Deep Image Destruction: Vulnerability of Deep Image-to-Image Models against Adversarial Attacks

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

Abstract—Recently, the vulnerability of deep image classification models to adversarial attacks has been investigated. However, such an issue has not been thoroughly studied for image-to-image tasks that take an input image and generate an output image (e.g., colorization, denoising, deblurring, etc.) This paper presents comprehensive investigations into the vulnerability of deep image-to-image models to adversarial attacks. For five popular image-to-image tasks, 16 deep models are analyzed from various standpoints such as output quality degradation due to attacks, transferability of adversarial examples across different tasks, and characteristics of perturbations. We show that unlike image classification tasks, the performance degradation on image-to-image tasks largely differs depending on various factors, e.g., attack methods and task objectives. In addition, we analyze the effectiveness of conventional defense methods used for classification models in improving the robustness of the image-to-image models.

I. INTRODUCTION

The deep learning technology brings us tremendous advantages in various computer vision fields. On the other side, recent studies have shown that deep learning-based algorithms are highly vulnerable to adversarial attacks, which add imperceptible noise to input images to fool the target deep model. Such vulnerability has been investigated chiefly on image classification models [1], [2], [3].

Meanwhile, deep image-to-image models have also been developed, which receive an input image and generate an output image. Unlike image classification models, image-to-image models cover much broader task objectives, including colorization [4], super-resolution [5], denoising [6], deblurring [7], and translation [8]. Image-to-image models are usually employed as pre-processing parts of computational systems or as standalone systems to generate images having better quality. A malicious attack can cause the whole system containing an image-to-image model to malfunction [9]; e.g., in a system performing image enhancement and then classification, attacking the enhancement model can affect its output so that the classification model produces a wrong result. Alternatively, an attack can change the visual content of the output image [10], e.g., changing the text or the color of a traffic light in the image, which can induce erroneous judgment of both human viewers and computers. Nevertheless, the vulnerability issue of deep models for image-to-image tasks has not been studied much compared to that of the image classification task. Because the way that image-to-image models deal with image data largely differs from that for image classification models, the characteristics of such a vulnerability issue may also differ from those for image classification.

Let X and Ā denote the original and attacked input images, respectively, i.e., Ā = X + Δ for a small perturbation Δ. For a given classification model fc(·), which outputs a class label y for input X, the goal of the adversarial attack is to achieve fc(Ā) ̸= y. By contrast, because the image-to-image models output images instead of class probability distributions, the goal of an adversarial attack can be set to maximize the difference between the original output fm(X) and the attacked output fm(Ā), where fm(·) is the target image-to-image model, i.e., maxΔ d(fm(X), fm(Ā)), where d(·, ·) is a distance measure. Due to this fundamental difference of adversarial attacks between classification models and image-to-image models, the following essential research questions need to be answered, which is the objective of this paper.

- How can we measure the success of adversarial attacks for image-to-image models? The metrics used in image classification models (e.g., success rate, fooling rate) cannot be used directly.
- What are the characteristics of adversarial attacks for image-to-image models? Because of the different properties of the models, the vulnerability of image-to-image models may appear differently in various factors.
- What are the right ways to improve the robustness of image-to-image models? The applicability of conventional defense methods used in image classification to image-to-image tasks needs to be examined.

We conduct comprehensive investigations on the vulnerability of deep image-to-image models against adversarial attacks. We expose five popular image-to-image tasks to malicious attacks and examine various factors of the vulnerability. In addition, we examine the feasibility of defense methods that are conventionally applied to image classification models.
II. RELATED WORK

Various adversarial attack methods for image classification models have been developed [2], [3]. Szegedy et al. [11] developed an optimization-based approach to make a given model produce wrong classification results with minimum amounts of input perturbation. Goodfellow et al. [1] proposed the fast gradient sign method (FGSM), which calculates the perturbation from the sign of the gradients obtained for a given model, and Kurakin et al. [12] extended it to an iterative approach, called I-FGSM. Ganeshan and Babu [13] proposed feature disruptive attack (FDA), which finds perturbation from the intermediate features of a given model. Other vulnerability issues also have been investigated such as finding universal perturbation [14], [15] and measuring transferability among different models [16], [17]. These mostly focus on the image classification task, and an in-depth study on vulnerability of image-to-image models has not been conducted.

