

Evaluating Robustness of Deep Image Super-Resolution Against Adversarial Attacks

Jun-Ho Choi¹, Huan Zhang², Jun-Hyuk Kim¹, Cho-Jui Hsieh², and Jong-Seok Lee¹

¹School of Integrated Technology, Yonsei University

{idearibosome, junhyuk.kim, jong-seok.lee}@yonsei.ac.kr

²Department of Computer Science, University of California, Los Angeles
huanzhang@ucla.edu chohsieh@cs.ucla.edu

Abstract

Single-image super-resolution aims to generate a high-resolution version of a low-resolution image, which serves as an essential component in many computer vision applications. This paper investigates the robustness of deep learning-based super-resolution methods against adversarial attacks, which can significantly deteriorate the super-resolved images without noticeable distortion in the attacked low-resolution images. It is demonstrated that state-of-the-art deep super-resolution methods are highly vulnerable to adversarial attacks. Different levels of robustness of different methods are analyzed theoretically and experimentally. We also present analysis on transferability of attacks, and feasibility of targeted attacks and universal attacks.

1. Introduction

Single-image super-resolution, which is to generate a high-resolution version of a low-resolution image, is one of the popular research areas in recent years. While simple interpolation methods such as bilinear and bicubic upscaling have been used popularly, the development of deep learning-based approaches, which is triggered by a simple convolutional network model named super-resolution convolutional neural network (SRCNN) [7], offers much better quality of the upscaled images. The improvement of the super-resolution technique extends its applications to broader areas, including video streaming, surveillance, medical diagnosis, and satellite photography [23].

While many deep learning-based super-resolution methods have been introduced, their robustness against intended attacks has not been thoroughly studied. The vulnerability of deep networks has been an important issue in recent years, since various investigations report that the attack can

fool the deep classification models and can cause severe security issues [9, 17]. The similar issues can be raised for the super-resolution applications, since the deteriorated outputs can directly affect the reliability and stability of the systems employing super-resolution as their key components.

In this paper, we investigate the robustness of deep learning-based super-resolution against adversarial attacks, which is the first work to the best of our knowledge. Our attacks generate perturbations in the input images, which are not visually noticeable but can largely deteriorate the quality of the outputs. The main contributions of this paper can be summarized as follows.

- We propose three adversarial attack methods for super-resolution, which slightly perturb a given low-resolution image but result in significantly deteriorated output images, including basic, universal, and partial attacks. The methods are based on the methods widely used in the image classification tasks, and we optimize them for the super-resolution tasks.
- We present thorough analysis of the robustness of the super-resolution methods, by providing experimental results using the adversarial attack methods. We employ various state-of-the-art deep learning-based super-resolution methods having variable characteristics in terms of model structure, training objective, and model size.
- We further investigate the relation of robustness to the model properties and measure the transferability. In addition, we provide three advanced topics, including targeted attack, attack-agnostic robustness measurement, and simple defense methods of the attack.

2. Related Work

Super-resolution. Recently, the trend of super-resolution researches has been shifted to the deep learning-based

methods. One of the notable methods that achieve much improved performance is the enhanced deep super-resolution (EDSR) model [12]. Later, Zhang *et al.* [25] propose a more advanced network model named residual channel attention network (RCAN), which applies an attention mechanism to exploit image features effectively.

While the aforementioned methods focus on achieving high performance in terms of peak signal-to-noise ratio (PSNR), some researchers argue that considering only such a distortion measure does not necessarily enhance perceptual image quality [5]. To deal with this, perceptually optimized super-resolution methods are proposed, which employ generative adversarial networks (GANs) [8]. One of the state-of-the-art methods is the enhanced super-resolution generative adversarial network (ESRGAN) [19], which generates more visually appealing outputs than other conventional methods, even though the PSNR values are lower. Choi *et al.* [6] develop the four-pass perceptual enhanced upscaling super-resolution (4PP-EUSR) method, which considers both the quantitative and perceptual quality to obtain more natural upscaled images.

Since super-resolution is also a useful component in mobile applications, some studies focus on economizing the computational resource while reasonable performance is maintained. For instance, Ahn *et al.* [3] propose the cascading residual network (CARN) and its mobile version (CARN-M), which employ cascading residual blocks with shared model parameters.

Adversarial attack. Recent studies show that deep image classifiers are vulnerable to various adversarial attacks. Szegedy *et al.* [18] propose an optimization-based attack method that aims to minimize the amount of perturbation with changing the classification result of a classifier. Goodfellow *et al.* [9] develop the fast gradient sign method (FGSM), which uses the sign of the gradients that are obtained from the classifier. Kurakin *et al.* [11] extend it to an iterative approach (I-FGSM), which shows higher success rate of the attack than FGSM. These attacks are known as strong attack methods that can fool almost every state-of-the-art image classifier with high success rate [17].

Some studies provide in-depth analysis of the robustness of deep learning models. Liu *et al.* [13] measure transferability of the adversarial images, which is to find out whether the perturbations found for a classifier also work for another classifier. Moosavi-Dezfooli *et al.* [15] investigate a universal perturbation that can be applied to all images in a given dataset. Weng *et al.* [20] proposed a theoretical robustness measure, which does not depend on a specific attack method.

Adversarial attack on super-resolution. Recently, combining the super-resolution tasks with adversarial attacks has emerged. Mustafa *et al.* [16] presents a method employing super-resolution to defense deep image classifiers

against adversarial attacks. Yin *et al.* [22] employ the adversarial attack on super-resolution to fool the subsequent computer vision tasks. However, these studies investigate the effectiveness of the adversarial attack for other tasks rather than the super-resolution task itself, including image classification, style transfer, and image captioning, where the super-resolution is used as a pre-processing step before the main tasks. Therefore, the robustness of super-resolution itself against adversarial attacks, which is investigated in this paper, has not been addressed previously.

3. Attacks on Super-Resolution

3.1. Basic attack

The goal of the adversarial attack on super-resolution models is to inject a small amount of perturbation in the given input image so that the perturbation is not visually perceivable but results in significant deterioration in the super-resolved output. To do this, we develop an algorithm based on the idea of I-FGSM [11], which is one of the most widely used strong attacks for classification models.

Let \mathbf{X}_0 denote the original low-resolution input image and \mathbf{X} denote the attacked version of \mathbf{X}_0 . From these images, we obtain the super-resolved high-resolution images $f(\mathbf{X}_0)$ and $f(\mathbf{X})$, respectively, via a given super-resolution model $f(\cdot)$. Our objective is to maximize the amount of deterioration in the super-resolved output, which can be defined as:

$$\mathcal{L}(\mathbf{X}, \mathbf{X}_0) = \|f(\mathbf{X}) - f(\mathbf{X}_0)\|_2. \quad (1)$$

To find an \mathbf{X} to minimize (1) with bounded ℓ_∞ -norm constraint ($\|\mathbf{X} - \mathbf{X}_0\|_\infty \leq \alpha$), we adopt the I-FGSM update rule that iteratively updates \mathbf{X} by:

$$\tilde{\mathbf{X}}_{n+1} = \text{clip}_{0,1}\left(\mathbf{X}_n + \frac{\alpha}{T} \text{sgn}(\nabla \mathcal{L}(\mathbf{X}_n, \mathbf{X}_0))\right) \quad (2)$$

$$\mathbf{X}_{n+1} = \text{clip}_{-\alpha, \alpha}(\tilde{\mathbf{X}}_{n+1} - \mathbf{X}_0) + \mathbf{X}_0 \quad (3)$$

where T is the number of iterations, $\text{sgn}(\nabla \mathcal{L}(\mathbf{X}_n, \mathbf{X}_0))$ is the sign of the gradient of (1), and

$$\text{clip}_{a,b}(\mathbf{X}) = \min(\max(\mathbf{X}, a), b). \quad (4)$$

The term α not only controls the amount of contribution that the calculated gradient provides at each iteration, but also limits the maximum amount of perturbation to prevent noticeable changes of the attacked input image. The final adversarial example is obtained by $\mathbf{X} = \mathbf{X}_T$.

3.2. Universal attack

Although an adversarial image can be found for each image as in Section 3.1, it is also possible to find an *image-agnostic* adversarial perturbation, which can affect any input image for a certain super-resolution method [15]. We

Method	# parameters	# layers	GAN-based
EDSR [12]	43.1M	69	-
EDSR-baseline [12]	1.5M	37	-
RCAN [25]	15.6M	815	-
4PP-EUSR [6]	6.3M	95	✓
ESRGAN [19]	16.7M	351	✓
RRDB [19]	16.7M	351	-
CARN [3]	1.1M	34	-
CARN-M [3]	0.3M	43	-

Table 1. Properties of the super-resolution methods.

apply this concept in our study by altering the formulation of our basic attack as follows.

