Evaluating Robustness of Deep Image Super-Resolution Against Adversarial Attacks

— Supplementary Material —

Jun-Ho Choi[1], Huan Zhang[2], Jun-Hyuk Kim[1], Cho-Jui Hsieh[2], and Jong-Seok Lee[1]

[1]School of Integrated Technology, Yonsei University
{idearibosome,junhyuk.kim,jong-seok.lee}@yonsei.ac.kr
[2]Department of Computer Science, University of California, Los Angeles
huanzhang@ucla.edu chohsieh@cs.ucla.edu

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 1 shows additional example low-resolution (LR) and super-resolved (SR) images obtained from an image of Set14 [7] with $\alpha = 8/255$. Undesirable artifacts similar to those observed in Figure 2 of the main paper can be found. Figure 2 shows the example images obtained by the EDSR model [4] 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 [4] and CARN [1] show higher transferability than the other models, whereas RCAN [8] and ESRGAN [6] show lower transferability. Here, we visually explain the transferability of these four super-resolution methods in Figures 3, 4, 5, and 6. In the figures, a LR image in the BSD100 dataset [5] is attacked with one of the super-resolution models and inputted to the other super-resolution models including EDSR [4], EDSR-baseline, RCAN, 4PP-EUSR [3], ESRGAN, RRDB [6], CARN, and CARN-M [1]. In Figures 3 and 4, 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 5 and 6, 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 3 and 4).

Transferability of the universal attack. We examine the universal attack across datasets, i.e., the universal perturbation obtained for the BSD100 dataset [5] is applied to the images of the Set14 dataset [7]. Figure 7 shows the super-resolved (SR) images obtained by the RCAN model [8], 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 8 shows the attack results where the perturbation is applied on the face region of an image in Set5 [2]. 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 9 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 10 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.

References

[1] 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. 1, 4

[2] 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. 1

[3] 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. 1

[4] Bee Lim, Sanghyun Son, Heewon Kim, Seungjun Nah, and Kyoung Mu Lee. Enhanced deep residual networks for single image super-resolution. In *Proccedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 136–144, 2017. 1, 3, 4

[5] 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. 1, 4, 5, 6

[6] 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. 1, 5

[7] 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. 1, 3, 6

[8] 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. 1, 5

Figure 1. 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 [7], (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.

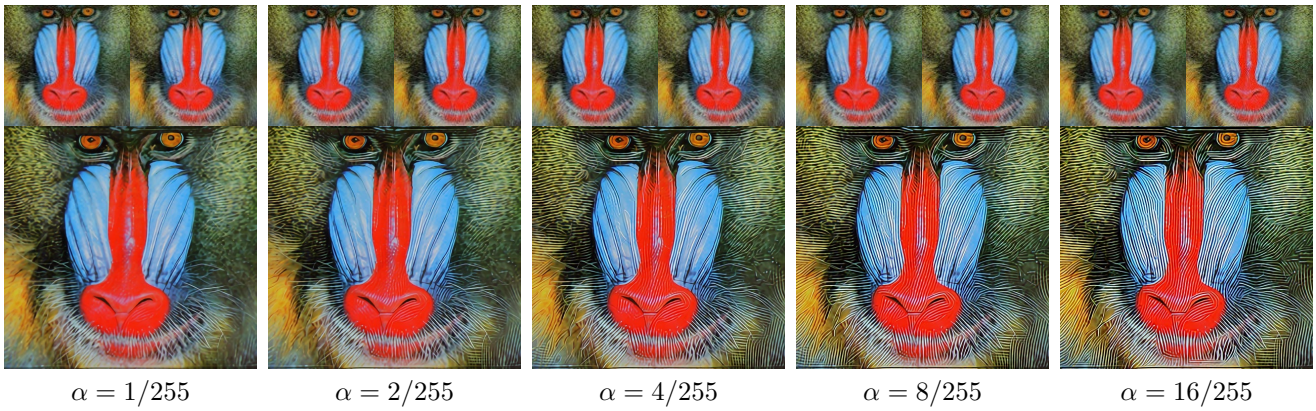


Figure 2. Visual comparison of the super-resolved outputs for the inputs attacked with different α values. In each case, (top-left) is the original input in Set14 [7], (top-right) is the adversarial input, and (bottom) is the output obtained on EDSR [4]. The input images are enlarged two times for better visualization.