Defense methods on classification models have also been proposed. One approach is to transform the input images before feeding them to a given model to reduce the amount of perturbation, including JPEG compression [18], bit depth reduction [19], and random resizing [20]. Another effective way is adversarial training, which uses images containing adversarial perturbation as training data [1], [12], [21].

III. METHODS

We consider 16 deep models for five popular image-to-image tasks: colorization (CIC [4]), deblurring (DeepDeblur [7]), denoising (three DnCNN models [6] for grayscale or color images with a single or multiple Gaussian noise levels (σ)), super-resolution (EDSR [5], RCAN [22], CARN [23], SRResNet [24], and SRGAN [24]), and translation (three CycleGAN models [8] for apple↔orange, horse↔zebra, and Van Gogh’s paintings↔photos). We consider two types of attack methods that are used for image classification models and are applicable to image-to-image models: feature-based attack (FDA [13]) and gradient-based attack (I-FGSM [10], [12]). The feature-based attack tries to reduce the variance of the intermediate activations in the target model. It does not rely on the model output to find perturbation, so it can be used to attack image-to-image models as well as classification models. The gradient-based attack iteratively finds the attacked input X from the gradient sign of the L2 difference between the original output and attacked output images. We set the pixel value scale of the images as [0, 255], and limit the $L_\infty$ norm of perturbation ε to ε = 8 in all cases except the universal attacks using ε = 16.

A. Quantitative Vulnerability Evaluation

One big challenge to evaluate image-to-image models is that it is hard to use evaluation methods used for image classification models directly (e.g., success rate or fooling rate). To this end, we design an evaluation metric based on the peak signal-to-noise ratio (PSNR), which is one of the most widely used metrics for evaluating the quality of image-to-image tasks. Given that the aim of attacks is to minimize the amount of perturbation in the input image but maximize the amount of deterioration in the output image, we first measure PSNR for both input and output images, denoted by $Q_i$ and $Q_o$, respectively. $Q_i$ is measured between the original input image $X$ and the attacked one $\tilde{X}$, indicating the amount of injected perturbation. $Q_o$ is measured between the output image for the original input image $f_m(X)$ and the output for the attacked input image $f_m(\tilde{X})$, quantifying the amount of deterioration due to the attack.

If an image-to-image model is more vulnerable to adversarial attacks than another, a smaller amount of perturbation injected into the input image can deteriorate the quality of its output image more significantly. In that case, $Q_i$ would be larger, while $Q_o$ would be smaller. Considering this, we define a vulnerability index ($VI$) that measures the degree of vulnerability, which is calculated as $VI = Q_i/Q_o$. A large VI indicates high vulnerability. We consider the degradations of both input and output images because we found that the amount of degradation in input images can significantly differ depending on the target image-to-image tasks and models, even under the same $\epsilon$ value. Note that other metrics such as SSIM can be used instead of PSNR for VI. Using SSIM, we observed similar trends to those using PSNR, which are omitted due to the space limit.

B. Transferability of Universal Perturbations

It is possible to find a universal perturbation that can affect any input image for a given model [14]. This can be obtained by applying a given attack method (i.e., FDA or I-FGSM) to the averaged deterioration, instead of the deterioration of each output image. In addition, because the universal perturbation does not rely on a specific input image, it enables us to investigate its transferability [16], [25], which refers to the applicability of the universal perturbation found for a model to another model. By measuring transferability of different perturbations, it is possible to determine whether some models have similar characteristics in terms of vulnerability.

We investigate the transferability of universal perturbations in two-fold. We examine the transferability of universal perturbations from one image-to-image model to another. In addition, we further investigate the transferability between image-to-image and classification models (VGG16 [26], ResNet-101 [27], MobileNetV2 [28]). These can clarify whether the perturbations are transferable across different tasks, not only between image-to-image tasks but also from/to the image classification task that is a completely different type of task.