Assume that there are K images in the dataset, where the k -th image is denoted as \mathbf{X}_0^k . With a universal perturbation Δ , we can obtain the adversarial images as:

$$\mathbf{X}^k = \text{clip}_{0,1}(\mathbf{X}_0^k + \Delta). \quad (5)$$

Then, we compute the average amount of deterioration as:

$$\mathcal{F}(\Delta) = \frac{1}{K} \sum_{k=1}^K \mathcal{L}(\mathbf{X}^k, \mathbf{X}_0^k). \quad (6)$$

Starting from $\Delta_0 = 0$, the universal perturbation is updated iteratively as:

$$\Delta_{n+1} = \text{clip}_{-\alpha, \alpha} \left(\Delta_n + \frac{\alpha}{T} \text{sgn}(\nabla \mathcal{F}(\Delta_n)) \right). \quad (7)$$

The final universal perturbation is obtained by $\Delta = \Delta_T$.

3.3. Partial attack

The basic attack in Section 3.1 finds a perturbation covering the whole region of the given image. We further investigate the robustness of the super-resolution methods by attacking only some part of the image, but measuring the amount of deterioration in the region that is not being attacked. With this experiment, we can examine how much the perturbation *permeates* into the adjacent regions during super-resolution.

Let \mathbf{M} denote a binary mask of the perturbation Δ , where only the region to be attacked is set to 1. The masked perturbation is $\Delta \circ \mathbf{M}$, where \circ denotes the element-wise multiplication. Then, (2) can be modified as:

$$\tilde{\mathbf{X}}_{n+1} = \text{clip}_{0,1} \left(\mathbf{X}_n + \frac{\alpha}{T} \text{sgn}(\nabla \mathcal{L}_{\mathbf{M}}(\mathbf{X}_n, \mathbf{X}_0)) \circ \mathbf{M} \right) \quad (8)$$

where

$$\mathcal{L}_{\mathbf{M}}(\mathbf{X}, \mathbf{X}_0) = \| (f(\mathbf{X}) - f(\mathbf{X}_0)) \circ (1 - \mathbf{M}_H) \|_2. \quad (9)$$

In (9), \mathbf{M}_H is a high-resolution counterpart of \mathbf{M} . The term $(1 - \mathbf{M}_H)$ ensures that the amount of deterioration is calculated only on the unperturbed regions. The final adversarial example is obtained by $\mathbf{X} = \mathbf{X}_T$.

4. Experimental Results

Datasets. We employ three image datasets that are widely used for benchmarking super-resolution methods: Set5 [4], Set14 [24], and BSD100 [14]. Each dataset consists of 5, 14, and 100 images, respectively.

Super-resolution methods. We consider eight state-of-the-art deep learning-based super-resolution methods having various model sizes and properties, including EDSR [12], EDSR-baseline [12], RCAN [25], 4PP-EUSR [6], ESRGAN [19], RRDB [19], CARN [3], and CARN-M [3]. Table 1 shows their characteristics in terms of the number of model parameters, the number of convolutional layers, and whether to employ GANs for training. EDSR-baseline is a smaller version of EDSR, RRDB is an alternative version of ESRGAN trained without the GAN, and CARN-M is a lightweight version of CARN in terms of the number of model parameters. In addition, we also consider the bicubic interpolation to compare its robustness against the adversarial attacks with that of the deep learning-based methods. We consider a scaling factor of 4 for all the super-resolution methods. In addition, we employ the pre-trained models provided by the original authors.

Implementation details. Our adversarial attack methods are implemented on the TensorFlow framework [2]. For all the attack methods, we set $\alpha \in \{1/255, 2/255, 4/255, 8/255, 16/255, 32/255\}$ and $T = 50$. For the universal attack, a perturbation Δ with a fixed spatial resolution is required in order to apply it to all images in a dataset. Therefore, we crop the center region of each input image with a fixed resolution. For the partial attack, we set the mask \mathbf{M} so as to attack the central part of the input image, i.e.,

$$\mathbf{M}_{(x,y)} = \begin{cases} 1 & \text{if } \frac{w}{4} \leq x < \frac{3w}{4}, \frac{h}{4} \leq y < \frac{3h}{4} \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

where $\mathbf{M}_{(x,y)}$ is the value of \mathbf{M} at (x, y) , and w and h are the width and height of the input image, respectively.

Performance measurement. We measure the robustness of the super-resolution methods against our adversarial attack methods in terms of PSNR. For low-resolution (LR) images, we calculate the PSNR values between the original and attacked images, i.e., \mathbf{X}_0 and \mathbf{X} . For super-resolved (SR) images, PSNR is measured between the output images obtained from the original and attacked input images, i.e., $f(\mathbf{X}_0)$ and $f(\mathbf{X})$. We report the averaged PSNR values for each dataset. For the partial attack, we calculate the PSNR values only for the outer region of the output image that corresponds to the masked region during the attack.

4.1. Basic attack

Figure 1 compares the performance of the super-resolution methods in terms of PSNR for the I-FGSM attack

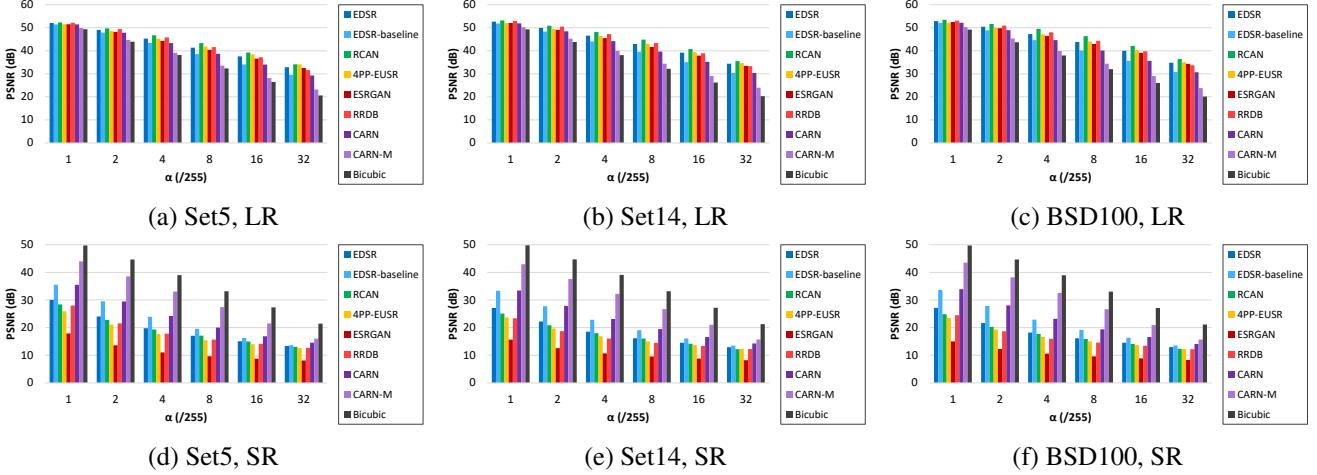


Figure 1. Comparison of the PSNR values of low-resolution (LR) and super-resolved (SR) images with respect to different α values for the basic attack on the Set5 [4], Set14 [24], and BSD100 [14] datasets.