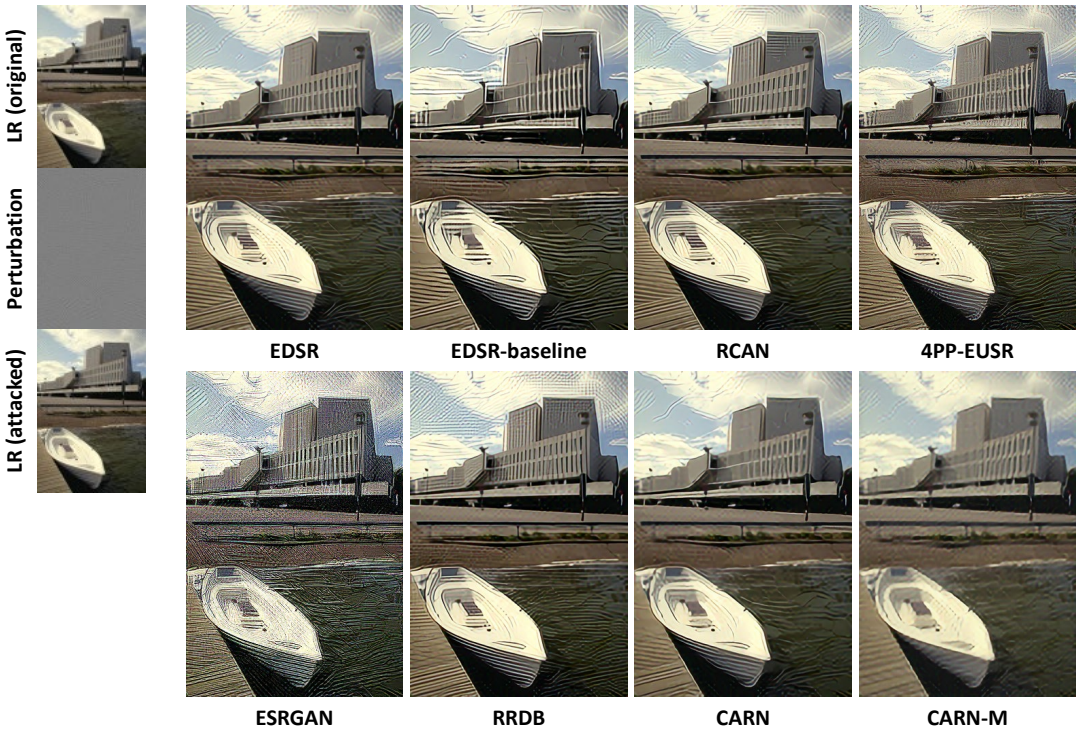


Figure 3. Visual examples of the transferred attack where EDSR-baseline [4] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [5] dataset is used.

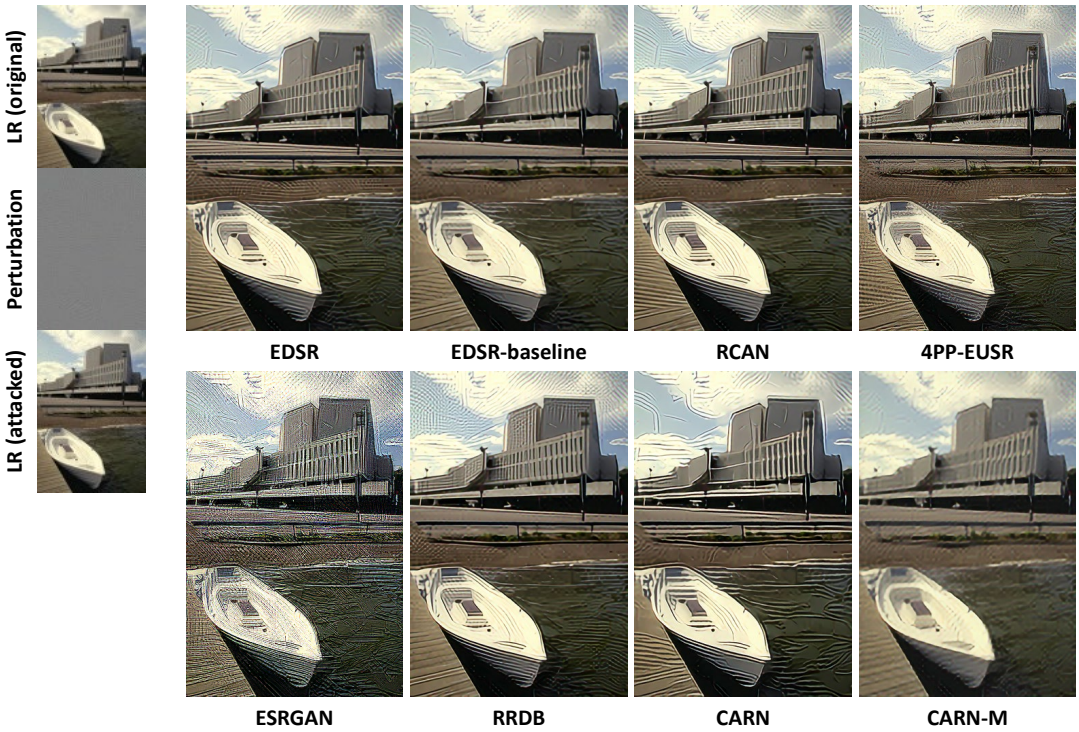


Figure 4. Visual examples of the transferred attack where CARN [1] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [5] dataset is used.