C. Characteristics of Adversarial Examples

Image-to-image models exhibit various patterns of qualitative degradation depending on the characteristics of the perturbations. In classification, different characteristics of attack approaches are revealed mainly through changes in quantitative performance (e.g., success rate and computational complexity). In image-to-image tasks, however, the differences according to the input perturbations can be directly observed through changes in the output images. In this regard, we investigate
the characteristics of adversarial perturbations by manipulating frequency components during the I-FGSM process, which was recently applied to the classification task [29], [30]. Let \( \Delta^{(i)} \in \mathbb{R}^{W \times H \times C} \) denote the perturbation for a given input image \( X \) found at the \( i \)-th iteration, where \( W \), \( H \), and \( C \) are the width, height, and number of channels of the input image, respectively. We obtain low-frequency components of the perturbation by \( \tilde{\Delta}^{(i)} = \text{IDCT}(\mathcal{M} \circ \text{DCT}(\Delta^{(i)})) \), where \( \text{DCT}(\cdot) \) and \( \text{IDCT}(\cdot) \) are the 2-D discrete cosine transform (DCT) and inverse DCT, respectively, \( \circ \) denotes the element-wise multiplication, and \( \mathcal{M} \in \mathbb{R}^{W \times H \times C} \) is a binary mask that extracts certain frequency components of the perturbation. For attacking the low-frequency subspace, the value of \( \mathcal{M} \) at \( (w, h, c) \) is set to one when \( w \leq rW \) and \( h \leq rH \), and zero otherwise. The parameter \( r \in [0, 1] \) controls the range of the low-frequency subspace. Finally, the I-FGSM update rule is applied from the gradient sign of the L2 difference between the original output image and the output image obtained from \( X + \Delta^{(i)} \). We also investigate the effectiveness of attacking the high-frequency subspace by setting the value of \( \mathcal{M} \) at \( (w, h, c) \) to one when \( w \geq (1-r)W \) and \( h \geq (1-r)H \), and zero otherwise.

IV. EXPERIMENTAL RESULTS

A. Vulnerability Evaluation

Fig. 1 compares the performance of the models in terms of VI. We also employ random uniform noise within \([-\epsilon, \epsilon]\) as a baseline attack method. The VI values for the random noise are near 1, meaning that the input and output images deteriorate similarly, and it is barely effective as an attack. By contrast, FDA and I-FGSM show significantly larger VI values than 1 except for DnCNN, proving that the adversarial attack approaches work well. In addition, FDA shows the VI value close to 1 for DeepDeblur, while I-FGSM shows a much larger VI value. These can also be observed in Fig. 2. Both FDA and I-FGSM fail to conceal changes in the input images of DnCNN, and FDA fails to deteriorate the output images of DeepDeblur, while I-FGSM results in significant degradation. It proves that VI is a reasonable measure to evaluate vulnerability of the image-to-image models. In the following, we examine vulnerability patterns in more detail.

First, diverse degradation patterns are observed depending on the task. For instance, in Fig. 2, the attacks add hardly perceivable perturbation to the input images for all tasks. However, significant quality deterioration can be found in the output images with various deterioration textures except denoising (as shown in Fig. 1). The failure to attack the denoising models is because they are trained to identify and remove the elements unrelated to the original visual content, which include the adversarial perturbations. On the contrary, the models for the other tasks try to find latent information from a given image (e.g., color for colorization and high-frequency details for deblurring and super-resolution) and recover the corrupted content, during which the perturbations are amplified. Besides, because the visual components that each model tries to recover differ, different patterns of destruction are observed in the output images of different models.
Second, degradation patterns in the output images reveal the characteristics of the attack methods. For example, in Fig. 2, the output images under I-FGSM look more corrupted than those under FDA overall. This is also shown as larger VI values for I-FGSM in Fig. 1. Fig. 3 shows another examples in denoising. FDA largely removes the sharp textures in the output image, which is mainly due to the reduction of the activations. By contrast, I-FGSM tries to deteriorate the quality of the output image as much as possible, resulting in noisy patterns. These kinds of characteristics cannot be found in image classification, where the attack results are usually represented only as success rates regardless of attack methods.

Third, some image-to-image models employ an exquisite training mechanism that is not so much used in image classification models, which is generative adversarial networks (GANs) [31] to improve perceptual quality of the output images. We observe that these models output more deteriorated images than the models without GANs when attacked. In Fig. 1, DeepDeblur and CycleGAN, which employ GANs, show relatively larger VI values. This is also confirmed when SRResNet and SRGAN are compared in Fig. 4. These two models share the same model structure, but SRGAN is trained with a GAN-based loss function for better perceptual quality. The models trained with GANs usually produce perceptually appealing sharp textures [32]. Thus, the perturbation in the input image also tends to be intensified in the output image.