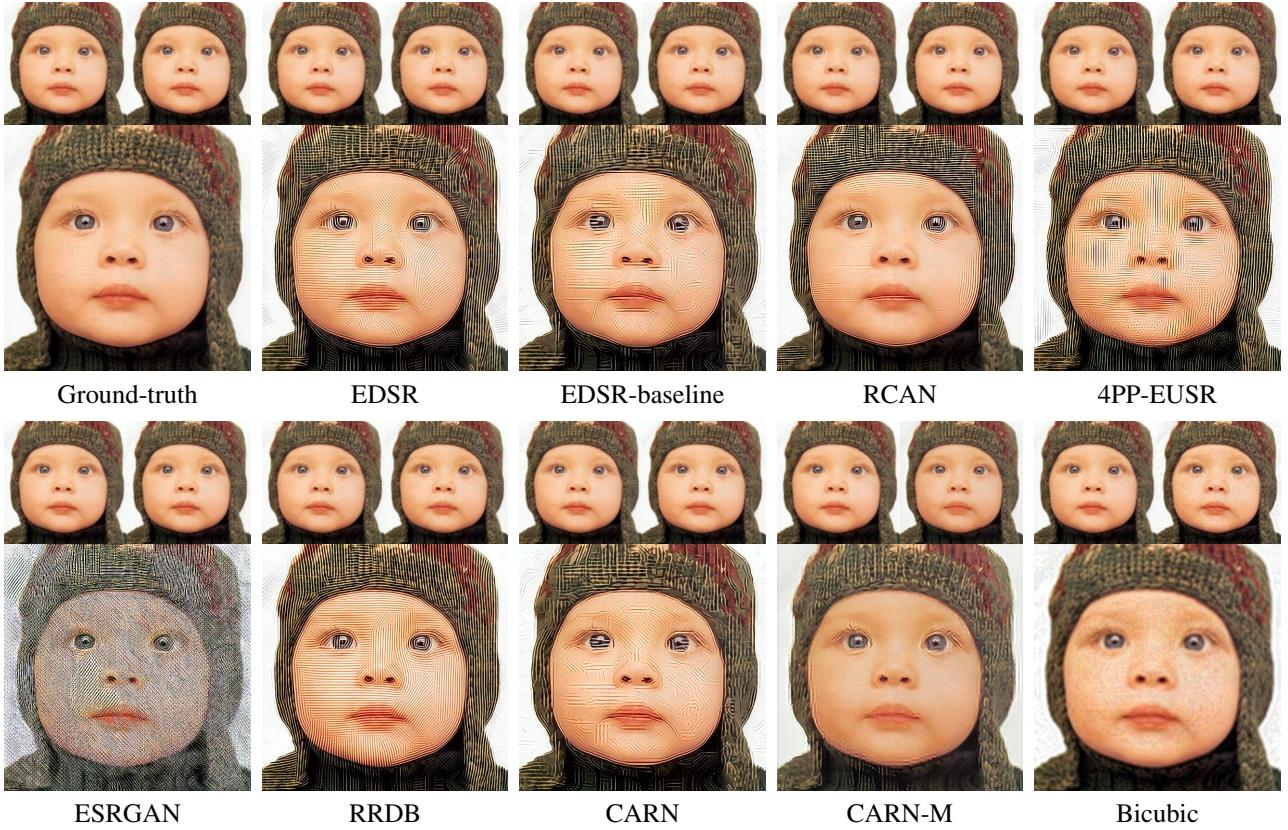


Figure 2. Visual comparison of the super-resolved outputs for the inputs attacked with $\alpha = 8/255$. In each case, (top-left) is the original input in Set5 [4], (top-right) is the adversarial input, and (bottom) is the output obtained from the adversarial input. The input images are enlarged two times for better visualization.

explained in Section 3.1. As α increases, quality degradation becomes severe in both the LR and SR images. However, it is much more significant in the SR images than the

LR images (i.e., lower PSNR values) except for the bicubic interpolation. For example, on the Set5 dataset, the PSNR values of LR and SR images for the EDSR model are 41.37

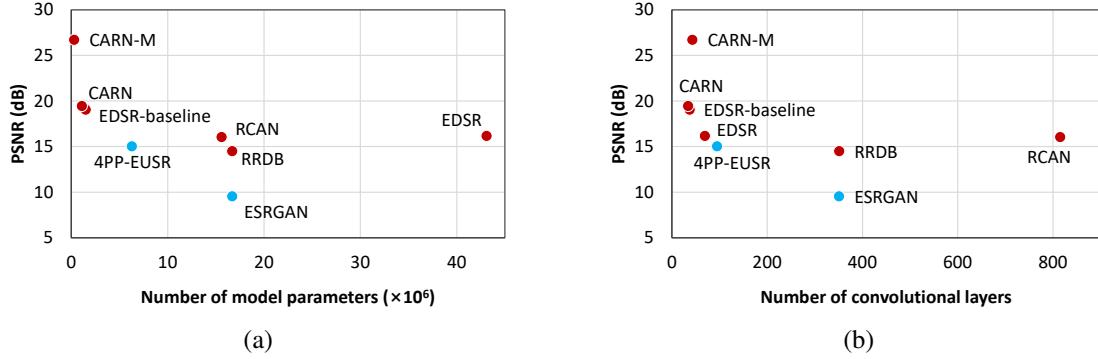


Figure 3. Comparison of the PSNR values of SR images for BSD100 [14] with respect to the model sizes in terms of (a) the number of model parameters and (b) the number of convolutional layers ($\alpha = 8/255$). Blue and red colors indicate the models trained with and without GANs, respectively.

Source model	PSNR (dB)	Target model							
		EDSR	EDSR-baseline	RCAN	4PP-EUSR	ESRGAN	RRDB	CARN	CARN-M
EDSR	16.14	32.88	25.82	23.86	16.26	23.57	32.80	37.74	
EDSR-baseline	24.44	19.19	23.65	21.23	15.15	22.29	26.82	33.62	
RCAN	30.57	35.49	15.89	26.60	18.94	29.74	35.57	40.46	
4PP-EUSR	27.25	32.76	26.83	15.02	16.16	24.71	32.87	37.97	
ESRGAN	28.64	33.11	28.59	24.28	9.57	24.46	33.30	36.56	
RRDB	25.55	33.09	25.31	23.77	15.86	14.59	32.91	38.11	
CARN	24.12	26.05	23.83	21.45	15.24	22.15	19.40	33.51	
CARN-M	27.34	28.20	27.20	23.49	16.27	26.77	28.20	26.66	

Figure 4. Comparison of the transferability in terms of PSNR for the BSD100 dataset [14] when $\alpha = 8/255$. Red and blue colors indicate the lowest and highest PSNR values (except the diagonal cells) for each target model, respectively.

and 17.05 dB, respectively, when $\alpha = 8/255$. Note that two images having a PSNR value higher than 30 dB can be regarded as visually identical images [10].

Figure 2 shows example LR and SR images for $\alpha = 8/255$. Overall, there is no obvious difference between the original and perturbed input images for all the super-resolution methods. However, significant quality deterioration can be observed in the SR images for all methods. ESRGAN shows the worst visual quality with degradation in all parts of the SR image, which can also be observed as the lowest PSNR values in Figures 1d, 1e, and 1f. For the other super-resolution models, fingerprint-like patterns are observed. This proves that all the deep learning-based super-resolution methods are highly vulnerable against the adversarial attack. In comparison, the bicubic method, although having lower super-resolution quality on clean data, is much more robust compared with the deep learning-based approaches.

Relation to model objectives. ESRGAN and 4PP-EUSR, which employ GANs for considering perceptual quality improvement, produce more significantly degraded outputs than the other methods. Since ESRGAN has exactly the

same structure as RRDB but is trained with a different objective (i.e., considering perceptual quality), the more significant vulnerability of ESRGAN than RRDB implies that differences of the training objectives affect the robustness against the adversarial attacks. It is known that the methods employing GANs tend to generate sharper textures than the other methods to ensure naturally appealing quality of the upscaled images [5]. Therefore, these methods amplify small perturbations significantly and produce undesirable textures, which makes them more vulnerable to the adversarial attacks than the methods without GANs.

Relation to model sizes. It is observed that the vulnerability of the super-resolution models is related to their model sizes. For example, EDSR-baseline, which is a smaller version of EDSR, shows higher PSNR values for SR images than EDSR, as shown in Figures 1d, 1e, and 1f. This is confirmed in Figure 3, where we compare the robustness with respect to the model size. The figure explains that the PSNR values of SR images tend to decrease when more model parameters or more convolutional layers are employed. Further analysis on this phenomenon is given in Section 4.3.

Transferability. In the classification tasks, the “transferability” means the possibility that a misclassified adversarial example is also misclassified by another classifier [13]. We also examine the transferability of adversarial attacks in super-resolution. In other words, an adversarial example that is found for a “source” super-resolution model is inputted to another “target” model, and the PSNR value of the output image is measured.

Figure 4 summarizes the transferability for the deep learning-based super-resolution models on the BSD100 dataset, where $\alpha = 8/255$. The figure shows that the adversarial examples are transferable between different models to some extent, and the level of transferability differs depending on the combination of the source and target models. The adversarial examples found for CARN and EDSR-

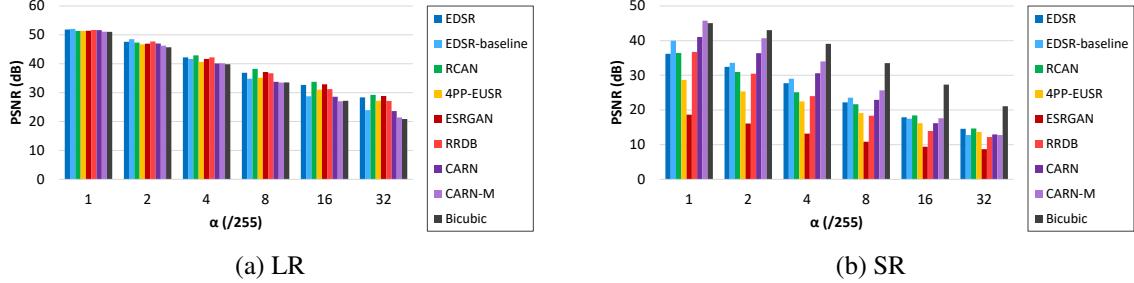


Figure 5. Comparison of the PSNR values of LR and SR images with respect to different α values for the universal attack on the BSD100 dataset [14].

