, we are interested in the case $p = 0.5$. In this case, the median is bounded between the percentile of $\overline{p} = \Phi(\frac{\epsilon}{\sigma})$ and $\overline{p} = \Phi(-\frac{\epsilon}{\sigma})$. On the one hand, we observe from Lemma 4.1 that a smaller distance between $\underline{q}_{\underline{p}}(x)$ and $\overline{q}_{\overline{p}}(x)$ indicates a more robust and well-certified model. On the other hand, the bounds of the interval depend on the value of $\frac{\epsilon}{\sigma}$ where ϵ represents the size of the perturbation against which we aim to certify. Therefore, the choice of σ depends on the adversarial attack and the perturbation that exists on LR images. Fortunately, at the inference phase, there is some flexibility in choosing the standard deviation of the Gaussian noise, σ, that will help us to get a robust and certified SR model.

4.2. Median Randomized Smoothing for SR

To create our RSR model which we call CertSR (Certified Super-Resolution) model, we need to go through three essential steps. First, we implement an initial SR model based on a Generative Adversarial Network (GAN) previously trained on clean LR images. Second, we need to fine-tune the SR model on noisy LR images using samples of i.i.d. Gaussians with a specified number of draws and standard deviations. This type of data augmentation will make the SR model more robust to noisy samples. We call this second step $MRS_{Fine-tuning}$. Finally, in a third step that we call $MRS_{Inference}$ phase, we use the median random smoothing method to certify the fine-tuned SR model with a sample of i.i.d. Gaussians associated to a standard deviation (see Figure 2).

Super-Resolution Model This study is based on the ESRGAN model [33], which is a generative adversarial network (GAN) used for super-resolving images. The generator adopts the Residual-in-Residual Dense Block (RRDB) [22] structure to improve the quality of the enhanced image. The resolution of the generated images will be enlarged by a factor of 4. We recall that several loss functions are applied during the training. Firstly, the L_1 loss is used to evaluate the pixel distance between the ground truth (GT) and the super-resolved image. Secondly, the perceptual loss L_{perc} [15] utilizes the activation features of the pre-trained VGG-19 [30] between the GT and the super-resolved image. This loss helps enhance the visual effect of low-frequency components. The third loss is the adversarial loss L_{adv}, employed to enhance the texture details of the super-resolved image and make it more realistic. The total loss function is the sum of these three losses:

$$L_{total} = L_1 + L_{perc} + L_{adv}.$$

The Discriminator is structured on a VGG-128 architecture [30] and operates under the same principle as the Relativistic GAN [16]. It estimates the probability that a real image appears more realistic than a fake one.

CertSR We use the pre-trained network generator of the ESRGAN model [33]. Subsequently, we propose to fine-tune this model, denoted $MRS_{fine-tuning}$, on LR images by adding samples of Gaussians noise. We use different standard deviations and for each of them, we choose the same amount of draws[1]. Then, we calculate the median of predictions associated with each standard deviation, following the procedure outlined in the fine-tuning phase of Figure 2. Finally, in the inference phase, denoted $MRS_{inference}$, we use the MRS to certify the SR model with a specific

[1]Note that in our experiments we observed that it is suitable to also use the original image (without adding Gaussian noise).