Fourth, different characteristics in terms of vulnerability and distortion patterns can be observed depending on the training dataset. In Fig. 1, the model converting apples to oranges shows the least vulnerability, while the model converting horses to zebras is the most vulnerable among the translation models. Note that these models share the same model structure and training mechanism. Fig. 5 depicts example images when FDA is employed. For converting apples to oranges, the model simply finds red regions and converts their colors to yellowish ones, meaning that it is primarily sensitive to color information in the input image. By contrast, due to grainy textures in the oranges, the model converting oranges to apples relies on complex analysis of the input image. Thus, there is more room for the attack to find effective perturbation for these models. This demonstrates that the vulnerability of the image-to-image models can differ depending on the employed dataset.

B. Universal Perturbations and Transferability

Fig. 6 shows the universal perturbation for each attack method and each task. First of all, the perturbation patterns differ depending on the attack method because of the differences in the attack objectives, i.e., while FDA tries to find perturbations that can reduce the intermediate activations, I-FGSM tries to maximize the amount of output deterioration directly. In addition, the overall patterns of the perturbation differ depending on the task. For example, grouped colorful pixels can be found in the perturbation of the denoising model for I-FGSM because Gaussian noise-like perturbations, obtained in the other tasks, tend to be removed in the output images due to the denoising process.

It is interesting to compare the perturbations obtained from the image-to-image models with those obtained from the image classification models. The former contains more granular textures, while the latter has thicker and bolder textures. One possible reason is due to the different architectures: The image classification models employ spatial aggregation operations (e.g., pooling) to reduce the number of intermediate features, while the image-to-image models usually do not employ such operations to minimize information loss that can directly damage quality of output images.

Fig. 7 compares VI for the universal attack. The results are similar to those obtained for image-specific attacks in Fig. 1, showing that the image-agnostic universal attack is successful although it is slightly less effective than the image-specific attack. Moreover, I-FGSM generates universal perturbations stronger than those of FDA because of the different objectives as in the image-specific attack.
We also investigate the transferability of universal perturbations. Specifically, we measure VI or a fooling rate of a target model for the universal perturbation found from another source model. Fig. 8 summarizes the results for I-FGSM. The universal perturbations are transferable between different models to some extent even across different tasks, although they are relatively more transferable when the source and target models are for the same task. For example, SRGAN for super-resolution is easily compromised by the universal perturbations found for the other tasks; especially, the VI values when the universal perturbations found for translation are used are as high as those when the universal perturbations found for the other super-resolution models are used. A similar observation can be made when CycleGAN (photo → Van Gogh) is set as a target model. It is thought that these two models are particularly vulnerable to universal perturbations because they try to recover complex textures using GANs.

In most cases, the transferability of universal perturbations between an image-to-image model and an image classification model is lower than that between image-to-image models or image classification models, which can be easily understood. Surprisingly, however, the universal perturbations are sometimes transferable between an image-to-image model and a classification model. For instance, the perturbation found in DnCNN can fool MobileNetV2 with a fooling rate of almost 50%. In Fig. 6b, the perturbation of DnCNN looks the most similar to that of the classification model in terms of texture patterns and variety of colors. However, the opposite is not applied; as the original attack on DnCNN does not work well, the perturbations of the classification models do not successfully attack DnCNN, either.

C. Characteristics of Adversarial Examples

We further investigate characteristics of perturbations by employing the frequency-aware attacks with $r = 1/4$. Fig. 9 shows the VI values for the low-frequency and high-frequency attacks. The results explain that the effectiveness of the frequency-aware attacks varies depending on the task. In particular, the deblurring model (DeepDeblur) is more vulnerable to the low-frequency attack than the high-frequency attack, while the super-resolution models are more vulnerable to the high-frequency attack.

Fig. 10 shows example images obtained from these models. In DeepDeblur, only the low-frequency attack significantly deteriorates the quality of the output image because the model exploits information in the low-frequency components (i.e., blurred regions) to recover missing details. By contrast, the output image of EDSR is affected by both the low-frequency and high-frequency attacks. However, the low-frequency attack also significantly deteriorates the quality of the input image. A super-resolution model finds latent textures through the high-frequency subspace of an input image; thus, the high-frequency attack works well. However, the low-frequency features in the input image, which do not have notable textures, appear in the output image without much change. Hence, the low-frequency attack tries to add a large amount of perturbation to degrade the output, which leads to clearly visible deterioration in both the input and output images. These show that the frequency-dependent degree of vulnerability is determined by the model’s mechanism optimized for the task.