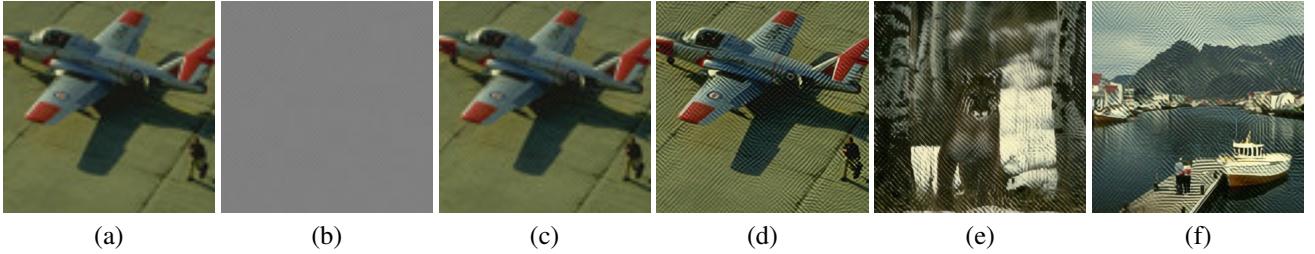


Figure 6. Visual examples of the universal attack with $\alpha = 4/255$ on the BSD100 dataset [14] for the RCAN model. (a) LR (original) (b) Perturbation (c) LR (attacked) (d) SR (e–f) Other examples obtained from the images attacked with the same perturbation

baseline are highly transferable, while those for RCAN are the least transferable. The result implies that RCAN has its own specific characteristics in recovering the textures from the input images, which makes the perturbations associated with such characteristics less effective in the other super-resolution methods.

4.2. Universal attack

Figure 5 compares the performance of the super-resolution methods for the BSD100 dataset with respect to different α values when the universal attack is applied. The figure confirms that the super-resolution models are also vulnerable to the image-agnostic universal attack, although the universal attack requires larger perturbations of the input images (i.e., slightly lower PSNR values in Figure 5a than in Figure 1c) and is slightly less powerful than the image-specific attack (i.e., slightly higher PSNR values in Figure 5b than in Figure 1f). Compared to the results of the basic attack (Figure 1), the same tendency is observed: both ESRGAN and 4PP-EUSR are the most vulnerable and the bicubic interpolation is the most robust.

Figure 6 shows visual examples of the universal attack for RCAN, where α is $4/255$. From all images of the BSD100 dataset, our attack method finds a universal perturbation (Figure 6b), which changes the input image shown in Figure 6a to the one in Figure 6c. While the attacked LR image has hardly noticeable differences from the original image, its upscaled version contains significant artifacts

as shown in Figure 6d. Similar artifacts can be observed in the other SR images attacked with the same perturbation, as shown in Figures 6e and 6f. This demonstrates that the state-of-the-art super-resolution methods using deep learning are also vulnerable to the universal perturbation.

4.3. Partial attack

Figure 7 shows the PSNR values of the SR images for the partial attack with respect to different α values. The rank of the super-resolution methods in terms of PSNR is the same to that for the basic attack, except that the PSNR values of the partial attack are much higher than those of the basic attack, since the region where PSNR is measured is not directly perturbed in the LR image. This shows that the propagation of the perturbation to the neighboring pixels during upscaling accounts for different levels of vulnerability of different super-resolution models. For instance, all the PSNR values of ESRGAN, except for $\alpha = 1$ in Set5, are lower than 30 dB due to the partial attack.

Figure 8 shows example SR images obtained from an image in the Set14 dataset that are partially attacked with $\alpha = 8/255$. The degradation due to the attack propagates outside of the attacked region, which are particularly noticeable for ESRGAN and RRDB. This is because the kernels of the convolutional layers operate on not only the pixel of a target position but also its adjacent pixels. Moreover, the propagation of the perturbation due to such operations is further extended through multiple convolutional layers,

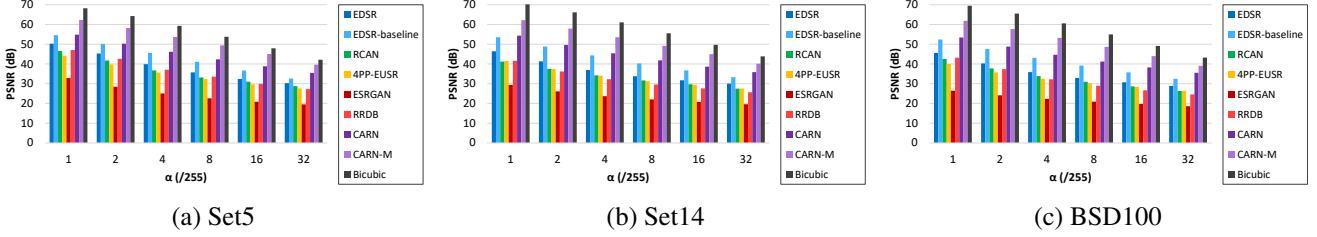


Figure 7. Comparison of the PSNR values of SR images with respect to different α values for the partial attack.

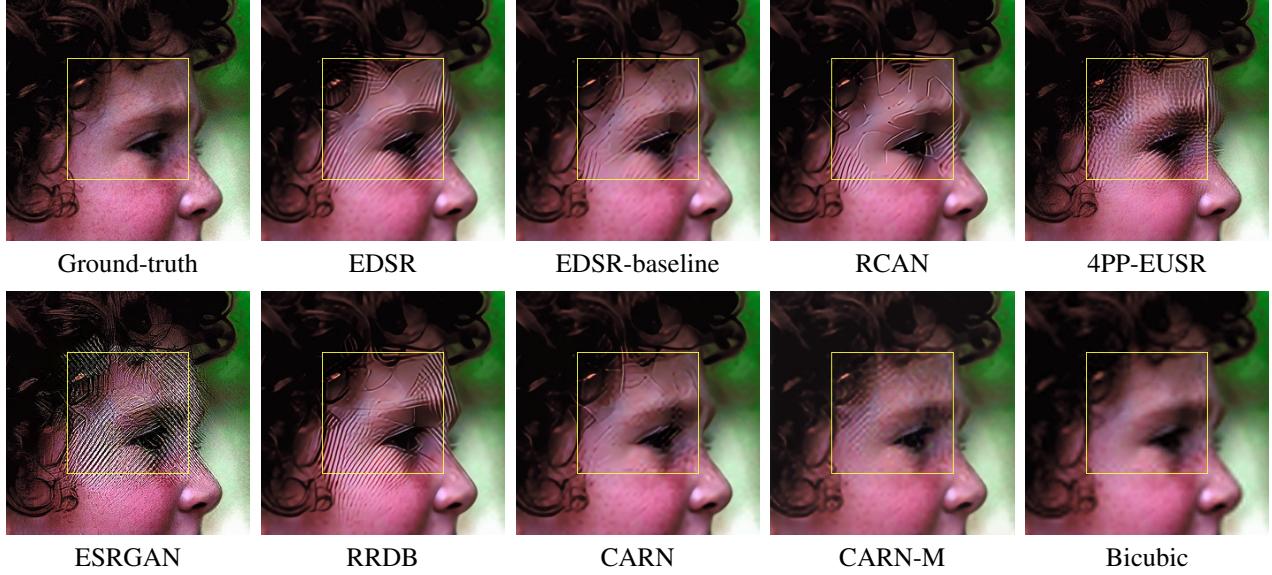


Figure 8. Visual comparison of the SR images for the partial adversarial attack with $\alpha = 8/255$ on an image of Set14 [24]. The regions marked with yellow boxes correspond to the regions where the attack is applied in the LR images.

which accounts for the result shown in Figure 3b.