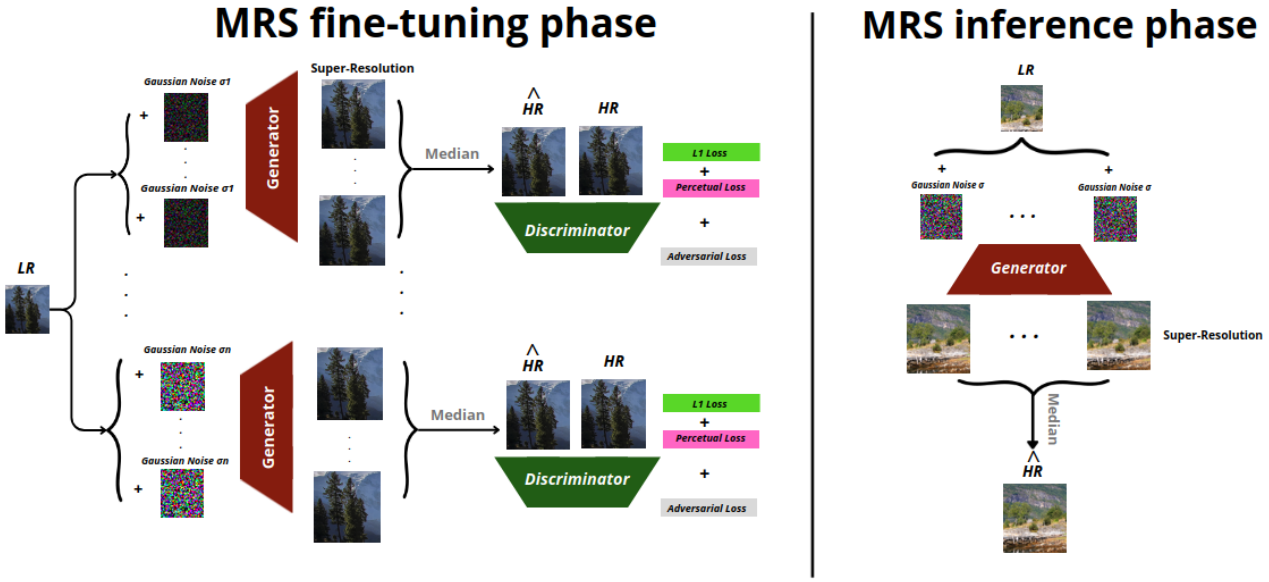

Figure 2. Framework of our proposed CertSR method. In the training part, we add different samples of i.i.d. Gaussians with different standard deviations to the same LR image. We then calculate the median of predictions associated with each standard deviation. In the test part, we use MRS to certify our generator by adding sample i.i.d. Gaussians with the same standard deviation.

standard deviation, which is a hyperparameter that must be selected to best suit each perturbation, as shown on the right of Figure 2. We emphasize that thanks to the small invariance of the pixel-wise loss on the super-resolved images, at this stage we draw only 21 Gaussian samples in all our experiments to certify our model, which allows us to control this invariability. Moreover, we rely on the LPIPS metric in this context to ensure that we have chosen the best standard deviation.

5. Experimental Results

In this section, we describe the experimental settings, including the utilized datasets and model configurations.

5.1. Evaluation Metrics

We evaluate the performance of different methods by calculating metrics such as Peak-Signal-to-Noise Ratio (PSNR), [34], Structural Similarity Index Measure (SSIM), [38], and Learned Perceptual Image Patch Similarity (LPIPS), [36]. PSNR and SSIM are widely used to evaluate image restoration and focus primarily on image fidelity rather than visual quality. LPIPS, on the other hand, places greater emphasis on assessing the similarity of visual features between images. To do this, it uses a pre-trained AlexNet [19] to extract image features, then calculates the distance between these features. As a result, a lower LPIPS value indicates a closer resemblance between GT and the generated image.

5.2. Dataset

Fine-tuning dataset We fine-tune the SR models on the DIV2K dataset [1, 32] which is a reference commonly used in traditional SISR. Its training set consists of 800 2K resolution images and their respective LR versions, generated by a bicubic downscaling process. These images incorporate no artificial perturbation. We crop the images into 480 × 480 sub-images for our experiments. A scaling factor of 4 was used between the HR images and the 120 × 120 LR images.

Inference dataset We assess the performance of our CertSR method on both the clean and the corrupted DIV2K validation dataset [1, 32], which contains 100 validation images. Specifically, we corrupt the validation dataset with sensor noise, which is simulated by adding pixel-wise independent Gaussian noise with a mean of 0 and a standard deviation of 0.03. We also corrupt this dataset by degrading LR images into blurry images. This operation is modeled by smoothing the images with the Gaussian kernel with 10 in size and a standard deviation of 0.3. Subsequently, we attack the inference dataset with the adversarial attacks defined in section 3.