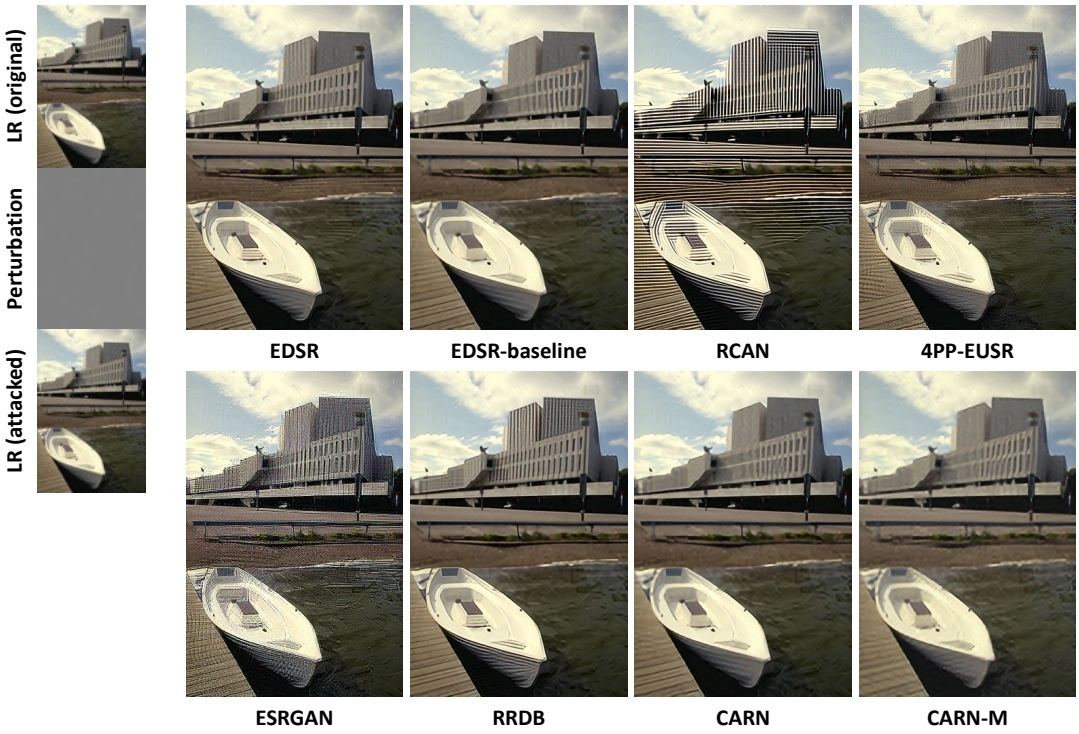


Figure 5. Visual examples of the transferred attack where RCAN [8] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [5] dataset is used.