5. Advanced Topics

5.1. Targeted attack

In the case of the classification tasks, it is possible to attack an image so that a classifier wrongly classifies the image as a specific target class. We present a showcase demonstrating that this concept can be also applied to the super-resolution methods. In other words, instead of degrading quality of the output image, the targeted attack makes a super-resolution method generate an image that is more similar to a target image than the original ground-truth one. For this, we modify (2) as:

$$\tilde{\mathbf{X}}_{n+1} = \text{clip}_{0,1} \left(\mathbf{X}_n - \frac{\alpha}{T} \text{sgn}(\nabla \mathcal{L}(\mathbf{X}_n, \mathbf{X}^*)) \right) \quad (11)$$

where \mathbf{X}^* is the target image.

For demonstration, we use two adjacent frames of a video named “foreman” [1]. Figure 9 shows the result for 4PP-EUSR, where $\alpha = 16/255$ and $T = 50$. The figure

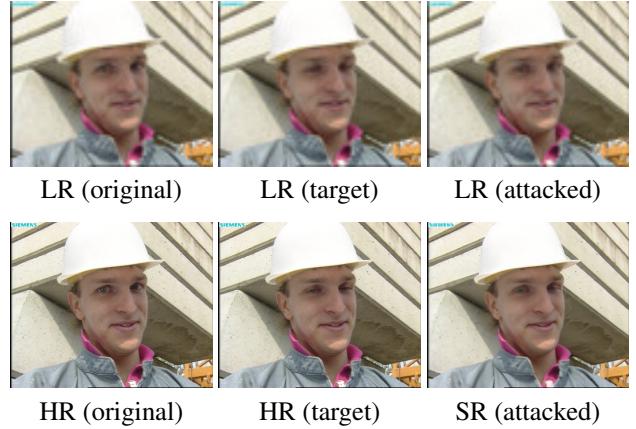


Figure 9. Result of the targeted attack with $\alpha = 16/255$ using two frames of a video “foreman” [1] for 4PP-EUSR [6].

shows that the targeted attack is successful: the perturbation is generated so as to make the super-resolution method produce the upscaled output (“SR (attacked)”), which looks

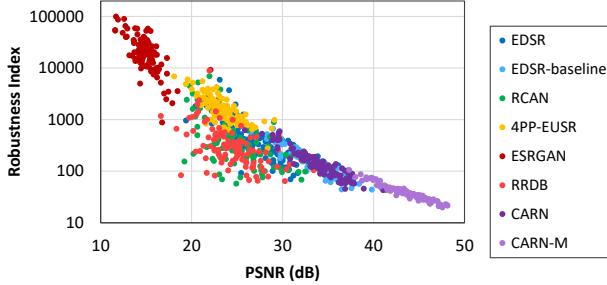


Figure 10. PSNR vs. the robustness index for the BSD100 dataset [14] when $\alpha = 1/255$. Each point corresponds to each image in the dataset.

more similar to the target high-resolution (HR) image with *half-closed eyes* (“HR (target)”) than the original ground-truth image with *open eyes* (“HR (original)”), while the attacked input image (“LR (attacked)”) still looks more similar to the original image (“LR (original)”) than the low-resolution version of the target image (“LR (target)”). In addition, we conduct a subjective test with 20 human observers, and 10 of them recognized the attacked output (“SR (attacked)”) as *closed eyes*. These results have serious security implications: attacks on super-resolution can not only compromise the fundamental goal of super-resolution (i.e., image quality enhancement) but also jeopardize further manual or automatic examination of the super-resolved images (e.g., identifying persons or objects in surveillance cameras, recognizing text in images, etc.).

5.2. Robustness measure

Recently, Weng *et al.* [20] propose an *attack-agnostic* robustness measure of classification models, called cross Lipschitz extreme value for network robustness (CLEVER), which does not depend on specific attack methods. It estimates the lower bound of robustness using the cross Lipschitz constant based on the extreme value theory. We apply the core idea of this method to the super-resolution tasks in order to theoretically validate the experimental results shown in Section 4.

Let \mathbf{X}_0 denote the original input image. We first obtain N_s random perturbations, which are within $[-\alpha, \alpha]$ for each pixel. Let $\Delta^{(i)}$ denote the i -th random perturbation. Then, we compute $b_i = \|\nabla \mathcal{L}(\mathbf{X}_0 + \Delta^{(i)}, \mathbf{X}_0)\|_1$ for all perturbations, where \mathcal{L} is defined in (1). Finally, we regard the maximum b_i as the robustness index; a large robustness index indicates high vulnerability. We set N_s and α to 1024 and $1/255$, respectively.

Figure 10 shows the PSNR values for SR images and robustness indices of the eight deep learning-based super-resolution methods for the BSD100 dataset, where the PSNR values are obtained from the basic attack with the same α value (Section 4.1). The result shows that the

robustness index is strongly correlated to PSNR. For instance, ESRGAN has the largest robustness indices, which shows the lowest PSNR values; the EDSR-baseline model has the similar robustness as the CARN model in terms of both PSNR and the robustness index. Furthermore, in each method, the robustness index successfully explains relative vulnerability of different images. The applicability of the CLEVER method for explaining the robustness of the super-resolution methods implies that the underlying mechanisms of the adversarial attacks share similarity between the classification and super-resolution tasks.

5.3. Defense

We show two simple defense methods against attacks. First, we adopt a resizing method [21] by reducing the size of the attacked input image by one pixel and then resizing it back to the original resolution, which is then inputted to the SR model. With this, PSNR for EDSR with $\alpha = 8/255$ increases from 16.14 to 25.01 dB. Second, we employ the geometric self-ensemble method used in the EDSR model [12]. With this, PSNR for EDSR with $\alpha = 8/255$ increases from 16.14 to 23.47 dB. More advanced defense methods can be investigated in the future work.

6. Conclusion

We have investigated the robustness of deep learning-based super-resolution methods against adversarial attacks, for which the attack methods for the classification tasks are optimized for our objectives. Our results showed that state-of-the-art deep learning-based super-resolution methods are highly vulnerable to adversarial attacks, which is largely due to the perturbation propagation through the convolutional operation. It was possible to measure different levels of robustness of different methods using the attack-agnostic robustness measure. We also showed the feasibility of generating universal attacks and transferring attacks across super-resolution methods. Furthermore, it was shown that the targeted attack can change the content of the image during super-resolution.

Acknowledgement

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the “ICT Consilience Creative Program” (IITP-2019-2017-0-01015) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation). In addition, this work was also supported by the IITP grant funded by the Korea government (MSIT) (R7124-16-0004, Development of Intelligent Interaction Technology Based on Context Awareness and Human Intention Understanding).