It is also crucial to evaluate our main method on real-world datasets containing various types of synthetic corruptions and sensor noise in LR images. Specifically, we evaluate our method using validation datasets from the NTIRE 2020 Real-World Image Super-Resolution Challenge, Track

1 [26], and the AIM 2019 Real World Super-Resolution Challenge, Track 2 [25]. The validation sets comprise artificially degraded versions of the 100 LR images in the DIV2K validation set, together with their corresponding GT. For simplicity, we abbreviate NTIRE 2020 and AIM 2019 as NTIRE and AIM, respectively.

5.3. Implementation details

Fine-tuning is based on the pre-trained ESRGAN [33]. We perform all the fine-tuning methods that we need on a node composed of 8 GPU A100 80Gb with 1.5 Terabytes of RAM and dual AMD processors. We use an Adam optimizer [18] with $\beta_1 = 0.9$ and $\beta_2 = 0.99$ for both the generator and discriminator with an initial learning rate of 10^{-4}. For the classical fine-tuning of ESRGAN, as well as for adversarial fine-tuning, we choose 18k iterations and 16 images per batch. Regarding the hyperparameters for adversarial learning, the choices are as follows: (i) Adversarial Learning with FGSM (AD-L-FGSM) has $\epsilon = 9/255$. (ii) Adversarial Learning with BIM (AD-L-BIM) uses the same ϵ as AD-L-FGSM with 2 iterations. (iii) Adversarial Learning with CW (AD-L-CW) employs $c = 10^{-2}$, 4 iterations, and utilizes Adam optimization for resolving 2 with a learning rate of 10^{-2}. (iv) Adversarial Learning with PGD (AD-L-PGD) uses the pre-trained model from [4]. Subsequently, for the $MRS_{Fine-tuning}$ step we take 59k as a number of iterations with 5 images per batch. During this phase, we duplicate the batch training set five times. For the first two batches, we add i.i.d. Gaussian samples with a standard deviation of $\sigma = 0.03$. For the next two batches, we add i.i.d. Gaussian samples with a standard deviation of $\sigma = 0.2$. The last remaining batch remains unchanged to ensure CertSR considers cleaned images as well. For more details on the hyperparameters of adversarial learning and the $MRS_{Fine-tuning}$ step, please refer to the supplementary material.

Comparison with State-of-the-Art We compare our main method CertSR[2] with other state-of-the-art methods to establish a universal robust baseline for SISR models. For this, we evaluate our results on both clean and corrupted images. We compare our results with ESRGAN [33], and AD-L-PGD [4]. To ensure a fair comparison, we fine-tune ESRGAN on the DIV2K training set. For real-world images, we also compare our results with the top-performing models on the NTIRE and AIM datasets: Impressionism [14] and ESRGAN-FS [10], respectively. We use pre-trained weights for Impressionism on NTIRE and DPED [13] datasets and for ESRGAN-FS on AIM and DPED datasets. Moreover, we fine-tune Impressionism on AIM and ESRGAN-FS on NTIRE, by employing default parameters from their works.

[2]See supplementary material for the ablation study.

Data	Clean			Noisy			Blurry		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ESRGAN [33]	27.48	0.75	0.12	20.25	0.29	0.67	22.23	0.62	0.48
AD-L-PGD [4]	26.60	0.71	0.22	22.63	0.47	0.37	22.15	0.60	0.50
AD-L-FGSM (ours)	26.28	0.70	0.34	24.84	0.57	0.32	21.95	0.59	0.53
AD-L-BIM (ours)	26.21	0.68	0.25	25.11	0.60	0.29	21.93	0.58	0.48
AD-L-CW (ours)	28.41	0.77	0.14	19.47	0.25	0.78	22.34	0.62	0.50
CertSR (ours)	28.24	0.76	0.12	26.35	0.70	0.19	22.11	0.60	0.44

Table 1. This table reports the quantitative results of robust and non-robust methods for clean, sensor noise (noisy), and blurry DIV2K validation dataset. In all the tables of this document, the arrows indicate if high ↑ or low ↓ values are desired. The best scores are displayed in Red and the second in Blue.