V. DEFENDING IMAGE-TO-IMAGE MODELS

Finally, we evaluate the feasibility of defending image-to-image models by employing defense methods that have been widely used for image classification models: image transformation and adversarial training. For transformation-based
approaches, we consider JPEG compression [18], random resizing [20], and bit reduction [19]. In addition, we consider geometric self-ensemble that was introduced in the super-resolution [5] to improve quality of the output images without additional training. For adversarial training, adversarial examples are obtained by I-FGSM.

Fig. 11 compares the performance of the defense methods in terms of VI, when I-FGSM is employed. Overall, the defense methods reduce influences of adversarial perturbations to some extent, i.e., the VI values are reduced. The JPEG compression shows the highest effectiveness. Nevertheless, the VI values are still larger than 1, which implies that deterioration in the output images is not completely removed.

We note that defending image-to-image models brings another concern: It is not desirable if the original performance without attack is degraded due to the application of a defense method. For classification models, slight changes in the unattacked input image due to, for instance, image transformation-based defense do not alter the image content much and thus the classification result is expected to remain the same. However, for image-to-image models, small changes in the input image due to defense can significantly affect the quality of the output image. In order to examine this, we compute original PSNR values from the ground-truth and output images when a defense is applied to unattacked input images. Fig. 12 shows that the defense methods tend to lower the original performance, especially when transformation-based approaches except the geometric self-ensemble are employed. The geometric self-ensemble most successfully preserves the original performance, which is followed by adversarial training. JPEG compression, which is the most effective defense method in Fig. 11, harms the original performance the most.

Fig. 13 provides visual examples for super-resolution (EDSR). It shows limitations of applying defense methods to image-to-image models, i.e., significant deterioration is introduced in the output images when the input image is not attacked, except the self-ensemble method that inherently aims to improve the original performance. Furthermore, all defense methods are not so effective when an adversarial attack is applied; the deterioration is still clearly visible although the amount of deterioration is reduced.

We can conclude that the existing defense methods developed for image classification models are not well-suited for image-to-image models because of the distinguished characteristics of defense of image-to-image models compared to image classification models. In other words, it is challenging for the defense methods to satisfy two conditions simultaneously: to completely remove the effect of the adversarial perturbation in the output image and to maintain the original performance when no attack exists.

VI. CONCLUSION

We presented in-depth analysis on vulnerability of deep image-to-image models against adversarial attacks. Our study aimed to 1) examine vulnerability patterns in various aspects with a proper measure of success of adversarial attacks, 2) find characteristics of image-to-image models in terms of vulnerability across different tasks, and 3) evaluate applicability of existing defense methods. To this end, we defined a measure for image-to-image models, vulnerability index (VI), and our results showed that most state-of-the-art image-to-image models are highly vulnerable to adversarial attacks and universal perturbations are transferable even across tasks. However, the attack results and the perturbation patterns have significantly different characteristics depending on the task type, attack method, training mechanism, and dataset. Finally, our results showed that many of the existing defense methods used for classification models are not suitable to image-to-image models and suggested the necessity to develop effective defense methods for image-to-image models in the future.

ACKNOWLEDGMENT

This work was supported in part by the Artificial Intelligence Graduate School Program, Yonsei University under Grant 2020-0-01361, and in part by the Ministry of Trade, Industry and Energy (MOTIE) under Grant P0014268.

1292
REFERENCES

[1] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and harnessing adversarial examples,” in Proceedings of the International Conference on Learning Representations, 2015.

[2] D. Su, H. Zhang, H. Chen, J. Yi, P.-Y. Chen, and Y. 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, 2018.

[3] Y. Dong, Q.-A. Fu, X. Yang, T. Pang, H. Su, Z. Xiao, and J. Zhu, “Benchmarking adversarial robustness on image classification,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.

[4] R. Zhang, P. Isola, and A. A. Efros, “Colorful image colorization,” in Proceedings of the European Conference on Computer Vision, 2016.

[5] B. Lim, S. Son, H. Kim, S. Nah, and K. M. Lee, “Enhanced deep residual networks for single image super-resolution,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2017.