References

- [1] YUV 4:2:0 video sequences. <http://trace.eas.asu.edu/yuv/index.html>. 7
- [2] Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. TensorFlow: A system for large-scale machine learning. In *Proceedings of the USENIX Symposium on Operating Systems Design and Implementation*, pages 265–283, 2016. 3
- [3] Namhyuk Ahn, Byungkon Kang, and Kyung-Ah Sohn. Fast, accurate, and lightweight super-resolution with cascading residual network. In *Proceedings of the European Conference on Computer Vision*, pages 252–268, 2018. 2, 3, 10, 12
- [4] Marco Bevilacqua, Aline Roumy, Christine Guillemot, and Marie Line Alberi-Morel. Low-complexity single-image super-resolution based on nonnegative neighbor embedding. In *Proceedings of the British Machine Vision Conference*, pages 1–10, 2012. 3, 4, 10
- [5] Yochai Blau and Tomer Michaeli. The perception-distortion tradeoff. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6228–6237, 2018. 2, 5
- [6] Jun-Ho Choi, Jun-Hyuk Kim, Manri Cheon, and Jong-Seok Lee. Deep learning-based image super-resolution considering quantitative and perceptual quality. *arXiv:1809.04789*, 2018. 2, 3, 7, 10
- [7] Chao Dong, Chen Change Loy, Kaiming He, and Xiaoou Tang. Learning a deep convolutional network for image super-resolution. In *Proceedings of the European Conference on Computer Vision*, pages 184–199, 2014. 1
- [8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 2672–2680, 2014. 2
- [9] Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations*, 2015. 1, 2
- [10] Rong Huang and Kouichi Sakurai. A robust and compression-combined digital image encryption method based on compressive sensing. In *Proceedings of the International Conference on Intelligent Information Hiding and Multimedia Signal Processing*, pages 105–108, 2011. 5
- [11] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial machine learning at scale. In *Proceedings of the International Conference on Learning Representations*, 2016. 2
- [12] 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. 2, 3, 8, 10, 11, 12
- [13] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *Proceedings of the International Conference on Learning Representations*, 2016. 2, 5
- [14] David Martin, Charless Fowlkes, Doron Tal, and Jitendra Malik. A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 416–423, 2001. 3, 4, 5, 6, 8, 10, 12, 13, 14
- [15] Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Universal adversarial perturbations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1765–1773, 2017. 2
- [16] Aamir Mustafa, Salman H. Khan, Munawar Hayat, Jianbing Shen, and Ling Shao. Image super-resolution as a defense against adversarial attacks. *arXiv:1901.01677*, 2019. 2
- [17] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy? – A comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision*, pages 631–648, 2018. 1, 2
- [18] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations*, 2014. 2
- [19] 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 Workshops*, pages 63–79, 2018. 2, 3, 10, 13
- [20] Tsui-Wei Weng, Huan Zhang, Pin-Yu Chen, Jinfeng Yi, Dong Su, Yupeng Gao, Cho-Jui Hsieh, and Luca Daniel. Evaluating the robustness of neural networks: An extreme value theory approach. In *Proceedings of the International Conference on Learning Representations*, 2018. 2, 8
- [21] Cihang Xie, Jianyu Wang, Zhishuai Zhang, Zhou Ren, and Alan Yuille. Mitigating adversarial effects through randomization. In *Proceedings of the International Conference on Learning Representations*, 2018. 8
- [22] Minghao Yin, Yongbing Zhang, Xiu Li, and Shiqi Wang. When deep fool meets deep prior: Adversarial attack on super-resolution network. In *Proceedings of the ACM International Conference on Multimedia*, pages 1930–1938, 2018. 2
- [23] Linwei Yue, Huanfeng Shen, Jie Li, Qiangqiang Yuan, Hongyan Zhang, and Liangpei Zhang. Image super-resolution: The techniques, applications, and future. *Signal Processing*, 128:389–408, 2016. 1
- [24] Roman Zeyde, Michael Elad, and Matan Protter. On single image scale-up using sparse-representations. In *Proceedings of the International Conference on Curves and Surfaces*, pages 711–730, 2010. 3, 4, 7, 10, 11, 14
- [25] Yulun Zhang, Kunpeng Li, Kai Li, Lichen Wang, Bineng Zhong, and Yun Fu. Image super-resolution using very deep residual channel attention networks. In *Proceedings of the European Conference on Computer Vision*, pages 286–301, 2018. 2, 3, 10, 13

Supplementary Material

In this supplementary material, we provide additional results that could not be included in the main paper due to the page limit.

More visual comparisons of the basic attack. We provide two additional visual comparisons of the basic attack shown in Section 4.1 of the main paper. Figure 11 shows additional example low-resolution (LR) and super-resolved (SR) images obtained from an image of Set14 [24] with $\alpha = 8/255$. Undesirable artifacts similar to those observed in Figure 2 of the main paper can be found. Figure 12 shows the example images obtained by the EDSR model [12] with different α values. As α increases, the upscaled images become more deteriorated, whereas the perturbed input images still look similar to the original image. The results support that the deep super-resolution methods are highly vulnerable against the adversarial attack in various cases.

Visualized results of transferability. In Section 4.1 of the main paper, we compared the transferability of the deep super-resolution methods in terms of peak signal-to-noise ratio (PSNR). According to Figure 4 in the main paper, EDSR-baseline [12] and CARN [3] show higher transferability than the other models, whereas RCAN [25] and ESRGAN [19] show lower transferability. Here, we visually explain the transferability of these four super-resolution methods in Figures 13, 14, 15, and 16. In the figures, a LR image in the BSD100 dataset [14] is attacked with one of the super-resolution models and inputted to the other super-resolution models including EDSR [12], EDSR-baseline, RCAN, 4PP-EUSR [6], ESRGAN, RRDB [19], CARN, and CARN-M [3]. In Figures 13 and 14, the attacked LR image successfully deteriorates the SR images obtained from the other methods, where similar fingerprint-like textures are observed as in Figure 2 of the main paper. On the other hand, in Figures 15 and 16, the perturbations found for RCAN and ESRGAN are not so effective for the other models; the amounts of deterioration in the SR images produced by the other models are much smaller than those triggered by the perturbations for EDSR-baseline and CARN (Figures 13 and 14).

Transferability of the universal attack. We examine the universal attack across datasets, i.e., the universal perturbation obtained for the BSD100 dataset [14] is applied to the images of the Set14 dataset [24]. Figure 17 shows the super-resolved (SR) images obtained by the RCAN model [25], where the perturbation shown in Figure 6b of the main paper is applied. This result verifies that the universal attack is transferable to unseen images.

Advanced partial attack. The objective of the partial attack in Section 4.3 of the main paper is to examine how

the perturbation planted in a region propagates spatially outside the region. Partial attacks with more complex masks can also be done using the proposed method. Figure 18 shows the attack results where the perturbation is applied on the face region of an image in Set5 [4]. It is observed that strong degradations are introduced around the face boundaries.

Additional example of the targeted attack. We provide an additional example of the targeted attack, which is explained in Section 5.1 of the main paper. Figure 19 shows the result. In the figure, the original number 87 in the original high-resolution image (“HR (original)”) is changed to 89 in the SR version (“SR (attacked)”). We conduct a subjective test with 20 human observers, and all the observers recognized the number in the red box of “SR (attacked)” as 89 instead of 87.

Robustness measure. We employed the “robustness index” in Section 5.2 of the main paper. Here we provide additional results obtained with different α values (i.e., $\alpha = 2/255$ and $\alpha = 4/255$). Figure 20 depicts the relationship between the PSNR values for SR images obtained with the basic attack (Section 4.1 of the main paper) and the robustness indices of the deep super-resolution models for the BSD100 dataset, where $\alpha = 2/255$ and $\alpha = 4/255$. When these figures and Figure 10 of the main paper are compared, increasing α results in decreasing the PSNR values and increasing the robustness index values, as expected. In addition, as in the result with $\alpha = 1/255$ (Figure 10 of the main paper), the robustness index is strongly correlated to PSNR regardless of the value of α , which supports the usefulness of the robustness index for explaining the relative vulnerability of the different super-resolution methods.