5.4. Evaluation on Clean and Corrupted Images

In Table 1, we present a comparison of PSNR, SSIM, and LPIPS values for our CertSR method, the non-robust SR model, ESRGAN , and various RSR models. In the quantitative experiments, we focus on the LPIPS measure, as it has the best correlation with image similarity. We see from Table 1 that our CertSR method performs well on all three inference datasets. It is important to note that on the clean and noisy dataset, we do not need to use $MRS_{inference}$, using only the $MRS_{fine-tuning}$ we achieve the same results. Furthermore, since the $MRS_{fine-tuning}$ includes both clean and noisy data simultaneously. We obtained a LPIPS value that is almost the same as that of ESRGAN. However, the LPIPS metric value of the ESRGAN model on the noisy dataset is the lowest. Concerning the blurry case, we use the $MRS_{inference}$ on this validation dataset with $\sigma = 0.05$. Moreover, we observe that the performance of our CertSR method surpasses that of all other RSR methods. Regarding the other robust models, we can see that AD-L-CW is the best RSR on the clean validation dataset, while AD-L-BIM performs better on the noisy and blurry datasets. Finally, we note that AD-L-FGSM performs better on noisy images than on clean images, which is attributed to the training conducted on attacked images.

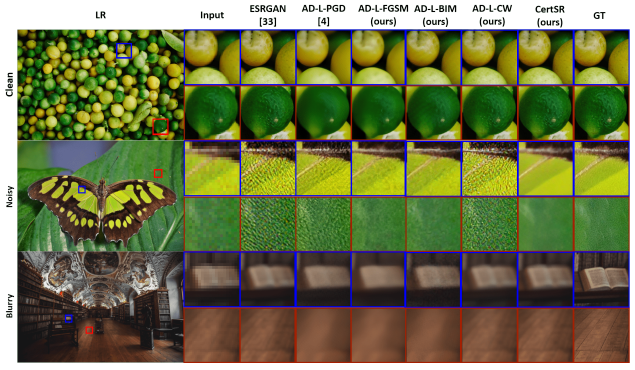

Figure 3. This figure presents the qualitative results of robust and non-robust methods for clean, sensor noise (noisy), and blurry DIV2K validation dataset.

Figure 3 represents the qualitative results of robust and non-robust methods with respect to the clean, sensor noise, and blurry DIV2K validation dataset. Our CertSR method provides clearer images with richer texture detail and without artifacts, showing that our method is the most robust against noisy and blurry perturbations. On the other hand, we observe that AD-L-PDG and AD-L-FGSM generate very smooth images, and AD-L-BIM introduces some little artifacts in the case where LR images are clean.

Adversarial attacks	FGSM			BIM			PGD			CW		
Method	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓	PSNR↑	SSIM↑	LPIPS↓
ESRGAN [33]	16.70	0.18	0.70	14.97	0.15	0.76	17.83	0.19	0.83	16.43	0.23	0.69
AD-L-PGD [4]	21.74	0.50	0.36	19.45	0.45	0.44	24.21	0.60	0.24	25.15	0.67	0.24
AD-L-FGSM (ours)	25.55	0.70	0.19	23.48	0.60	0.29	21.56	0.39	0.46	24.13	0.64	0.32
AD-L-BIM (ours)	24.17	0.60	0.27	23.79	0.59	0.26	24.65	0.59	0.33	25.57	0.65	0.25
AD-L-CW (ours)	4.72	0.23	0.99	12.83	0.09	0.91	15.39	0.13	0.95	18.37	0.33	0.61
CertSR (ours)	24.72	0.64	0.27	24.28	0.64	0.25	25.09	0.67	0.24	26.66	0.72	0.18

Table 2. This table shows the quantitative results concerning robust and non-robust methods against the most relevant adversarial attacks. The best scores are displayed in Red and in Blue.