[6] K. Zhang, W. Zuo, Y. Chen, D. Meng, and L. Zhang, “Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising,” IEEE Transactions on Image Processing, vol. 26, no. 7, pp. 3142–3155, 2017.

[7] S. Nah, T. Hyun Kim, and K. Mu Lee, “Deep multi-scale convolutional neural network for dynamic scene deblurring,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[8] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, “Unpaired image-to-image translation using cycle-consistent adversarial networks,” in Proceedings of the IEEE International Conference on Computer Vision, 2017.

[9] M. Yin, Y. Zhang, X. Li, and S. Wang, “When deep fool meets deep prior: Adversarial attack on super-resolution network,” in Proceedings of the ACM International Conference on Multimedia, 2018.

[10] I.-H. Chou, H. Zhang, J.-H. Kim, C.-J. Hsieh, and J.-S. Lee, “Evaluating robustness of deep image super-resolution against adversarial attacks,” in Proceedings of the IEEE International Conference on Computer Vision, 2019.

[11] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, “Intriguing properties of neural networks,” in Proceedings of the International Conference on Learning Representations, 2014.

[12] A. Kurakin, I. Goodfellow, and S. Bengio, “Adversarial machine learning at scale,” in Proceedings of the International Conference on Learning Representations, 2017.

[13] A. Ganesan and R. V. Babu, “FDA: Feature disruptive attack,” in Proceedings of the IEEE International Conference on Computer Vision, 2019.

[14] S.-M. Moosavi-Dezfooli, A. Fawzi, O. Fawzi, and P. Frossard, “Universal adversarial perturbations,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[15] C. Zhang, P. Benz, C. Lin, A. Karjaun, J. Wu, and I. S. Kweon, “A survey on universal adversarial attack,” in Proceedings of the International Joint Conference on Artificial Intelligence, 2021.

[16] Y. Liu, X. Chen, C. Liu, and D. Song, “Delving into transferable adversarial examples and black-box attacks,” in Proceedings of the International Conference on Learning Representations, 2016.

[17] X. Wang and K. He, “Enhancing the transferability of adversarial attacks through variance tuning,” in Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2021.

[18] G. K. Drizgaiate, Z. Ghahramani, and D. M. Roy, “A study of the effect of JPG compression on adversarial images,” arXiv:1609.00853, 2016.

[19] W. Xu, D. Evans, and Y. Qi, “Feature squeezing: Detecting adversarial examples in deep neural networks,” in Proceedings of the Annual Network and Distributed System Security Symposium, 2018.

[20] C. Xie, J. Wang, Z. Zhang, Z. Ren, and A. Yuille, “Mitigating adversarial effects through randomization,” in Proceedings of the International Conference on Learning Representations, 2018.

[21] T. Pang, X. Yang, Y. Dong, H. Su, and J. Zhu, “Bag of tricks for adversarial training,” in Proceedings of the International Conference on Learning Representations, 2021.

[22] Y. Zhang, K. Li, K. Li, L. Wang, B. Zhong, and Y. Fu, “Image super-resolution using very deep residual channel attention networks,” in Proceedings of the European Conference on Computer Vision, 2018.

[23] N. Ahn, B. Kang, and K.-A. Sohn, “Fast, accurate, and lightweight super-resolution with cascading residual networks,” in Proceedings of the European Conference on Computer Vision, 2018.

[24] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang et al., “Photo-realistic single image super-resolution using a generative adversarial network,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017.

[25] L. Wu and Z. Zhu, “Towards understanding and improving the transferability of adversarial examples in deep neural networks,” in Proceedings of the Asian Conference on Machine Learning, 2020.

[26] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” arXiv:1409.1556, 2014.

[27] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016.

[28] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, “MobileNetV2: Inverted residuals and linear bottlenecks,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.

[29] C. Guo, J. S. Frank, and K. Q. Weinberger, “Low frequency adversarial perturbation,” in Proceedings of the Conference on Uncertainty in Artificial Intelligence, 2019.

[30] Y. Sharma, G. W. Ding, and M. A. Brubaker, “On the effectiveness of low frequency perturbations,” in Proceedings of the International Joint Conference on Artificial Intelligence, 2019.

[31] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, “Generative adversarial nets,” in Proceedings of the Advances in Neural Information Processing Systems, 2014.

[32] Y. Blau and T. Michaeli, “The perception-distortion tradeoff,” in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018.