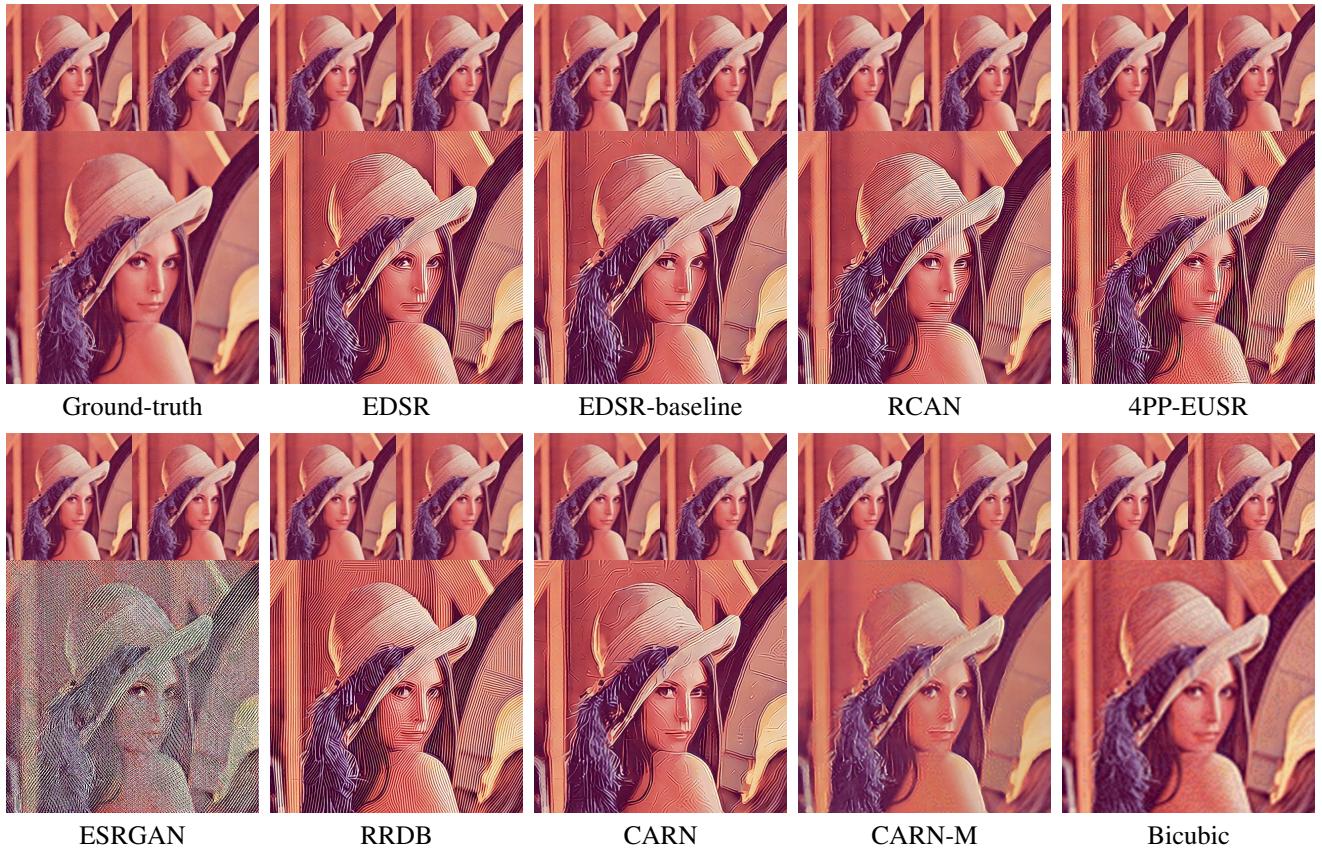


Figure 11. Visual comparison of the super-resolved outputs for the inputs attacked with $\alpha = 8/255$. In each case, (top-left) is the original input in Set14 [24], (top-right) is the adversarial input, and (bottom) is the output obtained from the adversarial input. The input images are enlarged two times for better visualization.

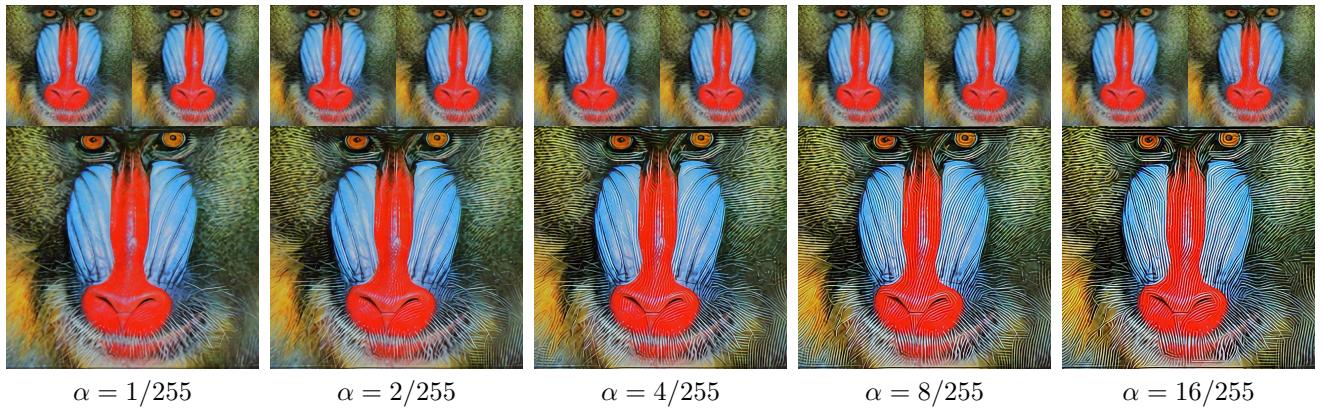


Figure 12. Visual comparison of the super-resolved outputs for the inputs attacked with different α values. In each case, (top-left) is the original input in Set14 [24], (top-right) is the adversarial input, and (bottom) is the output obtained on EDSR [12]. The input images are enlarged two times for better visualization.

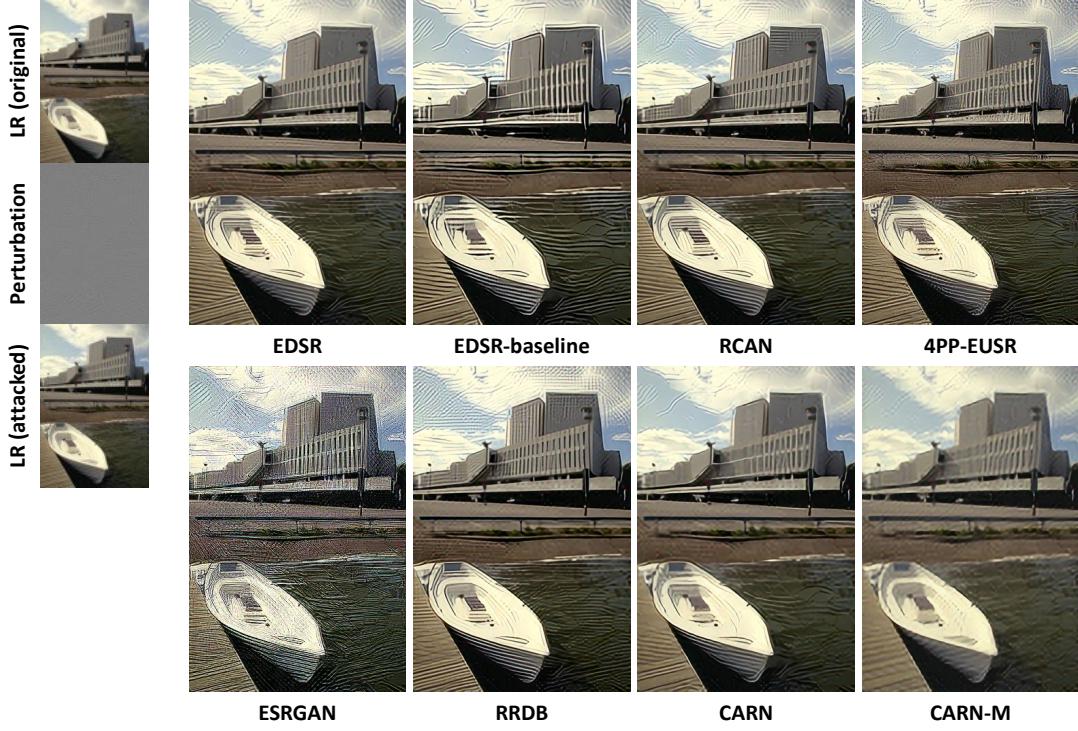


Figure 13. Visual examples of the transferred attack where EDSR-baseline [12] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [14] dataset is used.

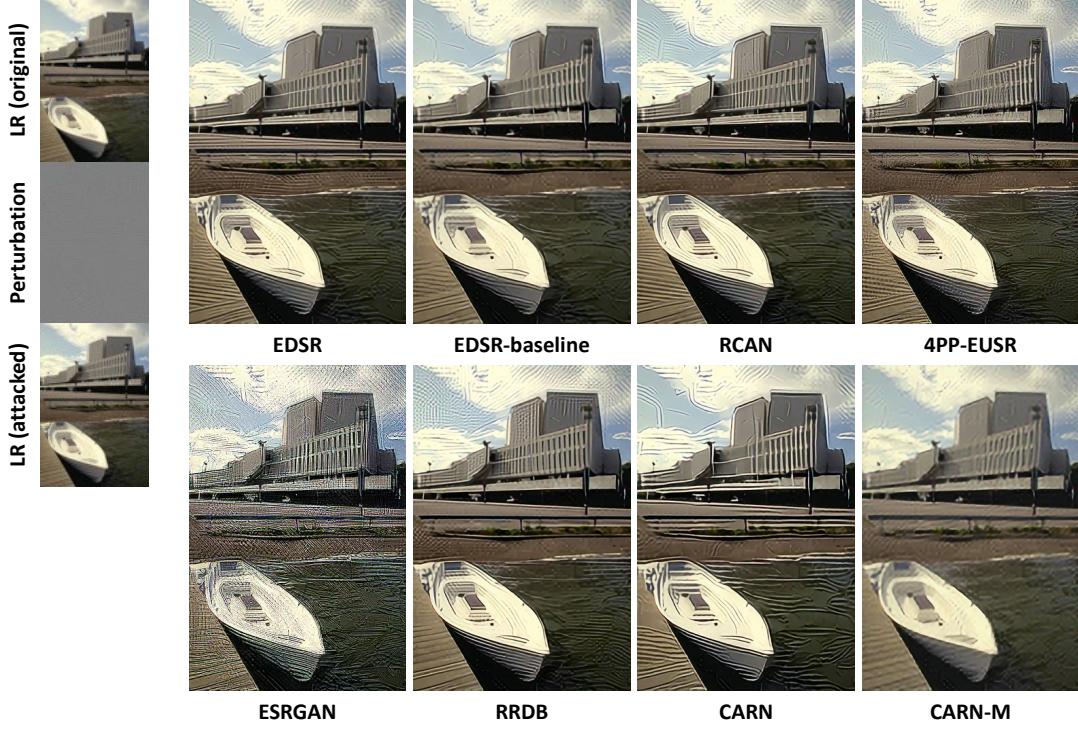


Figure 14. Visual examples of the transferred attack where CARN [3] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [14] dataset is used.