Table 2 presents the quantitative results of the robust and non-robust methods against the adversarial attacks. To study this, we place ourselves in the worst-case scenario, which means we test the universality of our CertSR's robustness against the same attacks that were used to build RSR models. It is important to mention that in the validation part, we use $MRS_{inference}$ against each adversarial attack with respect to different standard deviations. More precisely, against PGD (see 3.1) and FGSM (see 3.1) attacks, we certify our model with $\sigma = 0.06$. Against the BIM attack (see 3.1), we choose $\sigma = 0.07$, and against the CW attack (see 3.1), we use $\sigma = 0.03$ (please consult the supplementary material to see how these hyperparameters have been selected). Therefore, we see from Table 2 that our main method achieves the best performance against all adversarial attacks with respect to PSNR, SSIM and LPIPS metrics, except against ADV-L-FGSM, where CertSR is the second-best method against FGSM attacks. Therefore, we can say that CertSR is the most globally robust SR method against adversarial attacks.

In Figure 4, we present qualitative results concerning CertSR's robustness against the most relevant adversarial attacks. Visually, it is clear that CertSR produces super-resolved images that are superior to those of other RSR models. The images generated by these RSR models show noticeable artifacts. This figure illustrates that even models trained with a specific adversarial attack remain somewhat vulnerable when subjected to a similar attack. We observe that the weakest robust SR model is AD-L-CW. This is related to the fact that even CW attack has the advantage of being the optimal and strongest attack, it also has the disadvantage of being the most difficult to learn.

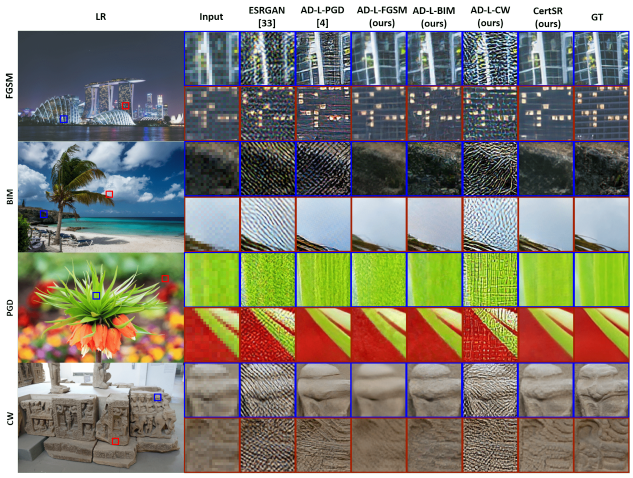

Figure 4. This figure provides qualitative results concerning robust and non-robust methods against the most relevant adversarial attacks.

5.5. Evaluation on Real-World Images

Table 3 presents the quantitative results of reference metrics for CertSR method, state-of-the-art methods and RSR models on both the NTIRE and AIM validation datasets. We observe that CertSR achieves the best LPIPS performance without any training or fine-tuning on these datasets. AD-L-CW and ESRGAN achieve the worst LPIPS on both validation datasets. We also observe that AD-L-BIM is more performant than AD-L-PGD on the AIM. These results are visually confirmed in Figure 5. For the $MRS_{inference}$ phase, we choose $\sigma = 0.03$ and $\sigma = 0.06$ for NTIRE and AIM respectively. Please refer to the supplementary material to see how these hyperparameters have been selected.

It is important to note that, we also test the proposed CertSR method on other SR models besides ESRGAN, on both NTIRE and AIM validation datasets, to demonstrate that the method can enhance the accuracy and robustness of other initial SR models. See supplementary material for more details.

6. Conclusion

In this work, we explore the fruitful relationship between Robust Super-Resolution (RSR) and real-world SR. Our main finding is the demonstration that the most universal model in terms of robustness to different adversarial attacks is also the more robust to unseen natural noise in the LR input real-world images. This important insight is based on a study conducted on two different types of RSR models: one type built from various adversarial training techniques (including the existing RSR model using PGD attack [4] and new RSR models that we built from FGSM, BIM and the CW attacks) and another original one built from a certifi-