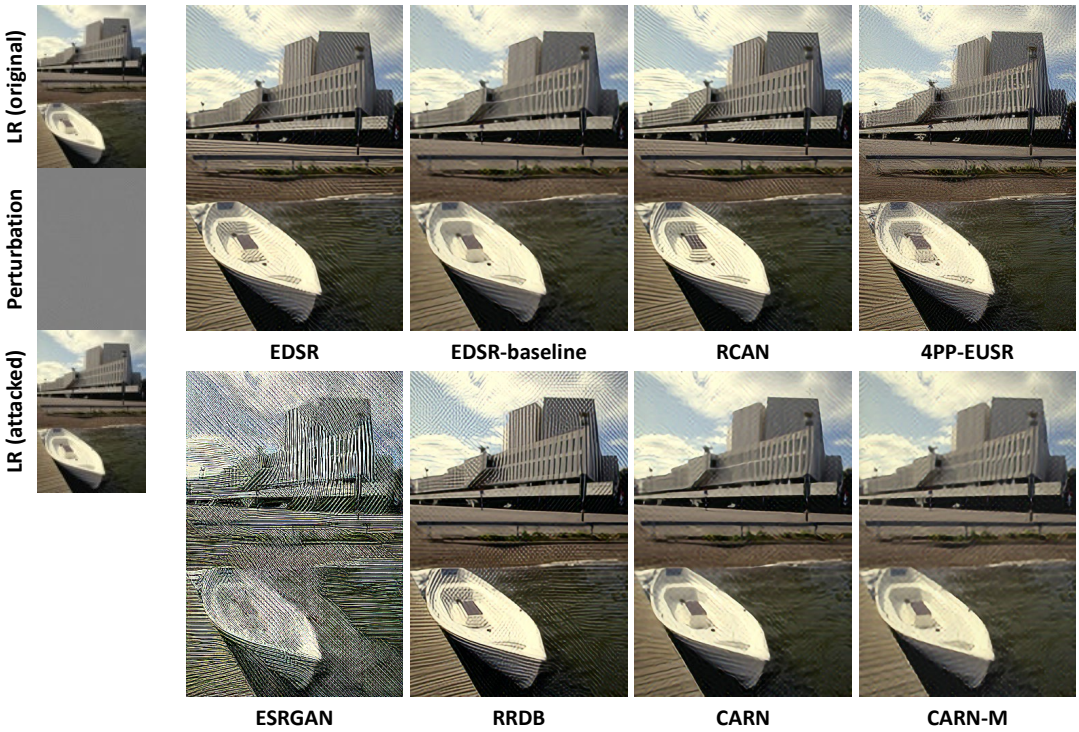


Figure 6. Visual examples of the transferred attack where ESRGAN [6] is used as the source super-resolution model with $\alpha = 8/255$. An image in the BSD100 [5] dataset is used.

Figure 7. Results of the universal attack applied to the Set14 dataset [7].

Figure 8. Results of the partial attack on the face region ($\alpha = 16/255$).

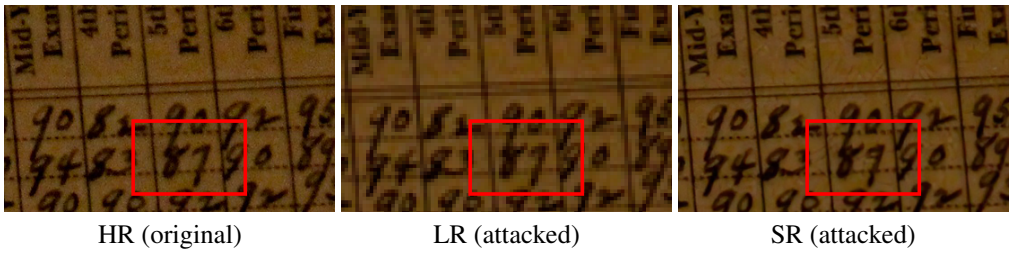
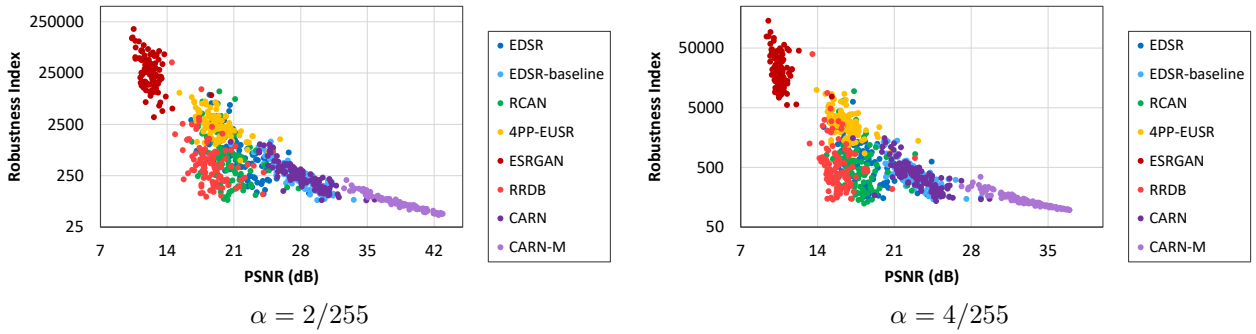


Figure 9. Targeted attack result using a score card image (Flickr, juggernautco, CC BY 2.0) with $\alpha = 16/255$ for ESRGAN. The attack targets to change the number in the red box to *89*.



$\alpha = 2/255$ $\alpha = 4/255$

Figure 10. PSNR vs. the robustness index for the BSD100 dataset [5] when $\alpha = 2/255$ and $\alpha = 4/255$. Each point corresponds to each image in the dataset.