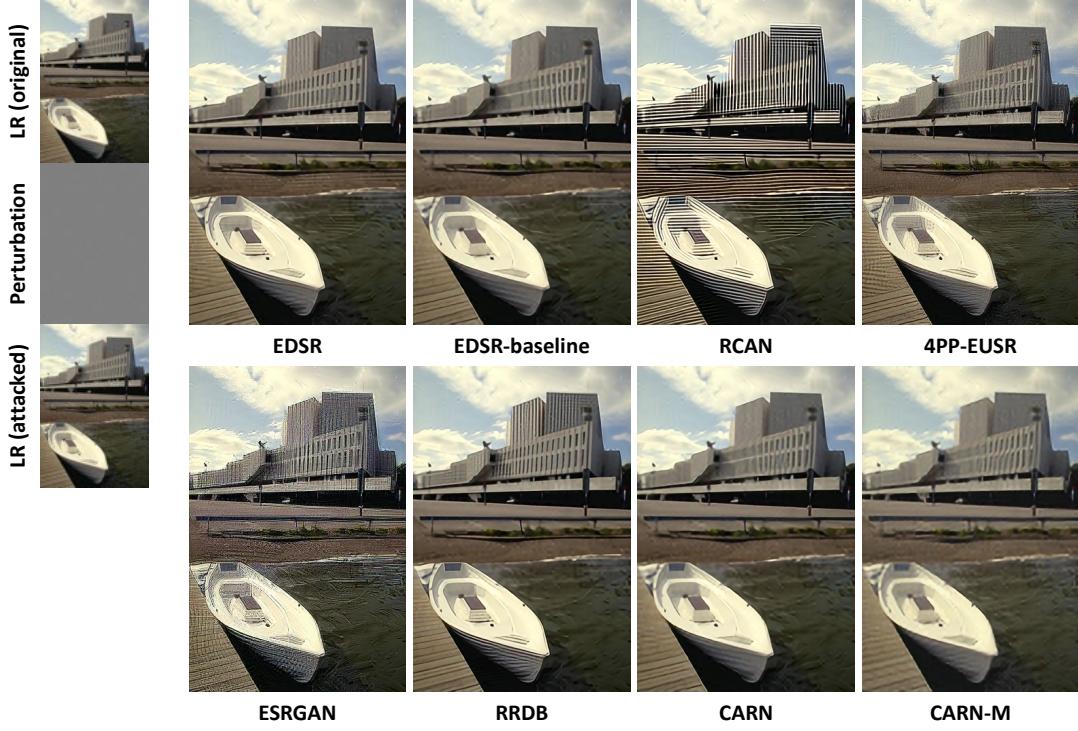


Figure 15. Visual examples of the transferred attack where RCAN [25] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [14] dataset is used.

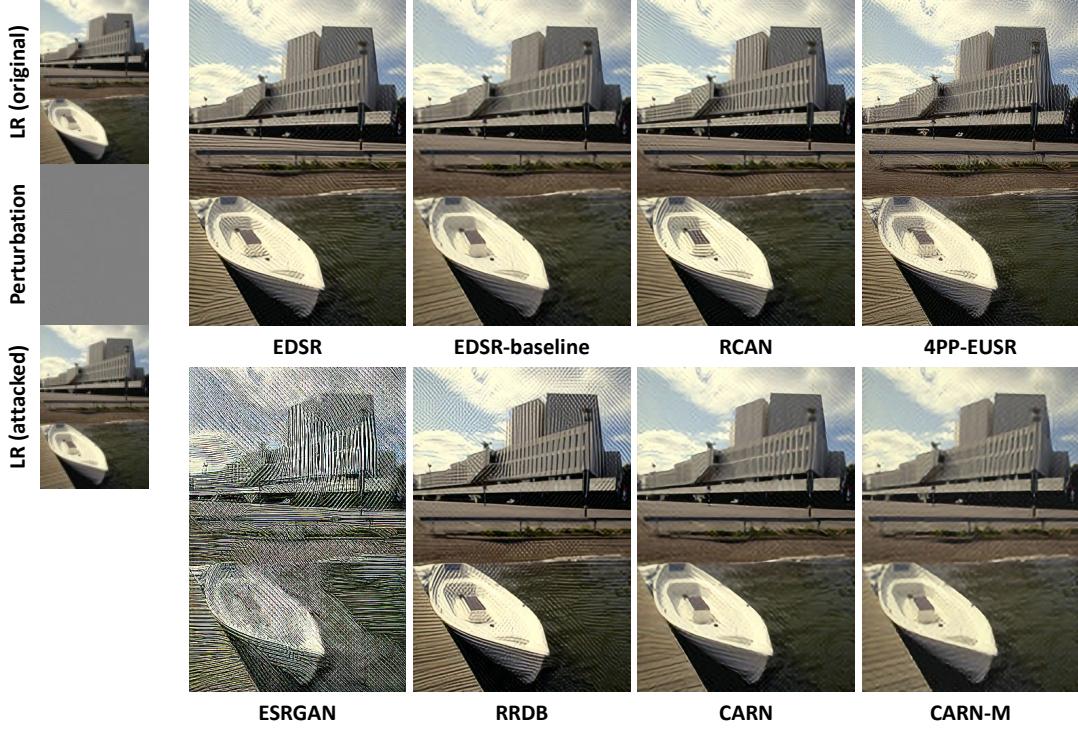


Figure 16. Visual examples of the transferred attack where ESRGAN [19] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [14] dataset is used.



Figure 17. Results of the universal attack applied to the Set14 dataset [24].

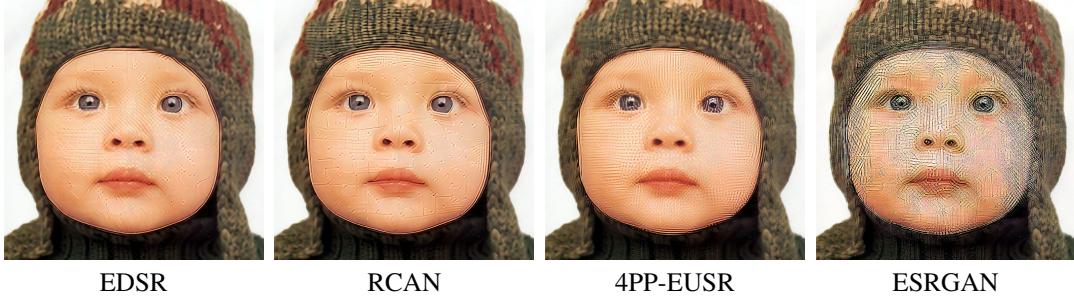


Figure 18. Results of the partial attack on the face region ($\alpha = 16/255$).



Figure 19. Targeted attack result using a score card image (Flickr, juggernautco, CC BY 2.0) with $\alpha = 16/255$ for ESRGAN. The attack targets to change the number in the red box to 89.

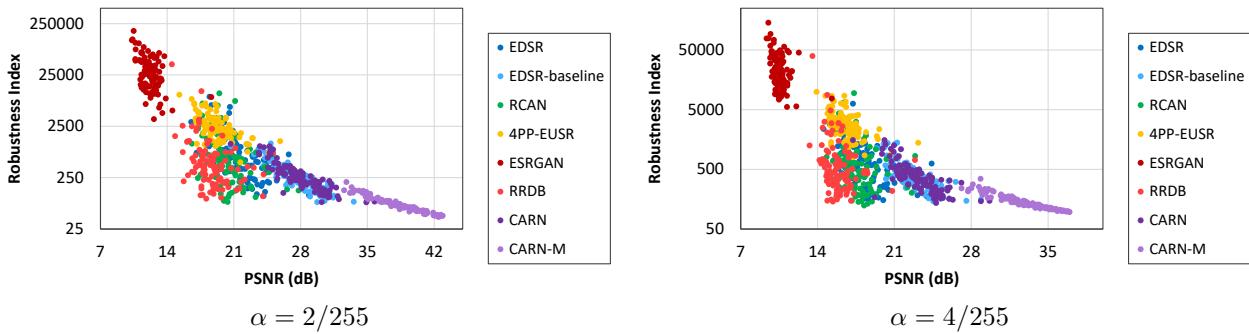


Figure 20. PSNR vs. the robustness index for the BSD100 dataset [14] when $\alpha = 2/255$ and $\alpha = 4/255$. Each point corresponds to each image in the dataset.