Method	Training Data	Fine-tuning Data	PSNR↑			SSIM↑			LPIPS↓		
			NTIRE	AIM	Avg	NTIRE	AIM	Avg	NTIRE	AIM	Avg
Bicubic			25.51	22.35	23.93	0.67	0.62	0.65	0.63	0.68	0.66
ESRGAN-FS [10]	Flickr2K	NTIRE	24.59	22.07	23.33	0.69	0.63	0.66	0.25	0.47	0.36
		AIM	19.56	20.82	20.19	0.31	0.51	0.41	0.56	0.39	0.48
		DPEP	17.79	20.15	18.97	0.34	0.53	0.43	0.51	0.47	0.49
Impressionism [14]	Flickr2K	NTIRE	24.82	21.47	23.15	0.66	0.54	0.60	0.23	0.52	0.37
		AIM	19.65	21.89	20.77	0.29	0.60	0.45	0.67	0.41	0.54
		DPEP	17.53	18.84	18.18	0.34	0.49	0.41	0.60	0.47	0.53
ESRGAN [33]	Flickr2K	DIV2k	21.94	21.95	21.03	0.39	0.55	0.49	0.56	0.51	0.53
AD-L-PGD [4]	Flickr2K	DIV2K	24.31	21.99	23.15	0.65	0.60	0.62	0.23	0.37	0.30
AD-L-FGSM (ours)	Flickr2K	DIV2k	25.55	22.70	24.20	0.65	0.63	0.64	0.30	0.42	0.36
AD-L-BIM (ours)	Flickr2K	DIV2K	25.35	22.31	23.95	0.63	0.59	0.61	0.26	0.36	0.31
AD-L-CW (ours)	Flickr2K	DIV2K	21.25	21.86	21.63	0.37	0.58	0.48	0.63	0.47	0.55
CertSR (ours)	Flickr2K	DIV2K	26.67	21.75	24.21	0.71	0.59	0.65	0.21	0.33	0.27

Table 3. **Quantitative results on Real-World Images.** We present the quantitative results of reference metrics between our method, state-of-the-art methods, and robust and non-robust models on NTIRE and AIM validation datasets. Red and Blue colors highlight the best two scores. **Bold** represents the best method for LPIPS metric for both datasets.

