Evaluating Robustness of Deep Image Super-Resolution Against Adversarial Attacks

Jun-Ho Choi, Huan Zhang, Jun-Hyuk Kim, Cho-Jui Hsieh, and Jong-Seok Lee

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 upsampling have been used popularly, the development of deep learning-based approaches, which can be 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 upsampling 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. Wang 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 $X_0$ denote the original low-resolution input image and $X$ denote the attacked version of $X_0$. From these images, we obtain the super-resolved high-resolution images $f(X_0)$ and $f(X)$, 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:

$$\mathcal{L}(X, X_0) = \|f(X) - f(X_0)\|_2. \quad (1)$$

To find an $X$ to minimize (1) with bounded $\ell_\infty$-norm constraint ($\|X - X_0\|_\infty \leq \alpha$), we adopt the I-FGSM update rule that iteratively updates $X$ by:

$$\bar{X}_{n+1} = \text{clip}_{0,1} \left( X_n + \frac{\alpha}{T} \text{sgn}(\nabla \mathcal{L}(X_n, X_0)) \right) \quad (2)$$

$$X_{n+1} = \text{clip}_{-a, a} (\bar{X}_{n+1} - X_0) + X_0 \quad (3)$$

where $T$ is the number of iterations, $\text{sgn}(\nabla \mathcal{L}(X_n, X_0))$ is the sign of the gradient of (1), and

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

The term $\alpha$ 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 $X = X_T$.

#### 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
Then, (2) can be modified as:

\[ X_{n+1} = \text{clip}_{0,1}\left( X_n + \frac{\alpha}{T} \text{sgn}(\nabla L_M(X_n, X_0)) \circ M \right) \]  

(8)

where

\[ L_M(X, X_0) = \| (f(X) - f(X_0)) \circ (1 - M_H) \|_2. \]  

(9)

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

### 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 \( \Delta \) 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 \( M \) so as to attack the central part of the input image, i.e.,

\[ M(x, y) = \begin{cases} 
1 & \text{if } \frac{w}{4} \leq x < \frac{3w}{4}, \frac{h}{4} \leq y < \frac{3h}{4} \\
0 & \text{otherwise}
\end{cases} \]  

(10)

where \( M(x, y) \) is the value of \( M \) 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., \( X_0 \) and \( X \). For super-resolved (SR) images, PSNR is measured between the output images obtained from the original and attacked input images, i.e., \( f(X_0) \) and \( f(X) \). 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 $\alpha$ 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 $\alpha$ 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.

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-
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 \( \alpha \) 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 \( \alpha = 4/255 \). 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 \( \alpha \) 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.
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:

\[
\tilde{X}_{n+1} = \text{clip}_{[0,1]} \left( X_n - \frac{\alpha}{T} \text{sgn}(\nabla L(X_n, X^*)) \right)
\]

where \(X^*\) 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 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
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 \( X_0 \) 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 \( b_i = ||\nabla \mathcal{L}(X_0 + \Delta^{(i)}, X_0)||_1 \) 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 \( \alpha \) 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 \( \alpha \) 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