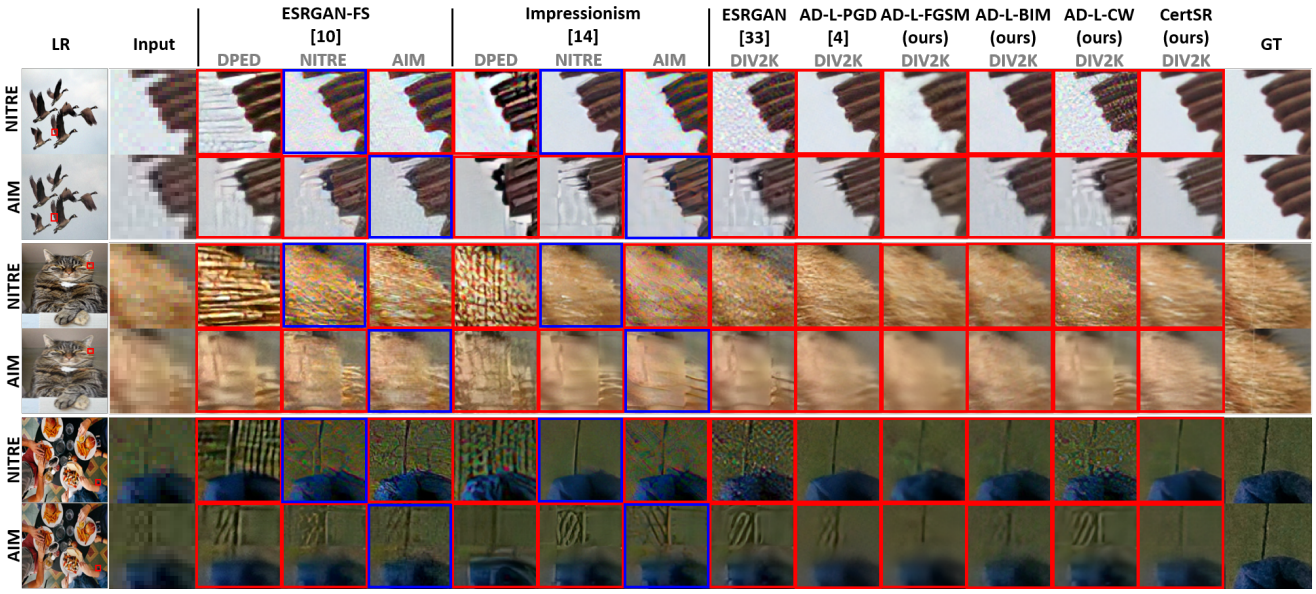

Figure 5. **Qualitative results on Real-World Images.** Comparison between the proposed methods including CertSR and state-of-the-art RSR method (AD-L-PGD [4]), for two corruption datasets: NTIRE and AIM. For reference, we show the input, the results of ESRGAN-FS method [10], Impressionism method [14] and the ground-truth (GT). Blue frames denote training and validation on the same dataset. Red frames denote training and validation on different datasets. The training dataset is indicated in gray just below the name of the methods.

cation technique that leverages MRS procedure with Gaussian noise. Our experiments on synthetic and real datasets show that, compared to the RSR models AD-L-PGD [4] AD-L-FGSM, AD-L-BIM, AD-L-CW, the proposed model CertSR, is the most universal in terms of robustness to adversarial attacks and is also the one that achieves the best results on real-world SR. We also show that the CertSR achieved state-of-the-art results in particular with the LPIPS metric. We expect that this finding will encourage further study of the RSR approach to tackle noise in real-world SR.

Acknowledgements This publication was made possible by the use of the FactoryIA supercomputer, financially supported by the Ile-de-France Regional Council. The authors thank Patrick Hede for his technical support in using FactoryIA.

References

[1] Eirikur Agustsson and Radu Timofte. Ntire 2017 challenge on single image super-resolution: Dataset and study. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 126–135, 2017. 3, 5

[2] Jan Allebach and Ping Wah Wong. Edge-directed interpolation. In *Proceedings of 3rd IEEE International Conference on Image Processing*, pages 707–710, 1996. 1

[3] Nicholas Carlini and David Wagner. Towards evaluating the robustness of neural networks. In *IEEE symposium on security and privacy*, pages 39–57, 2017. 3

[4] Angela Castillo, Juan Escobar, María C. Pérez, Andrés Romero, Radu Timofte, Luc Van Gool, and Pablo Arbelaez. Generalized real-world super-resolution through adversarial robustness. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1855–1865, 2021. 1, 2, 3, 6, 7, 8

[5] Ping-yeh Chiang, Michael Curry, Ahmed Abdelkader, Aounon Kumar, John Dickerson, and Tom Goldstein. Detection as regression: Certified object detection with median smoothing. In *Advances in Neural Information Processing Systems 33*, pages 1275–1286, 2020. 2, 4

[6] Jun-Ho Choi, Huan Zhang, Cho-Jui Kim, Jun-Hyuk Hsieh, and Jong-Seok Lee. Evaluating robustness of deep image super-resolution against adversarial attacks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 303–311, 2019. 1, 2

[7] Jun-Ho Choi, Huan Zhang, Cho-Jui Kim, Jun-Hyuk Hsieh, and Jong-Seok Lee. Adversarially robust deep image super-resolution using entropy regularization. In *Proceedings of the the Asian Conference on Computer Vision*, 2020. 1, 2

[8] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *European Conference on Computer Vision (ECCV)*, pages 184–199, 2014. 1

[9] Aurelien Ducournau and Ronan Fablet. Deep learning for ocean remote sensing: an application of convolutional neural networks for super-resolution on satellite-derived sst data. In *9th IAPR Workshop on Pattern Recogniton in Remote Sensing (PRRS). IEEE*, pages 1–16, 2016. 1

[10] Manuel Fritsche, Shuhang Gu, and Radu Timofte. Frequency separation for real-world super-resolution. In *IEEE/CVF International Conference on Computer Vision Workshop*, 2019. 1, 6, 8

[11] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014. 1, 3

[12] Yawen Huang, Ling Shao, and Alejandro F Frangi. Simultaneous super-resolution and cross-modality synthesis of 3d medical images using weakly-supervised joint convolutional sparse coding. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 6070–6079, 2017. 1

[13] Andrey Ignatov, Nikolay Kobyshev, Radu Timofte, Kenneth Vanhoey, and Luc Van Gool. Dslr-quality photos on mobile devices with deep convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 3277–3285, 2017. 6

[14] Xiaozhong Ji, Yun Cao, Ying Tai, Chengjie Wang, Jilin Li, and Feiyue Huang. Real-world super-resolution via kernel estimation and noise injection. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2020. 1, 6, 8

[15] 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, 2016. 4

[16] Alexia Jolicoeur-Martineau. The relativistic discriminator: a key element missing from standard gan. *arXiv preprint arXiv:1807.00734*, 2018. 4

[17] Robert Keys. Cubic convolution interpolation for digital image processing. *IEEE transactions on acoustics, speech, and signal processing*, 29(6):1153–1160, 1981. 1

[18] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 6

[19] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 372–386, 2012. 5

[20] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. *arXiv preprint arXiv:1611.01236*, 2016. 2, 3

[21] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, Andrew Aitken, Alykhan Tejani, Johannes Totz, Zehan Wang, and Wenzhe Shi. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 1

[22] Christian Ledig, Lucas Theis, Ferenc Huszár, Jose Caballero, Andrew Cunningham, Alejandro Acosta, and Andrew et al. Aitken. Photo-realistic single image super-resolution using a generative adversarial network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4681–4690, 2017. 4

[23] Xin Li and Michael T Orchard. New edge-directed interpolation. *IEEE transactions on image processing*, 10(10): 1521–1527, 2001. 1

[24] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 1

[25] Andreas Lugmayr, Danelljan Martin, Radu Timofte, Manuel Fritsche, Shuhang Gu, Kuldeep Purohit, Praveen Kandula, Suin Maitreya, A. N. Rajagoapalan, Joon Nam Hyung, Won Yu Seung, Kim Guisik, Kwon Dokyeong, Hsu Chih-Chung, Lin Chia-Hsiang, Huang Yuanfei, Sun Xiaopeng, Lu Wen, Li Jie, Gao Xinbo, Bell-Kligler Sefi, Assaf Shocher, and Irani Michal. Aim 2019 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshop (ICCVW)*, pages 3575–3583, 2019. 1, 6

[26] Andreas Lugmayr, Danelljan Martin, and Radu Timofte. Ntire 2020 challenge on real-world image super-resolution: Methods and results. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition workshops*, pages 494–495, 2020. 1, 6

[27] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017. 1, 2, 3

[28] Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32*, 2019. 2

[29] Assaf Shocher, Nadav Cohen, and Michal Irani. "zero-shot" super-resolution using deep internal learning. In *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pages 3118–3126, 2018. 1

[30] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 4

[31] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. 1

[32] Radu Timofte, Eirikur Agustsson, Luc Van Gool, Ming-Hsuan Yang, and Lei Zhang. Ntire 2017 challenge on single image super-resolution: Methods and results. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 114–125, 2017. 5

[33] Xintao Wang, Ke Yu, Shixiang Wu, Jinjin Gu, Yihao Liu, Chao Dong, Yu Qiao, and Chen Change Loy. Esrgan: Enhanced super-resolution generative adversarial networks. In *Proceedings of the European conference on computer vision (ECCV) workshops*, 2018. 1, 3, 4, 6, 7, 8

[34] Chih-Yuan Yang, Chao Ma, and Ming-Hsuan Yang. Single-image super-resolution: A benchmark. In *European Conference on Computer Visio (ECCV)*, pages 372–386, 2014. 5

[35] Liangpei Zhang, Hongyan Zhang, Huanfeng Shen, and Pingxiang Li. A super-resolution reconstruction algorithm for surveillance images. *Signal Processing*, 90(3):848–859, 2010. 1

[36] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 5

[37] Wenlong Zhang, Yihao Liu, Chao Dong, and Yu Qiao. Ranksrgan: Generative adversarial networks with ranker for image super-resolution. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3096–3105, 2019. 1

[38] Wang Zhou, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004. 5